Exploration of Regulatory Affairs for Medical Devices Using Data-Driven Approaches and AI



Yu Han

Reuben College

University of Oxford

A thesis submitted for the degree

Doctor of Philosophy

Michaelmas Term 2025

Acknowledgements

I am deeply grateful to my supervisor, Professor Jeroen H. M. Bergmann, for his guidance, encouragement, and steady support throughout this doctoral work. His insight shaped both the direction of this thesis and the way I approach research, and I could not have completed it without his patience and generosity with his time.

I owe a particular debt to Dr Aaron Ceross, who has guided and supported me since before I even began my doctorate. He mentored me through research step by step, with extraordinary patience and from the very basics — when I started, I did not even know how to compile a document in Overleaf. His advice on regulatory science and methodology was central to this project, and his careful reading improved every chapter on which we worked together. I also thank my co-authors and collaborators — in particular Dr Florence Bourgeois, Dr Paulo Savaget, and Dr Sarim Ather — for their contributions to the studies that underpin several chapters of this thesis. I am grateful to the members of the Natural Interaction Lab at Oxford for many helpful discussions and, in particular, for their generous help with mock vivas and feedback as I prepared for the examination. More broadly, I am thankful for the many people I have been fortunate to meet at Oxford, whose friendship, conversation, and kindness made these years far richer than I could have hoped.

I gratefully acknowledge the Edgell Sheppee Fund for its financial support of this research, and the University of Oxford and the Department of Engineering Science for providing the resources and environment in which it was carried out. I am especially grateful to Reuben College for its warmth and care, which gave me a happy and supportive college life throughout my time at Oxford.

Finally, and most personally, I thank my parents and my dear grandma and grandpa. As the first person in my family to pursue a doctorate, and as an international student far from home, I am profoundly grateful for their unwavering belief in me and their constant support throughout my studies. Last, I thank my husband, Xi Xiong, for his companionship through so many hours of research and for helping me carry its pressures more lightly.

Contents

List of Figures viii

List of Abbreviations xi

1 Introduction 1
- 1.1 Background of Regulatory Affairs for Medical Devices . . . 1
- 1.2 Motivation . . . 5
- 1.3 Research Problem . . . 9

2 A Preliminary Overview for Software Medical Devices 14
- 2.1 Introduction . . . 14
- 2.2 China Artificial Intelligence Development Plan . . . 15
- 2.3 Chinese Regulation on Medical Device Software . . . 16
- 2.4 Challenges Posed by Software and AI in Medical Devices . . . 21
- 2.5 Future Directions . . . 22

3 Exploration of Software Medical Devices Using a Chinese Regulatory Database 25
- 3.1 China's New Generation Artificial Intelligence Development Plan and Effect to Medical Devices . . . 25
- 3.2 Methods . . . 29
- 3.3 Results . . . 32
  - 3.3.1 Devices amount and category . . . 32
  - 3.3.2 Artificial intelligence devices and algorithms in medical software 35
  - 3.3.3 Exploratory Observations: Algorithm Types and Classification 37
  - 3.3.4 Observed Differences in International Regulatory Classification 43
- 3.4 Discussion . . . 45
- 3.5 Conclusions . . . 51

4 Overview of Quantitative Measurements of Linguistic Complexity for Regulations 55
- 4.1 Introduction . . . 55
- 4.2 Methods . . . 58
- 4.3 Results . . . 59
- 4.4 Discussion . . . 68

5 Qualitative and quantitative methods to explore complexity in Regulatory Affairs 72
- 5.1 Research design and methods . . . 74

5.1.1 Research Ethics . . . 74
5.1.2 Semi-structured interviews . . . 75
5.1.3 Data Collection . . . 77
5.1.4 Data Analysis . . . 77
5.2 Results . . . 79
5.2.1 Results of Open Coding . . . 79
5.2.2 Interview Theme 1 : Complexity in Regulatory Language . . . 80
5.2.3 Interview Theme 2: Complexity in Registration Process . . . 84
5.2.4 Interview Theme 3: Complexity in Available Database . . . 89
5.2.5 Interview Theme 4: Complexity with Product . . . 90
5.2.6 Interview Theme 5: Complexity at a Global Level . . . 92
5.2.7 Code analysis with NLP . . . 92
5.3 Conclusions . . . 97
5.3.1 The wider context of the interview findings . . . 97
5.3.2 Limitation . . . 99
5.3.3 Future work . . . 100

6 Evaluating Precision: The Role of Large Language Models in Medical Device Classification 103
6.1 Introduction . . . 103
6.2 Methods . . . 105
6.3 Results . . . 108
6.4 Discussion . . . 113
6.5 Limitation . . . 115
6.6 Conclusion . . . 116

7 Text Analysis, Machine Learning and Requirement Engineering 117
7.1 Introduction . . . 117
7.2 Related Work . . . 121
7.2.1 Rule-Based Systems . . . 121
7.2.2 Traditional Machine Learning . . . 121
7.2.3 Deep Learning Models . . . 122
7.2.4 Transformer-Based Models and LLMs . . . 122
7.2.5 Model Interpretability in Regulatory Contexts and Computational Cost . . . 122
7.2.6 Trade-off . . . 123
7.3 Theoretical Framework : Text classification methods . . . 124
7.3.1 Rule-Based Classification . . . 124
7.3.2 Traditional Machine Learning Models . . . 125
7.3.3 Deep Learning Models . . . 127
7.4 Experiments . . . 133
7.4.1 Dataset and Preprocessing . . . 133
7.4.2 Text Representation . . . 137
7.4.3 Train–Validation–Test Split . . . 138
7.4.4 Classifiers and Training Setup . . . 140
7.4.5 Interpretability . . . 141
7.5 Experiments and Results . . . 143

7.5.1 Model Accuracy . . . 143
7.5.2 Model Interpretability . . . 145
7.5.3 Model Computational Cost . . . 152
7.6 Discussion . . . 153
7.6.1 Limitations and Future Work . . . 169
7.6.2 Conclusion . . . 171

8 Multimodal AI Methods for Regulatory Affairs Task 173
8.1 Introduction . . . 173
8.2 Related Work . . . 176
8.2.1 Traditional Machine Learning Models for Regulatory/Healthcare 176
8.2.2 Multimodal Fusion Architectures . . . 176
8.2.3 Domain-Specific Models vs. Foundation Models . . . 177
8.2.4 Multimodal Learning in Healthcare and Regulation . . . 177
8.2.5 Gaps in Prior Work and Research Questions . . . 178
8.3 Methodology . . . 178
8.3.1 Problem Definition . . . 178
8.3.2 Dataset and Preprocessing . . . 179
8.3.3 Feature Extraction . . . 180
8.3.4 Multimodal Model Architecture . . . 182
8.3.5 Training Strategy . . . 185
8.3.6 Evaluation . . . 188
8.3.7 Explainability . . . 189
8.3.8 Data Representativeness and Selection Bias . . . 190
8.4 Experiments and Results . . . 191
8.4.1 Model Performance . . . 191
8.4.2 Robustness and Generalization . . . 197
8.4.3 Error Analysis and Explainability . . . 202
8.4.4 Computational Efficiency Comparison with CLIP Baselines . . 206
8.5 Discussion . . . 207
8.5.1 Key Findings in Regulatory Context . . . 207
8.5.2 Generalizability and Limitations . . . 209
8.5.3 Deployment Pathway and Regulatory Considerations . . . 213
8.6 Conclusion and Future Work . . . 216

9 Discussions 218
9.1 Thesis Summary . . . 218
9.2 Research Contributions . . . 220
9.3 Limitations . . . 221
9.4 Future Work . . . 222

Appendices

1 NMPA Medical Device Software Catalogue 225

2 AI-Enabled Medical Devices Identified in the NMPA UDI Registry 231

3 Statement of Authorship 234

References **237**

Abstract

Across jurisdictions, medical device regulation has grown more complex, yet there is still no objective way to measure that complexity or to ease the burden on the manufacturers and regulators who navigate it. This thesis develops AI methods for regulatory affairs, applying large language models alongside traditional algorithms to ease communication between authorities and manufacturers and support cross-regional device registration. The methods are validated using China's National Medical Products Administration (NMPA) as one application setting, but the methods and trade-offs they expose are general to regulatory science, not specific to any single authority.

The work proceeds in three stages. First, it characterises the regulatory landscape through a policy analysis of software- and AI-based device regulation, an exploratory NMPA-registry analysis isolating the small software- and AI-based subset, and a systematic review finding general-purpose readability formulas poorly suited to regulatory text. Second, it identifies the sources of complexity through semi-structured interviews that surface five recurring sources, and benchmarks large language models on classification, showing accuracy to be strongly jurisdiction-dependent. Third, it builds AI models for automatic device risk classification: a systematic comparison of rule-based, machine-learning, deep-learning, and large-language-model approaches across accuracy, interpretability, and cost; and a multimodal transformer fusing textual and visual device data with self-training, raising accuracy on a curated dataset from roughly four-fifths to over nine-tenths, with calibrated probabilities, no high-risk misclassifications, and SHAP auditability.

The central finding is that the trade-offs among accuracy, interpretability, and cost are intrinsic to the AI methods rather than artefacts of the NMPA calibration setting. The principal contribution is to show how AI methods can begin to measure and reduce regulatory complexity for manufacturers and regulators alike, and when such models could transfer across jurisdictions.

List of Figures

1.1 Hierarchy of Laws, Regulations, and Guidance, and Classification of Medical Devices in the USA. The pyramid indicates the hierarchical relationship, with laws at the base having the highest authority, followed by regulations and guidance documents. Standards (e.g. ISO 13485, IEC 62304, ISO 14971) are shown at the apex: they are voluntary rather than legally mandatory, but are "harmonised" in the EU (conferring a presumption of conformity) and "recognised" by the FDA in the US; where a harmonised standard is not applied, an "at least equivalent" approach is expected. 3

1.2 High-level schematic of the medical device regulatory lifecycle phases (N1–N7) that contribute to regulatory complexity. The figure highlights key stages from device classification to post-marketing surveillance, and identifies decision nodes in medical device lifecycle. Non-clinical *testing* precedes any *clinical trial*, since appropriate non-clinical test methods must be demonstrated before first human exposure. 5

1.3 Research question, Methodology and Contributions of Dissertation . . 12

2.1 Categories of medical device software 19

2.2 Medical device software classification flowchart 19

3.1 Medical device registration pathways set out by the NMPA. Imported and domestic devices relates to all class II and III devices, for Class I devices, only record-filing is required. 26

3.2 UDI code layout according to YY/T 1630-2018. Note: UDI-PI elements (e.g., version number [.V], manufacturing date [.M], expiration date [.E], lot [.L], sterilization batch [.D], and serial number [.S]) are optional and may not appear in all UDIs. 28

3.3 Flowchart for the selection process for the identification of software and AIMD 31

3.4 Distribution of domestic and imported SaMD and SiMD devices by risk class, stratified by AI and non-AI designation. Note: Due to substantial differences in device counts across categories (spanning more than an order of magnitude), each subplot uses an independently scaled y-axis to preserve visibility of within-category distributions. As a result, bar heights should be interpreted within, rather than across, subplots. . . . 33

3.5 A geographic map of non-foreign medical device software are registration 2020–2024 (n = 2,149). The areas include mainland provinces as well as the indication of , which indicates products from the Special Administration Regions. 34

3.6 An alluvial diagram illustrating the types of 43 AI medical devices and the risk classifications along with the approval pathways within the NMPA regulatory framework 36
3.7 Distribution of AIMD on different medical specialty and algorithm . . . 38
3.8 Applications of AI in Different Medical Specialties 41

4.1 PRISMA flowchart of systematic review. The process consists of identification, duplicate removing, screening and inclusion of relevant papers. 59

5.1 Navigating Complexity in Regulatory Affairs with Nvivo 80
5.2 Inter-topic distance map (produced with pyLDAvis). Each circle is a topic; the distance between circles approximates topic similarity via multidimensional scaling of the topic–term distributions. The expression at the foot of the figure is the pyLDAvis term-*relevance* metric, $r(w, t \mid \lambda) = \lambda\, p(w \mid t) + (1 - \lambda)\, p(w \mid t)/p(w)$, in which $\lambda \in [0, 1]$ trades off a term's probability within a topic against its overall corpus frequency when ranking the terms most characteristic of each topic. 96

6.1 Comparative Analysis of Large Language Model Performance in Medical Device Classification: Ground Truth Versus Predictive Accuracy Across Diverse Models 111
6.2 Accuracy Results Visualization 112

7.1 Text process, Model Training and Evaluation Pipeline 134
7.2 NMPA device Corpora Characteristics in Classification I, II and III Devices 136
7.3 Kernel density estimate of the character-count distribution across the whole corpus. The horizontal axis is the number of characters in a device description; the vertical axis is the estimated probability *density*, normalised so that the area under the curve integrates to one (it is a relative frequency per unit character count, not a raw device count). . . 137
7.4 Density of Corpora by classification 138
7.5 Vocabulary overlap between classifications. Values indicate the number of shared unique tokens between each pair of device risk classes. 139
7.6 SVM LIME Explanation: Text is analyzed by highlighting the most influential words based on their contribution to the predicted class. . . 149
7.7 Logistic Regression Anchor Explanation: Anchor words identified for each class with precision and coverage metrics. 149
7.8 RNN Integrated Gradients Explanation: Token-level attribution showing how each word influences the model's decision. 150
7.9 TreeSHAP-based explanations for BERT embeddings using ensemble models. 151
7.10 Example of Governance-Aware Interpretability in LLM-Based Medical Device Classification 152
7.11 AI Algorithm Decision Tree in Different Regulatory Affairs Scenarios . 168

8.1 Multimodal training and evaluation pipeline. Text features from BERT-base-Chinese and image features from EfficientNet-B5 are fused via a cross-attention Transformer with an SVM head. A self-training loop augments supervision 182
8.2 Comparison of classification accuracy across text-only, image-only, and multimodal models with and without self-training. 192
8.3 Dataset sufficiency and self-training efficiency. (Top) Learning curves exhibit diminishing performance gains as the proportion of labeled training data increases, with marginal improvements of less than 0.3% when expanding from 85% to 100%, suggesting that additional labeled data yields limited benefit under the current setting. (Bottom) The self-training curve indicates substantial performance improvements from semi-supervised learning, with gains gradually saturating after approximately 2,721 pseudo-labeled samples. 201
8.4 Class I (minority class) learning curve illustrating F1-score progression with increasing training set size. Performance improves from 0.60 at 30% of the training data to 0.74 when using the full dataset. A performance plateau is observed beyond 85% of the data, with only a marginal gain of 1.76% when increasing from 85% to 100%. These results suggest that minority-class performance continues to benefit from access to the full dataset, despite severe class imbalance (39 original samples). 202
8.5 SHAP attribution breakdown for three representative Class III to Class II errors, showing text-dominance pattern (93–94% attribution) and low prediction confidence (<0.55), well below the proposed 0.95 deployment threshold. 203
8.6 SHAP summary plot showing text (75%) vs. image (25%) feature contributions. 204
8.7 Grad-CAM visualizations highlighting discriminative regions on medical device images. Each heatmap shows where the model focuses when predicting device risk classes. 205
8.8 Calibration curve for the post–self-training SVM 206
8.9 Prediction Accuracy Rate Across Different Classes 212

List of Abbreviations

NMPA . . . . . National Medical Products Administration
FDA . . . . . Food and Drug Administration
TPLC . . . . . Total Product Life Cycle
IFU . . . . . Instruction for Use
PMC . . . . . Pre-market Certification
PMA . . . . . Pre-market Approval
IMDRF . . . . . International Medical Device Regulators Forum
GHTF . . . . . Global Harmonization Task Force
CRIE . . . . . Chinese Readability Index Explorer
SaMD . . . . . Software as a Medical Device
SIMD . . . . . Software in a Medical Device
MDSW . . . . . Medical Device Software
AI . . . . . Artificial Intelligence
ML . . . . . Machine Learning
NLP . . . . . Natural Language Processing
IVD . . . . . In Vitro Diagnostic

1 Introduction

1.1 Background of Regulatory Affairs for Medical Devices

Regulatory Affairs (RA) plays a central role in guiding the development and oversight of medical devices. According to the U.S. Food and Drug Administration (FDA), RA involves the science of creating tools, standards, and approaches for evaluating FDA-regulated products. As a multidisciplinary field, regulatory affairs encompasses activities that support product development, regulatory evaluation, and compliance throughout the device lifecycle, ultimately ensuring that medical products meet established regulatory and quality requirements [1]. It is a complex and ever-evolving process that involves iterative regulation of medical products, which governs the entire process of the creation of medical devices, from early-stage design and innovation to post-market monitoring and eventual disposal. This cross-functional process requires a deep understanding of various fields, including product development, medical research, engineering, legal and regulatory requirements, data science, product patents, legal disputes, and financial considerations. The ultimate goal of regulatory affairs is to deliver safe and effective products to patients who need them.[1]

Regulatory authorities, such as the FDA in the United States, enforce legal frameworks such as the Federal Food, Drug, and Cosmetic Act, which provides specific

[1]The standard against which devices are judged is expressed differently across jurisdictions—for example "safe and effective" in the United States (FDA), "quality, safety and performance" under the EU Medical Devices Regulation, and "safety and effectiveness" (and "quality") under China's NMPA. As this thesis concerns device *classification* rather than the underlying clinical evidence, these expressions ("safe and effective", "quality and safety", "performance, safety and quality") are used somewhat interchangeably; the distinctions are not material to the classification methods developed here.

pathways for device approval based on classification. For example, the FDA offers the 510(k) pathway for Class II devices and the Pre-Market Approval (PMA) pathway for Class III devices. In addition to the Federal Food, Drug, and Cosmetic Act, pivotal legislation, such as the 1997 Modernization Act and other acts, amendments, and provisions enacted by Congress, forms the backbone of the U.S. medical device regulatory system.

Regulations, developed in accordance with governing legislation, are binding requirements for stakeholders, including manufacturers, distributors, and healthcare providers. Meanwhile, guidance documents, though non-binding, provide authoritative and up-to-date insights, such as testing methods or dossier preparation guidelines, to aid industry compliance. Similar regulatory frameworks exist internationally, including Japan's PMDA Ninsho and Shonin pathways and the European Union's MDR regulations, and China's NMPA regulations. Figure 1.1 illustrates the hierarchy of laws, regulations, and guidance in the context of medical device governance in the USA, alongside examples of risk classifications for medical devices. Beyond laws, regulations and guidance, *standards* also play a part: although not legally mandatory, standards are "harmonised" in the EU—where demonstrating conformity to a harmonised standard confers a presumption of conformity—and "recognised" by the FDA in the US; where a harmonised standard is not applied, an approach at least equivalent to it is expected.

Per Section 201(h)(1) of the Food, Drug, and Cosmetic Act, as defined by FDA (Food and Drug Administration), a device is: An instrument, apparatus, implement, machine, contrivance, implant, in vitro reagent, or other similar or related article, including a component part, or accessory which is:

A Recognized in the official National Formulary, or the United States Pharmacopoeia, or any supplement to them,

B Intended for use in the diagnosis of disease or other conditions, or in the cure, mitigation, treatment, or prevention of disease, in man or other animals, or

C Intended to affect the structure or any function of the body of man or other animals, and which does not achieve its primary intended purposes through chemical action within or on the body of man or other animals and which is not dependent upon being metabolized for the achievement of its primary intended purposes. The term "device" does not include software functions excluded pursuant to section 520(o).

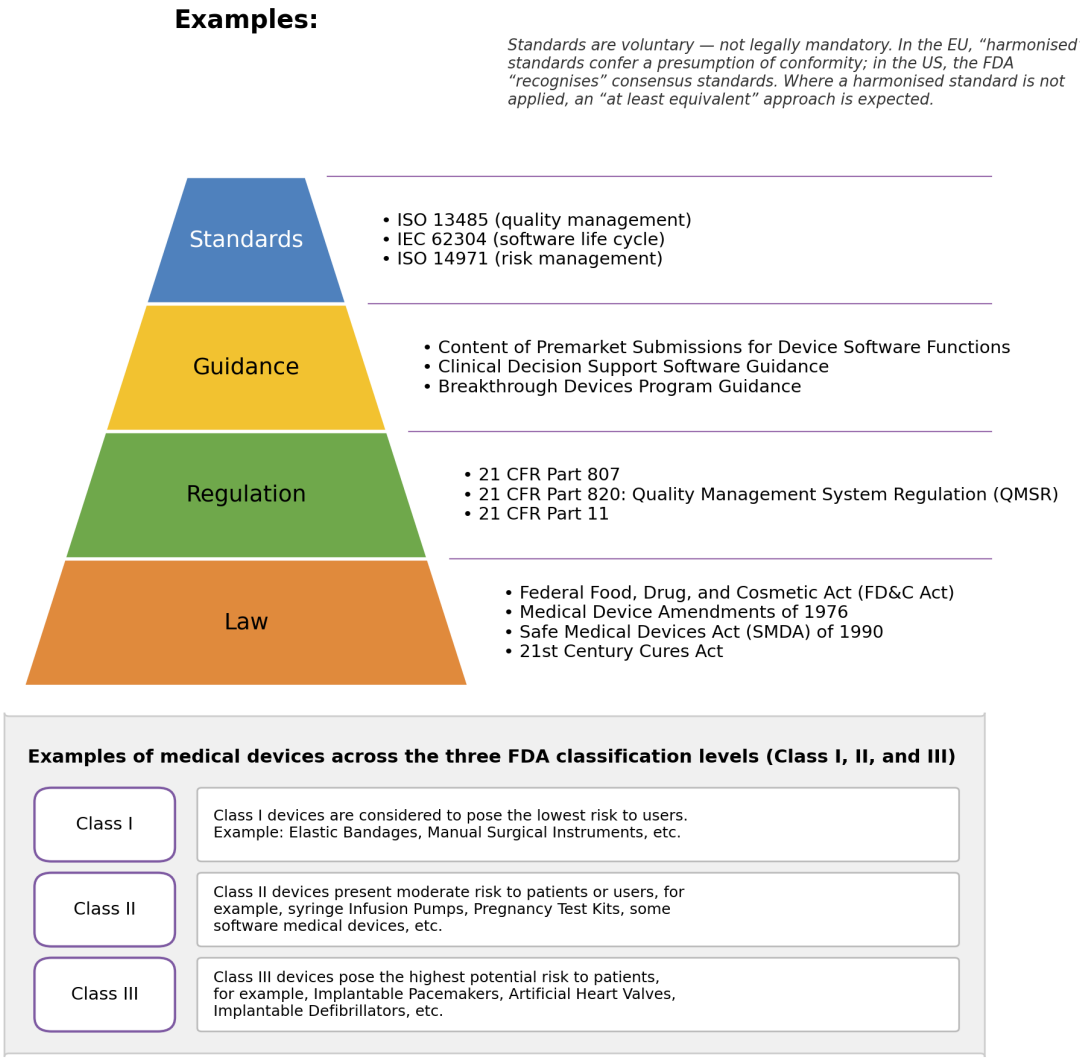


Figure 1.1: Hierarchy of Laws, Regulations, and Guidance, and Classification of Medical Devices in the USA. The pyramid indicates the hierarchical relationship, with laws at the base having the highest authority, followed by regulations and guidance documents. Standards (e.g. ISO 13485, IEC 62304, ISO 14971) are shown at the apex: they are voluntary rather than legally mandatory, but are "harmonised" in the EU (conferring a presumption of conformity) and "recognised" by the FDA in the US; where a harmonised standard is not applied, an "at least equivalent" approach is expected.

Key concepts in regulatory affairs (RA) are foundational to understanding the medical device approval process. The Total Product Life Cycle (TPLC) encompasses the entire lifespan of a medical device, from its initial design and development stages through commercialization, use, maintenance, and eventual obsolescence or disposal. This

comprehensive process involves key stages such as research and development, manufacturing, regulatory approval, marketing, distribution, and utilization by healthcare professionals and patients. Effective management of the TPLC is essential to ensure the safety, efficacy, and quality of medical devices throughout their lifecycle.

A submission dossier is a critical document submitted to regulatory authorities to seek approval for the legal marketing of a medical device. It contains detailed information about the device, including its materials, physical and chemical properties, and testing data. The dossier serves as evidence that the device meets safety and efficacy standards outlined in its Instructions for Use (IFU). The content varies by device classification; for Class III devices, clinical trials or evaluations are typically required.

The registration pathway is the structured process regulatory authorities provide to guide applicants through device approval. Each pathway is tailored to the specific requirements of different device classifications. For example, the U.S. FDA offers the 510(k) pathway for Class II devices and the Pre-Market Approval (PMA) pathway for Class III devices. In Japan, the PMDA uses the Ninsho pathway for Pre-Market Certification (PMC) of Class II devices and the Shonin pathway for Class III PMA. Similar pathways exist worldwide, providing manufacturers with clear steps and criteria for gaining market access.

At an international level, the International Medical Device Regulators Forum (IMDRF) fosters global harmonization of medical device regulations. Building on the Global Harmonization Task Force's (GHTF) legacy, the IMDRF includes regulatory authorities from nations such as Australia, Brazil, Canada, China, the EU, Japan, and the U.S. It serves as a collaborative platform for exchanging knowledge, sharing best practices, and working toward a unified global regulatory framework for medical devices.

To clarify these regulatory frameworks and their operational dynamics, Figure 1.2 presents a high-level roadmap that collectively represents key steps, stakeholders, and the relationships across the regulatory ecosystem. Additionally, the development phases of medical devices are systematically mapped out, illustrating the sequential stages from design and preclinical testing to clinical trials, regulatory approval, market launch, and post-market monitoring. Throughout the lifecycle of a medical device, stakeholders—particularly manufacturers and regulatory authorities—are engaged in continuous communication at each stage. Results from every phase are analysed to determine their compliance with the relevant regulations, and any gaps identified feed back into earlier stages, creating a dynamic feedback loop in which the outcome of one

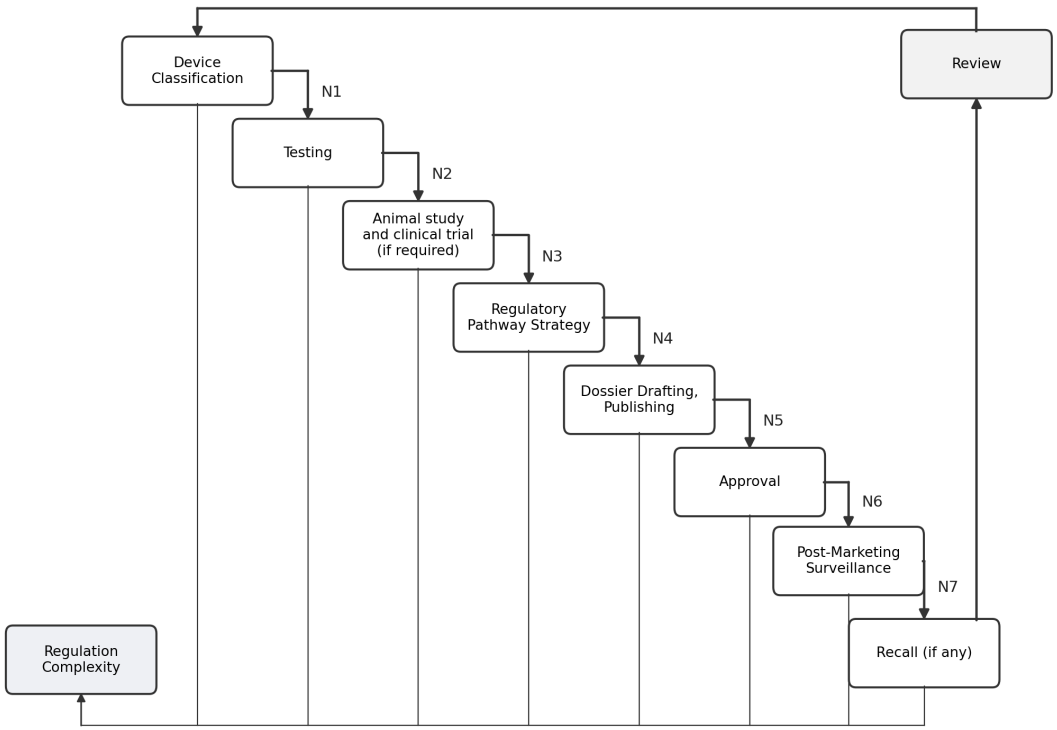


Figure 1.2: High-level schematic of the medical device regulatory lifecycle phases (N1–N7) that contribute to regulatory complexity. The figure highlights key stages from device classification to post-marketing surveillance, and identifies decision nodes in medical device lifecycle. Non-clinical *testing* precedes any *clinical trial*, since appropriate non-clinical test methods must be demonstrated before first human exposure.

phase informs the next. Furthermore, examples of medical devices are provided across different risk classifications to provide contextual grounding, as illustrated in Figure 1.1.

1.2 Motivation

According to the Oxford Dictionary, "complexity" refers to the state of being difficult to understand or intricate, involving many different and connected parts [2]. The complexity of regulations arises from the way they are designed, stated, and applied to the entities they regulate. This can cause difficulties in translating intention into execution [3]. The complexity of laws and regulations makes it challenging for people from different backgrounds to understand and communicate, leading to different interpretations. The issue of complexity in legal texts is significant because vague guidance can lead to confusion and misinterpretation by industries [4]. Regulations should provide guidance to all stakeholders in a clear and understandable framework.

Regulations serve as a critical platform for facilitating dialogue and development among various stakeholders, while also implementing legal principles and bringing legislation into effect. However, the efficacy of regulations has often been questioned, particularly in the context of their ability to mitigate systemic risks. For instance, insufficient regulatory frameworks were cited as a contributing factor to the 2008 financial crisis, or at the very least, as failing to adequately address its consequences [5]. The complexity of regulations has been widely debated and researched across fields, including general law [6], clinical settings [7–9], business management [10], financial regulation [11], and tax law [12]. While complexity in medical device design has been studied extensively [13], there remains a notable scarcity of research focusing on regulatory complexity in the field of medical device affairs. The goal of regulatory affairs is to ensure the quality and safety of medical products in a cost-effective manner, rather than imposing undue burdens solely for the sake of compliance. Nevertheless, ongoing debates persist regarding the effectiveness of current regulatory frameworks in both the European Union and the United States [14].

Medical device regulation is a set of laws and regulations that govern the listing and marketing process of medical devices in accordance with regional laws and regulations. The regulatory authority serves as a central point of control in the medical device industry, interacting with manufacturers to ensure that their products meet the required standards. Regulations ensure that products in the market are safe and effective, and manufacturers strive to produce products that meet the requirements set by the authorities. Regulations also serve as a call to action for producers, similar to a five-year plan or an industry plan. However, the complexity and variety of government regulations make it a critical task to understand them. In the medical device field, poor understanding or interpretation of regulatory requirements can lead to costly redesigns or repeated testing, which has been shown to hinder innovation and discourage technological development.[15]

Regulations serve as a platform for different entities to engage in dialogue and development. Their role is indispensable in implementing legal principles and bringing legislation into effect. Regulations have the potential to either support or hinder development. Proper implementation of regulations can aid in mitigating risks and ensure the smooth functioning of the system. Nonetheless, improper implementation may result in delays or even a complete halt to development[15].

Regulations also serve as a formal mechanism that structures the interaction between manufacturers and regulatory authorities. Regulatory agencies issue regulations to

safeguard patient safety, ensure product effectiveness, and maintain a well-functioning market environment. Manufacturers, in turn, must comply with these regulatory requirements to obtain approval and legally market their products. In the medical device industry, regulatory frameworks operate as a double-edged sword. On one hand, they ensure patient safety and device efficacy by regulating market entry and circulation. On the other hand, compliance with these regulations often imposes significant financial and operational burdens on manufacturers. In some cases, registration fees, clinical trials, and testing expenses exceed the revenue generated, leading to delisting or the withdrawal of devices from the market. This outcome not only limits innovation but also reduces access to potentially life-saving technologies for patients. For manufacturers targeting global markets, the problem is further compounded by redundant trials and compliance requirements across different jurisdictions. While organizations such as the International Medical Device Regulators Forum (IMDRF) aim to harmonize regulatory processes globally, inconsistencies persist. This underscores the urgent need for systematic evaluation of regulatory frameworks. This thesis proposes a framework for assessing regulatory complexity across the entire medical device lifecycle, aiming to enhance efficiency, reduce redundancy, and support innovation while maintaining the highest standards of safety and efficacy.

Navigating these regulations can be challenging due to their complexity, varied jurisdictional requirements in different countries, and fast evolving nature nowadays. Despite the critical importance of RA in the medical device industry, there is a notable gap in research that evaluates the complexity of regulatory processes and their impact on innovation. Manufacturers often struggle with vague or ambiguous guidance, inconsistent interpretations across jurisdictions, and the burdensome need to conduct redundant trials for global compliance. These challenges are exacerbated by the lack of a systematic framework for assessing and mitigating regulatory complexity. Previous research highlights the need for deeper investigations into these issues, especially within the context of readability, legal frameworks, and process efficiency in RA [15]. However, current studies have primarily focused on linguistic complexity within other legal domains, leaving the specific challenges in RA for medical devices largely unaddressed.

There have been a lot of debates surrounding the effectiveness of regulatory frameworks to ensure devices' performance, safety, and quality in the European Union and the United States [14].The complexity of medical device regulation has been researched and proven to be increasingly difficult to read over the years [15]. This thesis aims to unravel the intricate dynamics of regulatory complexity and propose actionable

solutions to bridge the gap between regulation and innovation, ultimately fostering a more effective and adaptive regulatory ecosystem.

Regulation will affect medical device innovation

While the medical device industry involves a broad range of activities, two domains are particularly central to discussions of regulatory science: the research and development of devices and the regulatory processes required for registration and compliance [16]. These domains are intricately intertwined and mutually reinforce each other in their functions. The journey of a medical device from its conceptualization (in the laboratory) to its commercialization involves multiple stages that necessitate collaborative efforts from professionals across diverse fields [17] as shown in Figure 1.2. Regulations serve as a vital framework that connects the various components of the medical device development process, and obtaining the final approval license represents a major milestone for all stakeholders, including innovators. Key activities—such as design, testing, and effective communication with regulatory bodies—are integral parts of this pathway. While regulations safeguard patient safety and device efficacy, they also impose substantial compliance costs on manufacturers. In some cases, the expenses associated with registration, clinical trials, and performance testing may even exceed the revenue generated from sales, potentially leading to the discontinuation or withdrawal of devices from the market [18]. This situation can significantly reduce the accessibility of these valuable devices to patients who may benefit from them. Particularly for manufacturers targeting the global market, the need to comply with varying regulations across different countries can result in duplicated trials or testing efforts. This happens because each regulatory authority may mandate its own formats, sample sizes, testing protocols, or validation criteria. As a result, clinical or performance data generated for one jurisdiction (e.g., the EU) may not satisfy the evidence requirements of another jurisdiction (e.g., the U.S. FDA), forcing manufacturers to repeat similar studies. Organizations such as the International Medical Device Regulators Forum (IMDRF) serve as mediators to facilitate communication and promote harmonization of regulatory processes across countries, aiming to enhance consistency and streamline regulatory procedures [19].

As noted earlier, while regulatory complexity has been extensively studied in other domains, there remains a significant gap in research pertaining to the assessment of medical device regulation, indicating the need for further investigation and evaluation in this domain. Despite being a significant sector within the healthcare industry,

medical device regulation has not been comprehensively examined from a complexity perspective [20]. Instead, much of the existing research is characterized by an ad hoc nature and lacks a theory-based approach aimed at identifying and addressing potential gaps and inefficiencies in order to enhance the overall regulatory process. The complexity and readability of regulatory documents, which serve as formal guidelines issued by regulatory authorities, have not been subjected to systematic evaluation using an appropriate framework or extensive discourse to ascertain their effectiveness in guiding manufacturers through the regulatory process. Furthermore, the absence of an objective metric for evaluating the impact of regulations limits the ability to assess the practical implications and efficacy of regulatory measures. Therefore, the development of a comprehensive framework for evaluating the quality of regulations is necessary. The proposed framework would assist manufacturers in assessing the anticipated regulatory difficulties involved in obtaining market approval for their devices. Additionally, it would offer regulators a means of self-reflection before developing and issuing regulatory documents. Furthermore, this framework would foster productive engagement between governments and manufacturers by offering an objective measure for evaluating regulatory practices. In the realm of global registration, the framework could enhance communication and cooperation among regulatory authorities in different geographical regions.

1.3 Research Problem

The issue of complex regulation in the medical device industry has garnered increasing attention in recent years. Manufacturers frequently encounter challenges when attempting to navigate the regulatory landscape and secure approval for their products. Unclear or ambiguous regulations are often cited as major obstacles to market entry. For example, the amount of MDR success approval is far lower than the number of application in recent years according to Notified Body Survey on Certificates and Applications [21] and serious review lagging problem [22]. This challenge is further compounded by the lack of research in this field, highlighting the need for new methods and tools to help manufacturers navigate the regulatory process with greater ease and efficiency. Addressing this challenge, this thesis aims to investigate how engineering methodologies can be utilized to assess the complexity of medical device regulation and how these findings can be leveraged to develop scientific tools that mitigate barriers to market entry. By establishing a more objective metric for evaluating regulatory

complexity and identifying specific areas for enhancement, this research has the potential to enhance the efficiency and efficacy of medical device regulation, benefiting both manufacturers and patients alike.

The primary research focus of this thesis is formulated as follows: *How can information engineering methods be used to assist medical device manufacturers and regulators to get devices to the market more easily? Framing the question around both manufacturers and regulators reflects the fact that the automated classification and analysis methods developed in the later chapters are of direct use to regulators as well as manufacturers, and affords the research question greater objectivity.* To address this main question, we focus on three interconnected objectives: (i) Exploration of the regulatory environment, (ii) Analyzing the origins and sources of complexity, and (iii) Development of engineering tools to reduce complexity for manufacturers.

In this dissertation, we address the following research questions.

RQ1: What characterizes the real-world regulatory environment? This question sets the stage by exploring the landscape of medical device registration and systematically reviewing current methods for evaluating legal texts in regulatory affairs. *Addressed in Chapters 2, 3, and 4.*

RQ2: What constitutes complexity in the field of medical device regulatory affairs, and can existing technologies effectively address downstream RA tasks? This question investigates the origins and sources of regulatory complexity through qualitative methods, and evaluates whether emerging technologies such as large language models can assist in regulatory tasks. The findings reveal that while LLMs show promise, their accuracy limitations necessitate the development of novel approaches. *Addressed in Chapters 5 and 6.*

RQ3: How can novel engineering tools be developed to reduce regulatory complexity for manufacturers? Building on the limitations identified in RQ2, this question focuses on developing and evaluating AI-based tools for downstream regulatory affairs tasks. This includes comparative analysis of traditional algorithms and LLMs, as well as the design of novel multi-modal AI frameworks for medical device classification. *Addressed in Chapters 7 and 8.*

These three research questions represent the core objectives of this dissertation, forming a progressive framework: RQ1 establishes foundational understanding of the regulatory landscape; RQ2 identifies complexity sources and evaluates existing technological solutions; and RQ3 advances novel engineering tools to address the

gaps identified. Together, they contribute to creating a more efficient, adaptive, and innovation-driven regulatory ecosystem.

The thesis presents novel recommendations targeted at key stakeholders, including regulatory authorities, manufacturers, and healthcare providers. These recommendations focus on leveraging engineering tools for regulatory affairs tasks, for example automatic device classification and the application of large language models in the medical device field. Indicators of success are broadly defined, including reduced submission errors, improved approval timelines, and enhanced post-market surveillance outcomes. This thesis contributes to both the academic literature and industry practice: theoretically, it advances the understanding of regulatory complexity and the application of engineering methodologies to compliance challenges; practically, it proposes implementable solutions that reduce human error, streamline workflows, reduce redundancies, and foster an innovation-friendly regulatory ecosystem.

The structure of this dissertation is organized as follows:

Part I: Understanding the Regulatory Landscape (RQ1)

- **Chapter 2** provides a preliminary overview of the regulatory framework for AI-based software medical devices under China's NMPA, serving as a pilot study to highlight key regulatory challenges.
- **Chapter 3** applies exploratory data analysis (EDA) to the NMPA's Unique Device Identifier (UDI) database, offering empirical insights into regulatory trends.
- **Chapter 4** presents a systematic literature review of existing metrics for evaluating legal and regulatory texts, identifying critical gaps in measuring regulatory complexity.

Part II: Identifying Complexity and Evaluating Existing Tools (RQ2)

- **Chapter 5** employs qualitative semi-structured interviews to uncover real-world challenges faced by industry stakeholders, mapping the origins and sources of regulatory complexity.
- **Chapter 6** evaluates the potential of large language models (LLMs) for medical device classification tasks. While LLMs demonstrate moderate accuracy and enhanced interpretability, their limitations motivate the development of novel approaches.

Part III: Developing Novel Engineering Tools (RQ3)

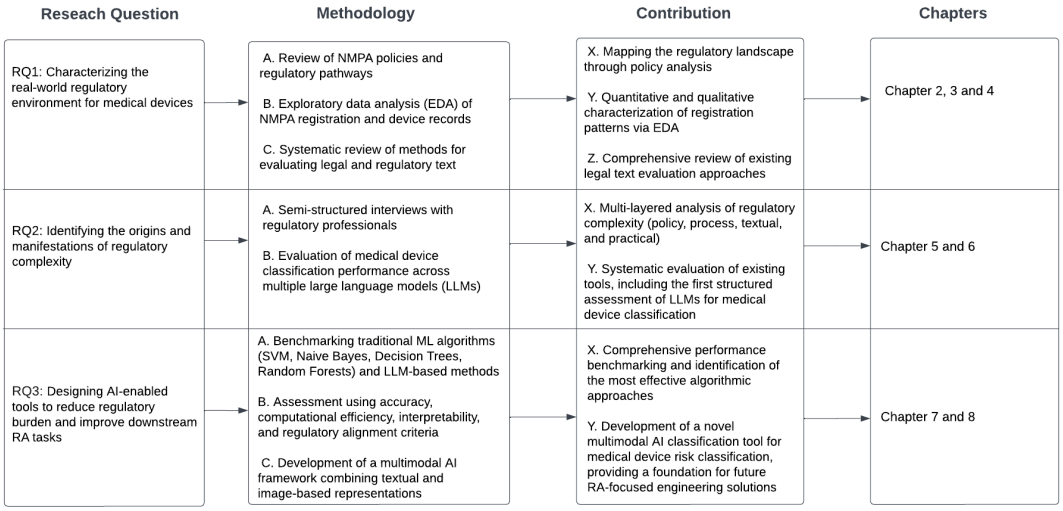


Figure 1.3: Research question, Methodology and Contributions of Dissertation

- **Chapter 7** conducts a comprehensive requirements engineering analysis, comparing traditional machine learning algorithms with LLMs for regulatory classification tasks, focusing on accuracy, computational efficiency, and interpretability.
- **Chapter 8** presents a novel multi-modal AI framework that integrates textual and visual information for automatic regulatory risk classification of medical devices.

Chapter 9 synthesizes the findings, discusses research contributions, and outlines future directions including multi-agent LLM frameworks for regulator-manufacturer modeling and cross-jurisdictional reasoning tools. The mapping between research questions, methodologies, and contributions is illustrated in Figure 1.3.

Building on the overarching research questions outlined above, the next Chapter 2 presents a preliminary exploration into the regulatory environment surrounding software-based medical devices. As a foundational step toward understanding regulatory complexity (RQ1), we chose to focus this pilot study on China—a country with a vast population, an expanding healthcare technology market, and growing global influence. Despite these factors, there remains a significant gap in scholarly research dedicated to China's medical device regulatory affairs, particularly concerning software and AI-enabled products.

In contrast, regulatory frameworks in the United States and the European Union have been widely studied and are often treated as reference points for international best practices. However, the Chinese regulatory system is evolving rapidly and presents

unique characteristics, institutional arrangements, and policy priorities that are underrepresented in existing literature. By examining software regulations in China, the research provides grounded insights into a complex, understudied jurisdiction. These insights serve as a contextual anchor for the later analysis in this dissertation, particularly the examination of regulatory complexity in RA (RQ2) and the evaluation of algorithmic tools for compliance and the development of data-driven models for regulatory support (RQ3).

2

A Preliminary Overview for Software Medical Devices

This chapter is based on: Yu Han; Aaron Ceross; Jeroen Bergmann, Regulatory Frameworks for AI-Enabled Medical Device Software in China: Comparative Analysis and Review of Implications for Global Manufacturers, 2024, Journal of Medical Internet Research (JMIR AI), published [23]

Author contributions (CRediT): Yu Han — Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft; Aaron Ceross and Jeroen Bergmann — Conceptualization, Writing – review & editing, Supervision. (The published JMIR AI *article is a Viewpoint and carries no formal contributions section; the roles stated here are the candidate's own.)*

2.1 Introduction

As a first step toward addressing RQ1, we conducted a study on the regulatory environment for software-based medical devices in China. Software regulations have been selected as the focus of this pilot study, aiming to explore the regulatory landscape and environment. While extensive research has been conducted on regulations in various contexts, there is a notable lack of literature specifically addressing software regulations in China, despite its status as a rapidly growing market. This study aims to fill this research gap by providing insights into the unique aspects of software regulations and the regulatory environment in China. By examining this specific domain, valuable knowledge and understanding can be gained to inform future research and contribute to the broader field of regulatory studies.

New software solutions that are being developed, especially medical devices that combine artificial intelligence (AI) and machine learning (ML), show a huge potential for patient benefit. These kind of applications can be used across different medical conditions, with the potential for easy scale-up to larger populations. It can reduce the burden on health care professionals and decrease the possible risk of missing vital information. For example, in radiology software is used to screen and diagnose large amounts of X-ray images [24]. A combined AI and ML approach can also be applied in e.g. oncology for the next-generation of sequencing [25], in ophthalmology for image recognition [26] or as support system for general medical decision-making [27]. ML models have been used from anything such as improving outcomes for diabetic patients [28] to tuberculosis diagnosis [29]. Many of these approaches should be applicable at a global scale and thus there is a growing interest to apply these solutions across borders. This has led to clinicians, academics and manufacturers to look at China and its medical device regulatory environment. However, China's regulations are deemed complex, due to language barriers, geographical distance and a general unfamiliarity with the regulatory system. These issues are augmented for innovative products that can have unconventional regulatory requirements. Reducing these barriers will allow for solutions to more easily flow between countries and it will provide an opportunity to learn from one another. This paper will provide a regulatory view of China, a large and rapidly growing market for medical device software, and discusses the implications for global manufacturers.

2.2 China Artificial Intelligence Development Plan

The 21st century has seen a rapid development of the Chinese economy and its ability to produce, manufacture, and distribute technology. In 2017, China State Council published a white paper discussing a *New Generation Artificial Intelligence (AI) Development Plan* [30]. The document indicated that the number of AI scientific papers published and invention patents granted in China ranked second worldwide. Several domain-specific applications that were developed in China have gained widespread attention, including intelligent monitoring, bio-metric recognition, industrial robots, service robots, and unmanned driving. The AI Development Plan clearly states China's support for smart medical care, products and services that utilize AI. Moreover, it is stated that this even should be developed as a priority. The vision is to establish a major medical system that leverages AI and ML.

China has become a major global producer and consumer of medical devices [31]. With the world's largest population (1.426 billion in 2022) [32], the need is obvious in terms of access to medical technology. In 2019, the Chinese medical device market had an estimated revenue value of ¥629 billion RMB ($88.7 billion USD), which more than double of what it was in 2015 (¥308 billion RMB) [33]. This coincides with a growing trend of medical device software (MDSW) registrations [34]. One factor driving this trend is the potential that digital health offers in terms of ease of scalability, which provides an opportunity to advance healthcare more sustainably. Those global manufacturers that want to sell into China need to understand the local medical device software environment. This requires an understanding of the registration requirements, authority supervision, differences with authorities in other regions and the current device mapping in China. Researchers and clinicians that need to stay on top of the latest developments by being aware of what is going on in China. The pandemic has shown clearly the need to understand policies and regulations in other countries (such as China) in order to effectively address a global crisis. Software based solutions can benefit from a similar global viewpoint.

2.3 Chinese Regulation on Medical Device Software

After *The New Generation Artificial Intelligence (AI) Development Plan* was introduced, China's medical products regulatory authority - National Medical Products Administration (NMPA) - released many regulations to fit the plan's theme. In 2022, the NMPA launched a program on digital health. Two medical device software guidelines were published as part of this program. Table 2.1 shows a series of regulatory documents published with regards to medical device software and AI enabled software. The NMPA released the first document in 2015, while a more up to date document was made public in 2022. This updated version raised more detailed requirements on the whole life cycle management of these technologies, as well as for the quality management, verification, raw code analysis and safety management.

Several standards are referenced in the regulation and they include (but are not limited to) standards on the the risk level of software (YY/T0664-2008), on the software engineering (GB/T 19003-2008) and those that describe the medical device quality management requirements (YY/T 0287-2003). These standards can help with the compliance to these new regulations and this provide a useful function in the regulatory pathway.

Date of Publication	Regulatory Document
08/2015	Guidelines of medical device software registration and review [35]
07/2019	Key Points of Deep Learning Decision-making Assisting Medical Device Software review [36]
07/2021	Guidelines for the classification and designation of artificial intelligence medical software [37]
03/2022	Guidelines of medical device software registration and review[38]
08/2022	Guidelines for the classification and designation of artificial intelligence medical software [39]

Table 2.1: Chinese regulatory documents for medical device software

In China, Medical device software (MDSW) includes "Software as Medical Device" (SaMD) and "Software in a medical device" (SIMD). The term "Software as a Medical Device" is defined by the International Medical Device Regulators Forum (IMDRF) as software intended to be used for one or more medical purposes that perform these purposes without being part of a hardware medical device [40]. Essentially this means that the software itself is a medical device. "Software in a medical device" means that the software works as a component of the whole device. For example, it helps to run a MRI scanner, X-Ray machine, or insulin pump. In these cases the software and other components all fall under the same registration license. In China, software that uses AI or ML can be present in both the SaMD and SIMD categories.

Regulatory environment in China

The CMDE (Centre for Medical Device Evaluation), which falls under the NMPA, is the agency that regulates medical devices in China. NMPA's regulatory framework for medical devices is enacted upon through Chinese laws, regulations and guidance. This framework includes all the elements that are regulated for market access, including the device designation, device classification, registration review dossier content and post-market surveillance. Manufacturers will need to interact with the NMPA on all these aspects.

Medical devices are regulated via a risk management framework, that classifies products as low risk (class I) to high risk (class III). For manufacturers that are importing into China, national authorities are responsible for the reviewing process, whilst some domestic devices are regulated by provincial authorities. To classify your device in

China, you will need to match your device description with information that is held in the Medical device catalog [41].

Manufacturers can design innovative equipment using two main approaches, which sometimes can be combined. One is starting purely with the patient needs (need-led innovation) to create a "new" technology, while the other is developing a "new" technology for which a subsequent patient need will have to be identified [42]. These innovations can occur either before or after appropriate regulations have been set [43]. It is common that transformative ideas initially do not have suitable regulations in place and that this mismatch can lead to either delays in market adoption or concerns in terms of device performance and safety. However, any medical software company that wants to bring their product to the market must comply with the current regulations. A clear understanding of the classification and regulatory requirements for a particular market is therefore crucial, as innovation and regulation are highly interactive. A good understanding is particularly important, as it has been suggested that the complexity of medical device regulations can increase whenever new regulations are formed [15]. If manufacturers misidentify the risk classification of their product and the related regulatory requirements, they will end up spending more time and money to correct for this. Regulations should thus be considered as soon as possible, as many decisions with regards to the final product are already made at the early stages of the research and development process.

Specific rules for software and AI

Medical software is basically divided into auxiliary diagnosis and treatment devices according to the intended use, and the overall categorisation of medical device software is summarised in Figure 2.1. A detailed translation of the software catalog can be found in Appendix 1. The SaMD (which begins with code 21) are categorized into six categories: Treatment Planning Software (21-01), Image Processing Software (21-02), Data Processing Software (21-03), Decision Support Software (21-04), In Vitro Diagnostic Software (21-05), and other software (21-06). If the device to be registered is not included in the list, then it has to be re-classified through the device designation pathway [44]. A simple flowchart for classification of the software is shown in Figure 2.2.

There are two branches for SaMD, which are split between AI and other technologies. If AI is applied, then a further decision is made according to the level of maturity of the algorithm. A high maturity level of the algorithm means the algorithm's safety and effectiveness profile has been well established, while a low maturity level would

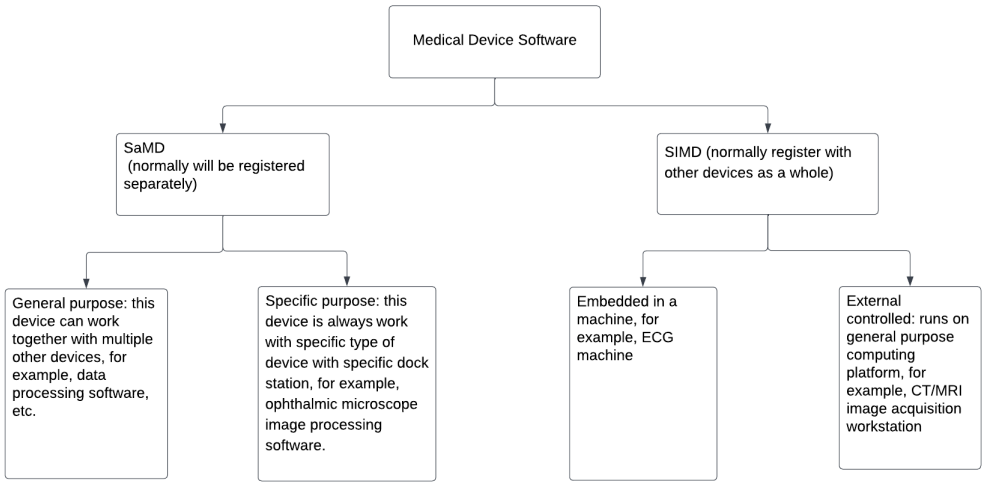


Figure 2.1: Categories of medical device software

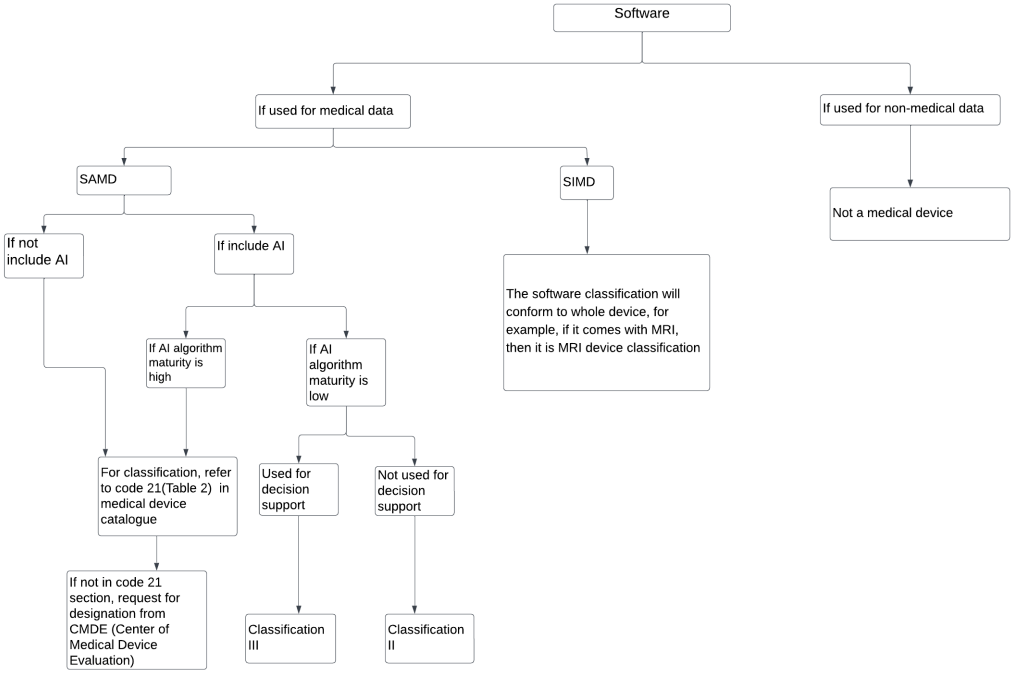


Figure 2.2: Medical device software classification flowchart

indicate that this is not the case. A pre-consultation meeting could be used to discuss the maturity level with the NMPA. If the AI algorithm has a well established profile then manufacturers can refer back to Appendix 1, code 21 [41] for classification. A request for designation could then be sent to the NMPA, if the device is out of scope. If the maturity is low, then there are two classifications possible. The device could be classified as a class III device if it is used for decision support, otherwise the device will be assigned a classification of II, which represents a lower risk class. According

to the Medical Device Classification Catalogue [41], a class II device classification is given when the software does not contain any AI and the medical software is only used for image/data processing, thus not used for e.g. diagnostics. If it would be used for diagnostic purposes then the classification would become III. The degree of risk for diagnostic software is determined by the level of maturity, registration of the applied algorithm in their database, as well as the "object" of interest (referring to a particular disease, such as a certain type of cancer) [45].

However, if the software just provides diagnostic suggestions through its algorithm (in other words, it only has an auxiliary diagnostic function and does not directly provide a diagnostic conclusion) then the device can be regulated as a class II medical device. Yet, if the diagnostic software automatically recognises, for instance, a lesion site through its algorithm and provides clear diagnostic prompts, a class III classification would be assigned due to the increased relative risk. In general, medical software using AI technology is currently managed by designating it the highest possible classification in China. This is driven by the novelty of the technology, as well as the lack of in-depth and complete evaluations of the clinical risks. China has been focusing more on reviewing the algorithm itself, whilst in the USA the attention has shifted towards the manufacturers themselves [46].

It should be noted that not all software applied in the medical field is regulated as a medical device by the NMPA. If the software is used to process medical device data for measurement, model calculation or analysis, then it is deemed medical device software and thus regulated by NMPA. If the software is used for non-medical device data, it will not be regulated as a medical device under the NMPA. This is the case when software is used for processing of general patient information or for patient testing reports, which both are not seen as medical device data.

China's and the International Medical Device Regulators Forum (IMDRF) criteria share many similarities on how to determine if the software is a SaMD. According to the IMDRF [47], the SaMD definition should include a clear statement about the intended use of the device and the following aspects need to be described in order to be able to be regulated as SaMD;

- The "significance of the information provided by the SaMD to the healthcare decision", which is used to identify the intended medical purpose of the SaMD.
- The "state of the healthcare situation or condition" that the SaMD is intended for.

In alignment with the IMDRF, under the European Union's Medical Devices Regulation (MDR) only products whose intended use includes a medical purpose and influences the patient's healthcare situation can be deemed to be a medical device. Products such as, medical information management software (which is a hospital management tool) are also not designated as a medical device. This is similar to China, since it then doesn't meet the definition of a medical device.

2.4 Challenges Posed by Software and AI in Medical Devices

As AI technology develops further, regulators will also face the challenge of applying regulatory safeguards to these novel technologies. The technical complexity of certain medical software solutions warrant the description of these system as a "black box", due to their inherent opacity [48, 49]. In addition, traditional frameworks for regulation are not suitable for adaptive AI and ML technologies, since the algorithm is constantly learning and making changes [50]. Therefore, digital healthcare solutions provide a different set of challenges to the regulators and the traditional fixed regulatory framework is not suitable for this type of AI device. At present, governmental agencies in the USA, the European Union and China have all issued new regulatory methods or frameworks for medical device software to help cope with the changing landscape.

The regulation of AI devices is to ensure safety, quality and reliability requirements are met. One key concern is the "unlocked" nature of these devices. "Locked" devices means that the algorithm provides the exact same result for a (specific) given input [50]. This is in contrast to a "unlocked" algorithm, which represents a continuous learning algorithm. The "unlocked" algorithm is also known as an adaptive algorithm and it changes its behaviour using a pre-defined learning process, which provides time-based updates from new data with the overall aim to improve its clinical performance. This algorithm continuously changes the input-output relationship. Thus, for a given set of inputs, the output may be different before and after these changes are implemented. This means that after an "unlocked" device has been approved and given access to the market, the device can continue to self-learn and thus alter its performance in comparison to when it was first approved. In this situation, it is difficult for the clinicians or the authority to fully trust the device before they use it in practice. So far, the FDA has not yet approved a device which integrates continual learning AI, as they have only granted approval to locked systems [50].

The FDA has enacted the Digital Health Innovation Action Plan[51], with the aim to build a more dynamic approval process with pre-certification for companies that then will have the ability to change characteristics of a product without needing ongoing FDA assessment. This enterprise-based approach (pre-cert program) is very different from the traditional medical device regulation. The FDA adopted the pre-Cert program together with the Total Product Life Cycle (TPLC) database to screen for eligible organisations. They also adopted a "predetermined change control plan". This plan provides a complete approach based on total product lifecycle in a way that manages the risk to patients in a controlled manner.

The EU also enacted new directives to regulate this fast-changing technology domain. They include the General Data Protection Regulation (GDPR), Cybersecurity Directive, Medical Devices Regulation (MDR) and In Vitro Diagnostic Medical Device Regulation (IVDR). The GDPR and the Cybersecurity Directive have taken effect in May 2018, whilst the MDR was applied in May 2021 with IVDR following suit a year later. These recent changes further highlights the moving landscape of regulations at a global scale.

2.5 Future Directions

Since China joined the IMDRF in 2013 [52], China has adopted and referenced international regulatory methods when formulating their own regulations. Regulatory similarities between China and other countries have been witnessed and demonstrated. However, China also has its own local requirements, standards, and regulatory ideologies, which can be an additional layer of complexity for global manufacturers who want to bring their medical device to the Chinese market.

There are different aspects for global manufacturers to pay attention to, when they want to leverage US or EU experience for the Chinese market. In China the focus is more on the maturity of algorithm, which is different from the FDA sponsor qualification program. Differences in sample populations upon which the algorithm is build is another key consideration, in addition to requirement to ensure data confidentiality and protection of patients in a specific region. In the Regulatory Science Action Plan (RSAP) issued by NMPA in 2019 [53], there is a clear focus on AI, which suggest more regulations might be developed with an increased level of harmonisation with the US, EU or other markets. Nonetheless, regulatory inconsistency still exists between countries. The same device can be regulated very differently across borders, which poses global manufacturers with big challenges. Large well-founded medical device companies

usually have global regulatory affairs professionals that deal with this situation, but innovation may also arise from small research teams at universities or innovative small and medium enterprises. In this situation, the complexity of the regulatory environment will hinder the potential of influential new products to enter into market. The regulatory strategy will need to differ from region to region to provide be best possible match for each.

For example, in the USA a high-tech device could be registered as a class II device, as long as it is similar to a predicate device which already has been registered. In this case, the characteristics need to be the same and there shouldn't be any cause for concern with regards to the safety and effectiveness of the device. However in China, manufacturers will need to refer to the classification catalogue, which aims to classify the device based on its own safety and effectiveness. If it is a high tech device, then it becomes more likely that it seen as a class III device in China. This means that the device will face more stringent registration requirements, including clinical evaluation and even trials. Manufacturers need to consider this when they start to map their market potential globally, as it could become a regulatory barrier for them.

Strategically, some manufactures would choose to register their device first in USA and then explore China or other markets. The USA regulation is also focused on the sponsor criteria and "Current Good Manufacturing Practice" alongside the assessment of the software algorithm itself, which makes it more organisation-centric [54]. Another registration strategy could be to register half of the medical devices that are in the development stage (also called "pipeline products") in the USA and the other half in China. After getting feedback from both authorities, they can switch them over. In the USA, applicants of new devices can go through the De Novo premarket pathway or Breakthrough Device designation to register their technology [54]. In China, there exists a "Green Channel" for software with urgent medical needs.

It is critical that international manufactures and authorities work together to determine the best possible regulatory pathway for each AI/ML product. It is therefore essential for everyone to understand the differences between regions in order to create an environment in which all stakeholders can learn from one another. Increasing the understanding of global regulations will create new opportunities in healthcare, which in the long-term will bring more patient benefits.

Among these factors, the regulatory approval process—often centered on registration—is particularly crucial, as it determines whether a product can enter a given market. Importantly, such registration processes rely on formal legal texts and requirements. For

manufacturers, understanding and interpreting these legal documents is a necessary step toward successful market access. However, the linguistic complexity and structural variability of these regulatory texts can pose significant barriers, potentially affecting how efficiently companies can comply with requirements and bring their products to market.

Having established the foundational understanding of China's software medical device regulatory environment in Chapter 2, the next step is to examine how these regulatory dynamics manifest in practice. Chapter 3 presents a data science analysis exploration based on a national regulatory database, aiming to uncover the scope, classification, and algorithmic characteristics of software-based medical devices in China.

3

Exploration of Software Medical Devices Using a Chinese Regulatory Database

This chapter is based on: Yu Han; Aaron Ceross; Sarim Ather; Jeroen Bergmann, Data-Driven Analysis of AI in Medical Device Software in China: Trends of Deep Learning and traditional AI Based on Regulatory Data, in preparation

Author contributions (CRediT): Yu Han led the work (Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Visualization, and Writing – original draft); the co-authors contributed to Conceptualization, Writing – review & editing, and Supervision.

3.1 China's New Generation Artificial Intelligence Development Plan and Effect to Medical Devices

Building on the conceptual overview in Chapter 2, this chapter provides a data-driven investigation into the landscape of software medical devices in China, with particular attention to the regulatory influence of national AI policy and the characteristics of algorithmic applications in submitted products.

While prior research has explored the deployment of AI-enabled medical devices (AIMD) in various regulatory landscapes, including notable work on regulatory frameworks and device categorization methods [55], our study introduces an innovative, semi-automated approach. By leveraging advanced data science techniques, our methodology filters targeted devices directly from the NMPA regulatory database. The automated filtering system enables rapid identification of candidate AIMD in under one minute, followed by a manual verification step to eliminate false positives arising from ambiguous

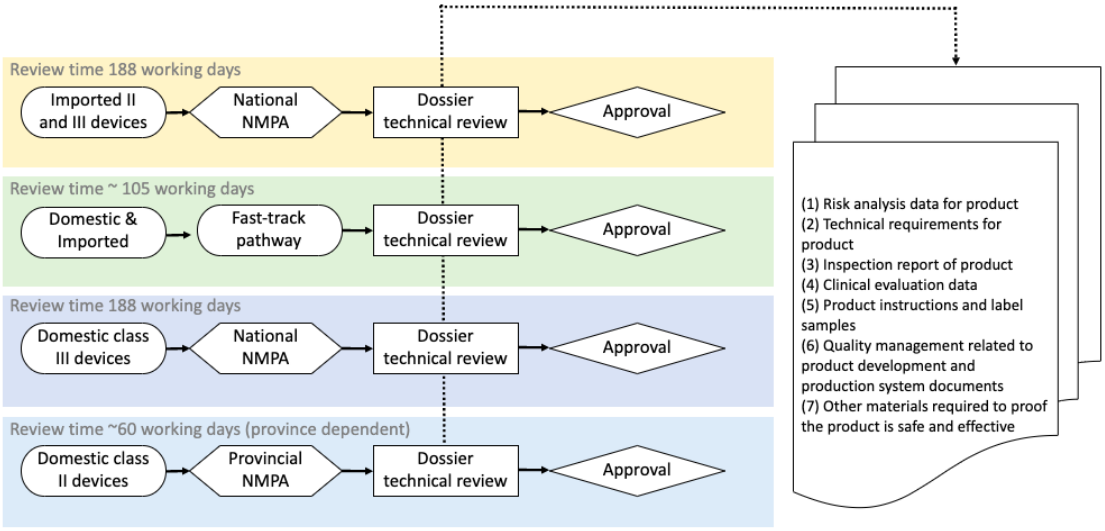


Figure 3.1: Medical device registration pathways set out by the NMPA. Imported and domestic devices relates to all class II and III devices, for Class I devices, only record-filing is required.

regulatory terminology. While not fully automated, this approach significantly enhances efficiency compared to previous methods that relied primarily on manual data collection.

China Software Medical Device Registration Process

Figure 3.1 provides an overview of the medical device registration process administered by the NMPA for various types of devices. The NMPA has also established expedited pathways, including the Innovation Approval [56], Priority Review [57], and Emergency Approval [58] processes. The final version of the "Innovation Approval Procedure for Medical Devices," published on 5 November 2018, outlines three criteria for fast-track approval: a) The applicant must have completed preliminary research on the product and developed a finalized prototype with comprehensive and traceable research data. b) The primary technical mechanism of the product must be novel, inventive, and hold significant clinical value, meaning that the product's performance or safety offers distinct advantages over existing alternatives. c) The applicant must either hold the patent for the core technology of the product in China or have obtained the right to use the patent in China through technology transfer [56]. Both domestic and foreign manufacturers are subject to the same application process for innovative device registration.

Unique Device Identification System and Database Background

The Unique Device Identification (UDI) system is a framework proposed by IMDRF and its intention is to adequately identify medical devices from manufacturing through distribution and patient use. Since joining the IMDRF in 2013 [52], China has been adopting and referencing international regulatory methods when formulating China regulations. China officially adopted the UDI System in 2019, as outlined by two key documents: (i) Pilot Project Plan for UDI System for Medical Devices (No.56, 2019) [59] and (ii) Rules for Unique Device Identification System (No. 66, 2019) [60]. The NMPA is implementing the adoption of UDI in phases. The strategy is to start with devices that carry the highest risk (Class III) and then proceed to add lower risk devices in each new phase. In January 2021, nine categories of high-risk Class III medical devices were included in the first batch. From June 2022, the second batch of medical devices was included. As of August 2024, there are 4,045,039 products registered with UDI system, with most being general medical devices (96.34%), whilst in vitro medical devices (3.66%) are only making up a small amount of the total registrations[61]. The NMPA's UDI database is publicly accessible, allowing patients, healthcare providers, and regulatory authorities to quickly access information about a particular medical device.

Figure 3.2 illustrates the meaning of the number codes in China's UDI system [62]. A UDI code typically consists of a Device Identifier (UDI-DI) and a Production Identifier (UDI-PI). The UDI-DI refers to a unique numeric or alphanumeric code that is specific to a particular model of medical device and is also used as an access key to information stored in the database. The UDI-PI, on the other hand, identifies the unit of device production with a numeric or alphanumeric code. In China, the UDI-DI is composed of two parts: part I is assigned by the code-issuing agency, while part II is issued by the manufacturer.

Research Aim and Contribution

This study aims to evaluate how real-world regulatory database can be used to generate insights into the landscape, classification, and governance of AI-enabled medical devices. While we introduce a rule-based identification method to support this analysis, our core objective is to apply data science in the service of regulatory science.

We are particularly interested in exploring how the Chinese regulatory authority (NMPA) classifies AI-enabled medical devices, and whether there are systematic patterns in device classification that correlate with technical and regulatory factors such as device type (SaMD vs. SiMD), algorithm category (e.g., deep learning vs. traditional AI), clinical

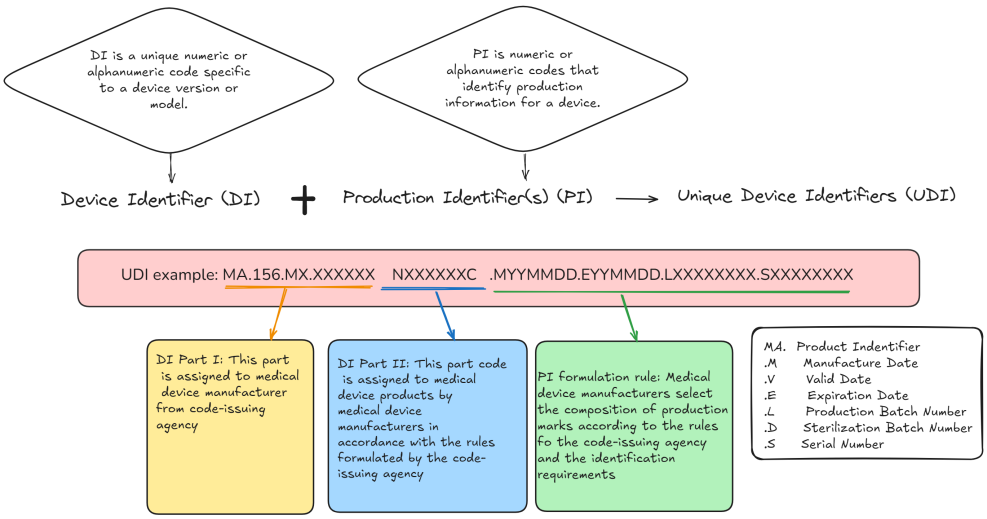


Figure 3.2: UDI code layout according to YY/T 1630-2018. Note: UDI-PI elements (e.g., version number [.V], manufacturing date [.M], expiration date [.E], lot [.L], sterilization batch [.D], and serial number [.S]) are optional and may not appear in all UDIs.

function (e.g., diagnosis, monitoring), and product origin (domestic vs. imported). To investigate this, we develop and apply a reproducible, rule-based method that integrates regulatory logic, product classification structures, and naming conventions. This method allows for efficient extraction, identification, and the analysis of AI-related submissions contained within a comprehensive national-level medical device registry.

To guide this investigation, we pose the following primary research question:

> *What can publicly available regulatory data reveal about the development, deployment, and oversight of AI-enabled medical devices in China?*

This question is further explored through the following sub-questions:

- *What types of medical device software (MDSW), including SaMD and SiMD, have been registered in China that involve AI components?*
- *Which domestic and international manufacturers are contributing to this domain?*
- *What kinds of AI techniques (e.g., deep learning, rule-based systems) are being used, and in what medical specialties or disease areas?*
- *How do observed trends correspond with regulatory policy changes, such as updated guidelines or new classifications?*
- *Explore how different AI algorithm types are currently represented in regulatory classifications, and whether observable patterns emerge?*
- *What methodological challenges arise when using registry data to analyze emerging AI-enabled devices?*

Our approach is rooted in a scalable exploratory data analysis (EDA) framework. We use the rule-based algorithm to classify AI-enabled medical devices, including SaMD and SiMD, and generate a set of structured observations from the UDI registry. This enables rapid identification of relevant entries at scale and supports systematic monitoring of evolving AI adoption patterns. However, the tool is a means rather than an end. Our central contribution lies in the use of this automated pipeline to empirically examine regulatory practice, track industrial behavior, and understand the emerging medical AI ecosystem through the lens of real-world registry data. In addition to the core objective of evaluating classification patterns using automated identification methods, we also include an exploratory analysis of algorithm types and observed regulatory variations, to support future hypothesis generation. These findings are descriptive only and are not intended to make normative or causal claims.

Our contributions are fourfold:

1. We develop and validate a scalable, reproducible method for extracting AI-related entries from a national regulatory database, providing a foundation for systematic surveillance of AI-enabled medical technologies.

2. We analyze descriptive patterns in regulatory classification, manufacturer distribution, algorithmic design, and clinical domains, offering empirical insights into how AI technologies are integrated into healthcare products.

3. We propose a conceptual framework for leveraging regulatory data to study oversight behavior, highlighting the potential for empirical approaches to inform broader digital health governance.

4. We demonstrate how real-world regulatory data can be used to infer the actual deployment of AI techniques within specific medical device contexts. This is particularly valuable given the limited transparency and explainability of many AI systems, and may inform future policy development and standards-setting.

3.2 Methods

Definition of MDSW, SiMD, SaMD and AI software

As defined in Chapter 2.5, Medical Device Software (MDSW) includes both Software as a Medical Device (SaMD) and Software in a Medical Device (SiMD).

According to NMPA Regulation Guiding Principles for Registration Review of Artificial Intelligence Medical Devices[63], an AI medical device referred to medical device that utilises artificial intelligence technology to achieve its intended use (i.e., medical purposes) by using "medical device data." Medical device data refers to objective data generated by medical devices for medical purposes. This includes medical imaging data generated by medical imaging devices (such as X-ray, CT, MRI, ultrasound, endoscope, optical imaging and so on), physiological parameter data generated by the medical devices (such as ECG, EEG, blood pressure and so on), and in vitro diagnostic data (such as pathological images, microscopic images, invasive blood glucose waveform data and so on). Therefore, an AI medical device refers to a medical device that utilises artificial intelligence technology to process and analyse medical (device) data for medical purposes.

Pre-processing

Description of database

All data analysis was conducted using RStudio (version 2022.12.0, Posit Software, PBC). The exploratory data analysis was performed on the `UDID_FULL_RELEASE_20240801.zip` dataset obtained from the NMPA website [64]. The dataset contains information on medical devices registered with the NMPA UDI system from 2016 to 2024, consisting of approximately 4 million products, each of product labelled with 48 variables, such as product description, version number, generic name, specification, product number, device classification and so on.

Search Strategy

The filtering process for identifying AI-enabled medical devices (AIMD) was conducted in a structured, multi-stage approach to systematically refine the dataset from the UDI database, which contains approximately 4 million medical devices, as shown in flowchart in Figure 3.3.

First, medical device software (MDSW) was extracted by isolating both Software as a Medical Device (SaMD) and Software in a Medical Device (SiMD). SaMD was identified using the six-digit classification code rule, where codes beginning with **"21-"** denote standalone software-based medical devices. SiMD, which lacks a specific classification code, was identified using a keyword-based filtering approach applied to product descriptions, ensuring relevance to medical applications. The keywords

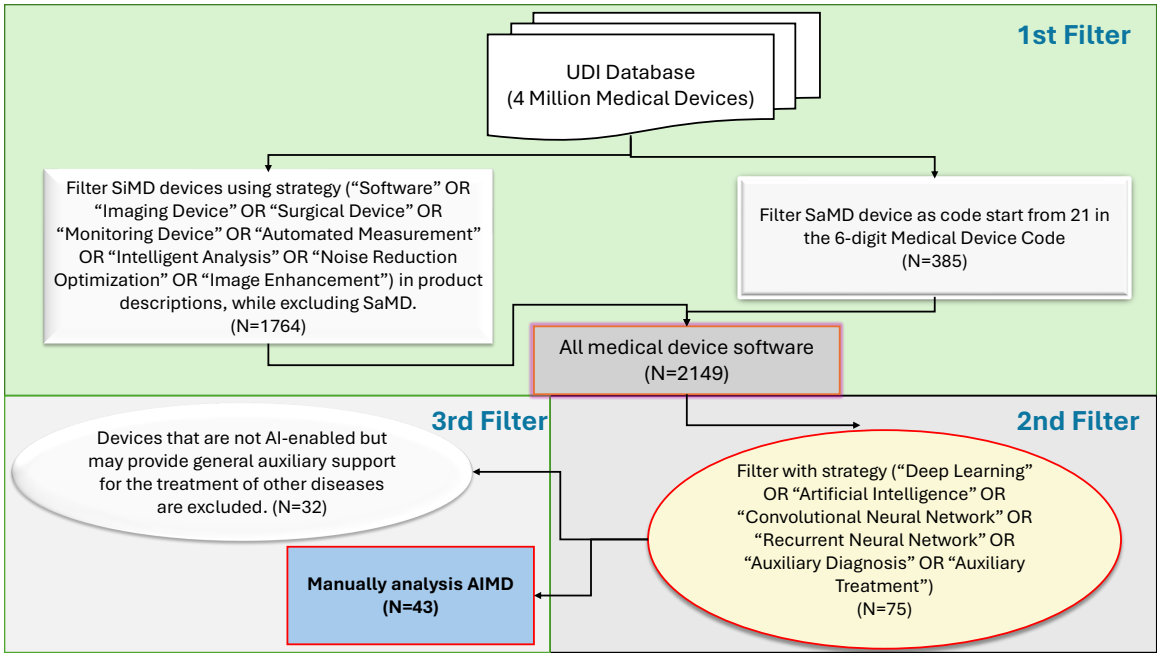


Figure 3.3: Flowchart for the selection process for the identification of software and AIMD

we adopted included **("software" OR "imaging device" OR "surgical device" OR "monitoring device" OR "automated measurement" OR "intelligent analysis" OR "noise reduction optimization" OR "image enhancement"**.Devices already categorized as SaMD (21-code rule) were explicitly excluded to prevent misclassification. This process resulted in a total of 2,149 MDSW devices.

To further identify AI-enabled medical devices, we applied a naming-based criterion. According to the Chinese regulation *Guiding principles for naming of medical software* [65], devices that incorporate artificial intelligence must include in their name either the terms (i) Auxiliary diagnosis or (ii) Auxiliary treatment. For example, as indicated in the NMPA regulation, if the radiotherapy contouring software uses non-artificial intelligence technology, it is named "radiotherapy contouring software"; if it uses artificial intelligence technology, it is named "radiotherapy contouring auxiliary decision-making software". So we adopt keywords as searching strategy to filter devices based on the results from first step: an additional filtering layer was applied to the MDSW dataset using strategy: **("Artificial Intelligence" OR "Machine learning" OR "Deep learning" OR "Reinforcement learning" OR "Auxiliary diagnosis" OR "Auxiliary treatment" OR "Convolutional Neural Network (CNN)" OR "Recurrent Neural Network (RNN)**. This automated step initially identified 75 candidate AIMD. However, due to regulatory wording, certain false positives emerged, particularly among devices labeled with "Auxiliary Diagnosis" or "Auxiliary Treatment," which, in some cases,

referred to general software support rather than AI-driven functionalities. To refine the dataset, a manual verification step was conducted, eliminating 32 devices that did not exhibit AI functionality, resulting in a final count of 43 AI-enabled medical devices. This structured filtering approach ensured both high precision and recall, effectively distinguishing genuinely AI-enabled medical devices from broader medical software applications.

3.3 Results

3.3.1 Devices amount and category

A total of 2,149 MDSW were identified, with 385 classified as SaMD and 1,764 classified as SiMD devices, as shown in Figure 3.4 . The majority of devices are domestic products, with the largest category being domestic SiMD Non-AI devices, comprising over 60% of the total of whole medical software. Imported SiMD non-AI devices constitute approximately 11% of the total. AI-enabled devices are predominantly concentrated in domestic SaMD, making up more than 90% of all AI devices identified. Notably, no SiMD device was found to incorporate AI.

In the filtered device list, we labeled each device register city and province, annotation for AI, non-AI, whether SaMD and SiMD and put into China map as shown in Figure 3.5, we can see in each province software registration amount in different province of china map.

The analysis of medical device software (MDSW) reveals distinct patterns in the distribution and classification of Software as a Medical Device (SaMD) and Software in a Medical Device (SiMD). The dataset indicates that SaMD is predominantly domestic, with the majority classified under Class II, suggesting that most standalone software applications in the medical field are moderate-risk solutions rather than high-risk interventions. This aligns with regulatory expectations, as standalone software is often used for decision support, diagnostics, or treatment recommendations, rather than direct intervention in critical medical procedures. In contrast, SiMD is more evenly distributed across domestic and imported categories, with a higher proportion classified as Class III, indicating its integration into high-risk medical equipment.

In our analysis, AI-enabled devices were predominantly identified within standalone software (SaMD), with very few detected among software integrated into hardware (SiMD). This outcome is not due to a limitation of the search strategy, which was carefully designed, transparently documented, and reproducible. Rather, it reflects the

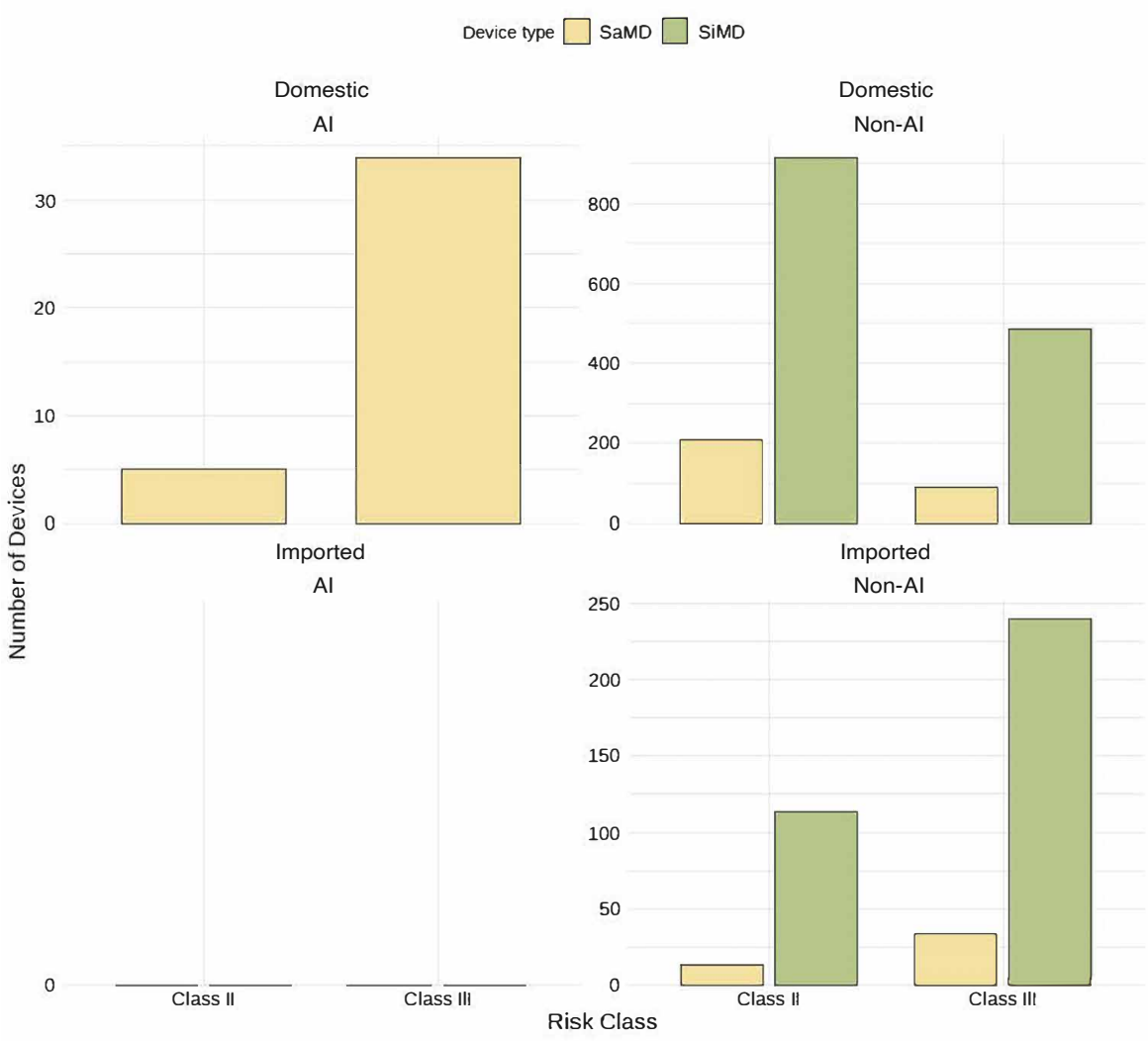


Figure 3.4: Distribution of domestic and imported SaMD and SiMD devices by risk class, stratified by AI and non-AI designation. Note: Due to substantial differences in device counts across categories (spanning more than an order of magnitude), each subplot uses an independently scaled y-axis to preserve visibility of within-category distributions. As a result, bar heights should be interpreted within, rather than across, subplots.

fact that many SiMD product entries in the UDI database lack explicit references to AI-related terms such as "machine learning" or "intelligent analysis." For example, well-known devices known to incorporate AI (e.g., deep learning reconstruction in imaging equipment) do not contain any such keywords in their official product descriptions. As such, even rigorous search methods are constrained by what is documented in the database. This finding highlights an important reality: the absence of identifiable AI in certain categories is not necessarily due to analytical shortcomings, but to how information is entered and structured in the registry. Future regulatory database enrichment efforts could help improve visibility and traceability of AI functionalities.

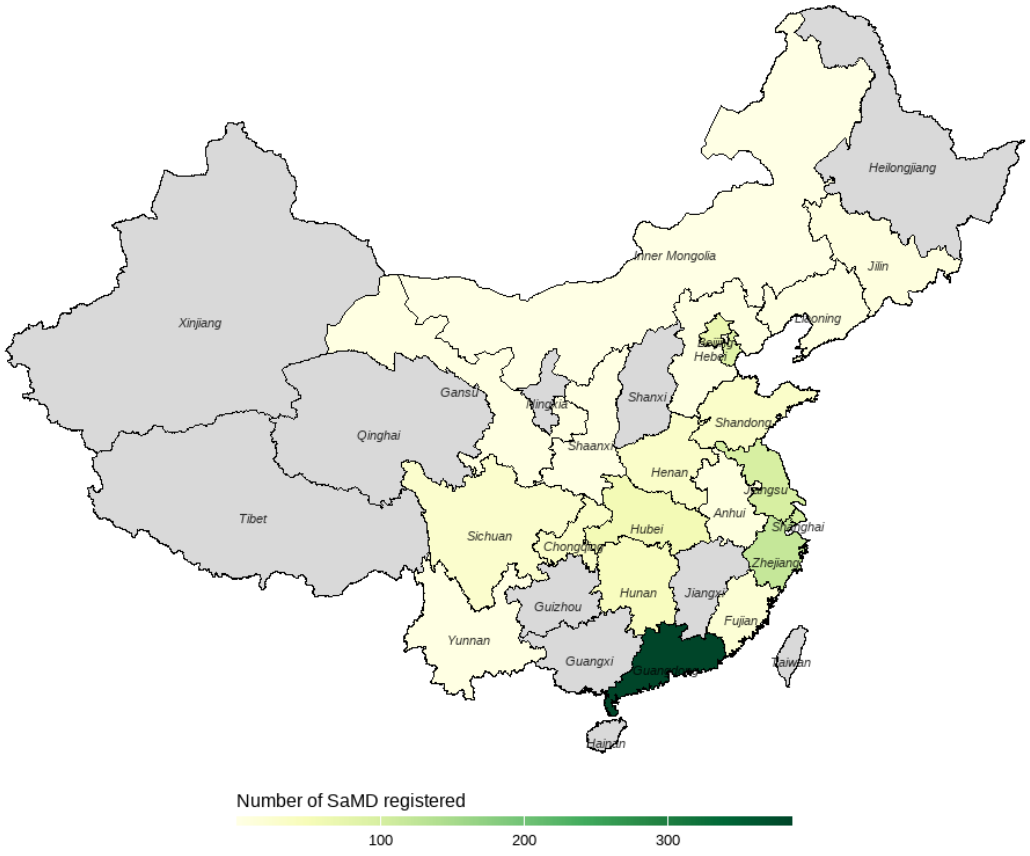


Figure 3.5: A geographic map of non-foreign medical device software are registration 2020–2024 (n = 2,149). The areas include mainland provinces as well as the indication of , which indicates products from the Special Administration Regions.

For AI devices, the majority of registrations are for domestic Class III devices, which constitute approximately 70% of the total. This figure includes manufacturers both within and outside of Beijing, as per regulatory requirements mandating that all Class III devices must be registered in Beijing (national level). For Class II devices, the numbers decline progressively, with Guangdong and Hubei each accounting for around 8% of the total registrations. Additional contributions come from Zhejiang (5%), Tianjin (3%), and Jiangsu (2%), with Shanxi and Sichuan provinces each representing 1% of the total registrations.

3.3.2 Artificial intelligence devices and algorithms in medical software

The authors cross-checked and validated the machine-filtered information, followed by a manual review to ensure accuracy. This process led to the development of a comprehensive database, which includes 43 AI-enabled devices, as shown in Figure 3.6. A detailed overview of these devices and their associated algorithms is provided in Appendix 2.

Our approach introduces greater granularity by incorporating the specific anatomical or disease-related focus of AI-enabled medical devices. This distinction is valuable for both manufacturers and patients, as it provides clearer insights into the availability of advanced technologies for particular medical conditions. A medical specialty refers to a distinct branch of medical practice that focuses on a specific group of diseases, organ systems, diagnostic techniques, or therapeutic interventions [66]. Specialties are often defined by the type of medical expertise required, the technology employed, and the patient population served. In the context of AI-enabled medical devices, specialty classification can therefore be based on the primary body system or disease area targeted by the device, rather than solely on its function [67].

This specialty classification aligns with the dominant logic used in international clinical taxonomies such as those defined by the American Medical Association (AMA) and the American Board of Medical Specialties (ABMS), where specialties are organized around organ systems, disease areas, and diagnostic or therapeutic domains. In contrast, the NMPA regulatory database structures device categories according to administrative and product-type considerations. Table 3.1 highlights these distinctions by comparing the AMA/ABMS clinical framework, the disease- and system-oriented categorization adopted in this thesis, and the administrative device categories used by NMPA. This comparison demonstrates that our approach is grounded in established medical specialty theory while remaining compatible with the regulatory structure of the Chinese device registry.

The top three medical specialties represented among AI-enabled medical devices in China are respiratory (20.5%), ophthalmology/endocrinology (12.8%), and orthopedics (10.3%). Six AI-enabled devices (6 out of 43) have successfully utilized the innovation fast-track pathway, underscoring the rapid advancement and regulatory prioritization of frontier technologies in critical clinical areas. All six devices are classified as Class III. All devices in the dataset were registered after 2020, with the majority registered

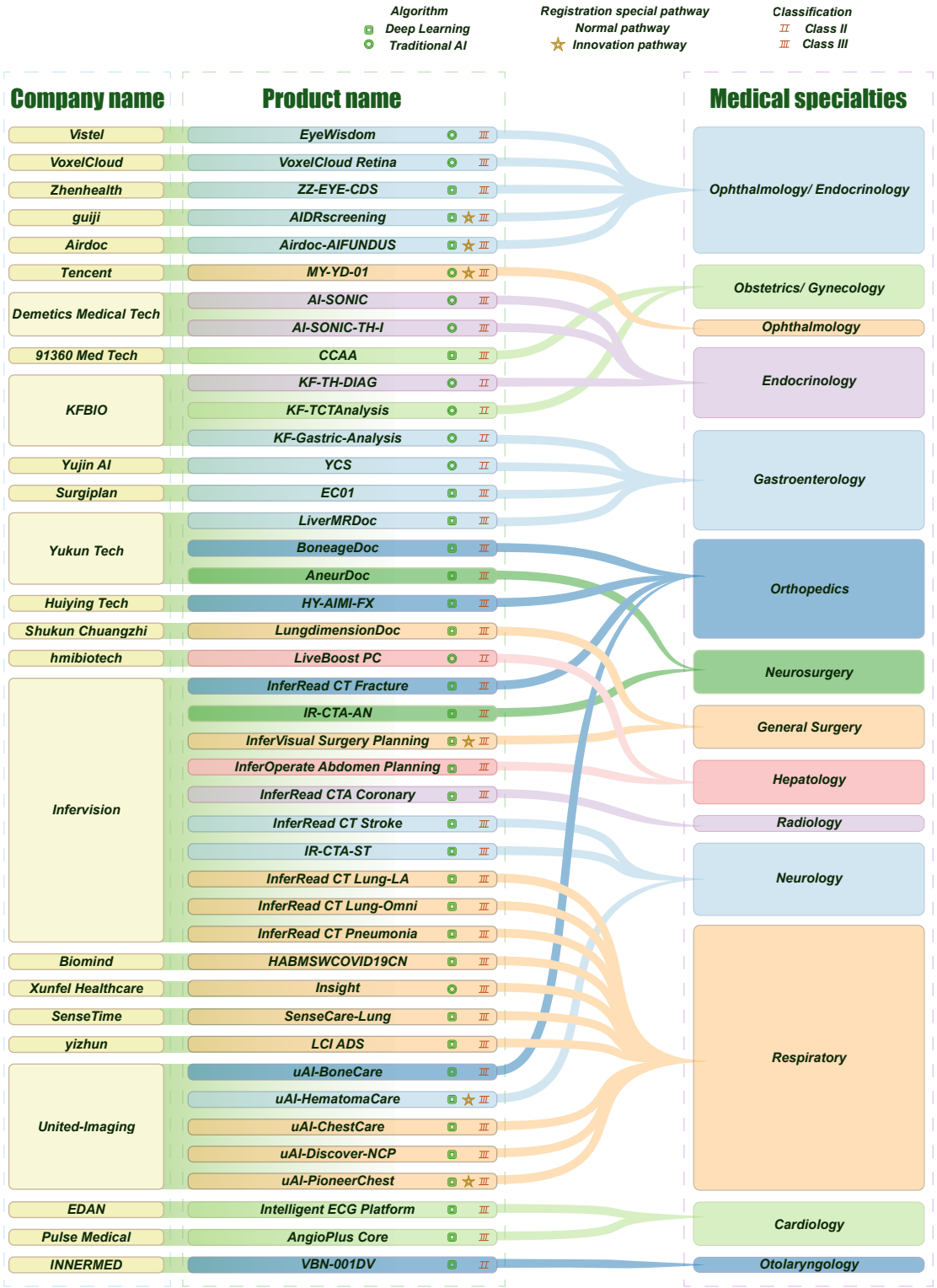


Figure 3.6: An alluvial diagram illustrating the types of 43 AI medical devices and the risk classifications along with the approval pathways within the NMPA regulatory framework

in 2023 and 2024. This trend indicates that device registration activities are aligned with the regulatory reforms outlined in Table 2.1.

Table 3.1: Comparison of Medical Specialty Classification Frameworks: AMA/ABMS, This Thesis, and NMPA

AMA / ABMS Clinical Specialties	Specialty Classification in This Thesis	NMPA Regulatory Classification
Organ-system based specialties such as cardiology, pulmonology/respiratory medicine, gastroenterology, neurology, ophthalmology, orthopedics, and endocrinology.	Disease-area or body-system targeted by the AI-enabled device, including respiratory diseases, ophthalmic disorders, endocrine and metabolic diseases, orthopedic conditions, and neurological disorders.	Administrative and product-type groupings, such as respiratory devices, ophthalmic devices, orthopedic devices, endocrine devices, and imaging devices.
Disease-focused subspecialties including oncology, infectious diseases, and metabolic disorders.	No unified oncology category; cancer-related AI devices are grouped by affected organ system.	Organ-based categories; cancer-related devices appear under their corresponding clinical department.
Diagnostic vs therapeutic specialties (radiology, pathology, radiation oncology).	Primary classification based on disease/system; diagnostic modality as secondary attribute.	Regulatory categories include both diagnostic and therapeutic device types.
Training-based, profession-defined clinical structure.	Application-oriented classification reflecting disease/system focus.	Administrative, product-type-focused structure for regulatory processes.

3.3.3 Exploratory Observations: Algorithm Types and Classification

We examine descriptive trends across devices labeled as deep learning or traditional AI to identify any emerging differences in their regulatory pathways. These analyses are intended to generate hypotheses for future study rather than confirm causality. We reviewed each device in detail to label their underlying technology. In this thesis, *deep learning* refers to methods based on multi-layer (deep) neural networks, whereas *traditional AI* refers to non-deep-learning approaches—including rule-based systems

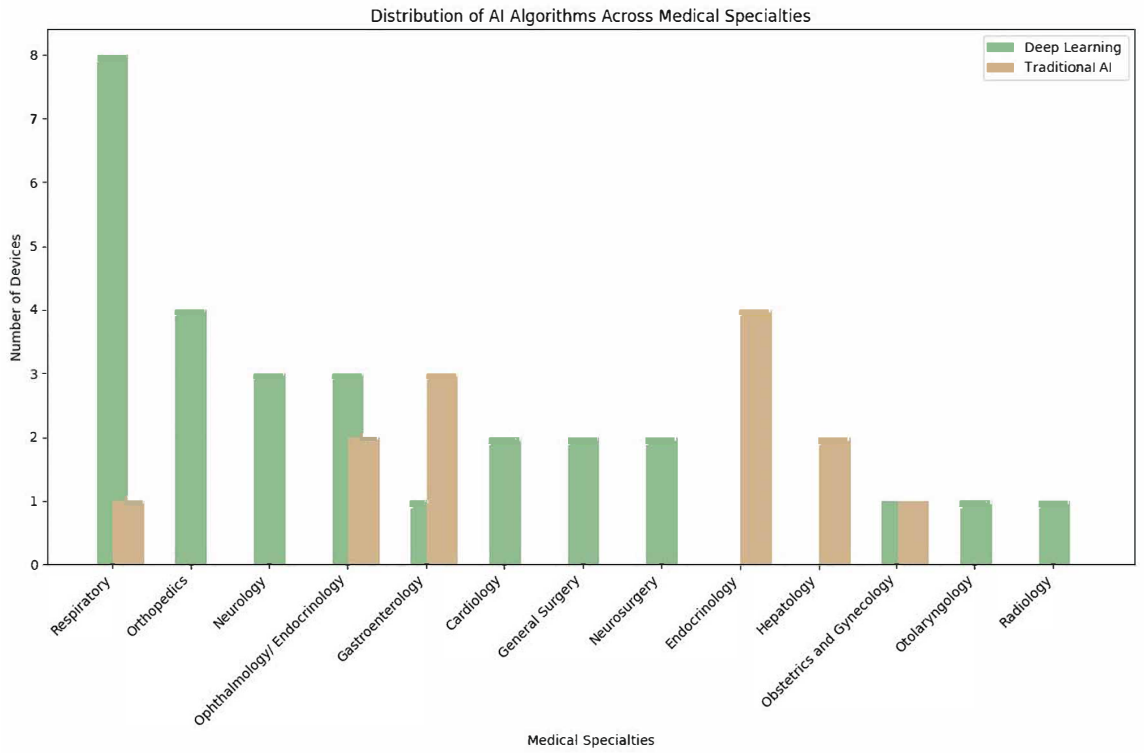


Figure 3.7: Distribution of AIMD on different medical specialty and algorithm

and classical statistical machine-learning algorithms such as logistic regression, support vector machines, decision trees and random forests. Among the 43 AI-enabled medical devices (AIMD), we found that 32 utilize deep learning (DL), while 11 use traditional AI, as shown in Figure 3.7.

While our study was not designed to formally test the relationship between algorithm type and regulatory classification, this trend may warrant further investigation. We found that deep learning (DL) is applied across a variety of medical specialties, including respiratory imaging, orthopedics, neurology, and the combined fields of ophthalmology and endocrinology, as well as in cardiology, general surgery, and neurosurgery. In contrast, fields such as gastroenterology, obstetrics and gynecology, and otolaryngology tend to rely more on traditional AI techniques. This distribution raises important questions about why certain AI technologies are adopted in specific medical specialities.

Medical imaging techniques vary by machine and medical application. CT provides detailed cross-sectional views of bones, tissues, and blood vessels[68]. MRI captures multi-planar images without radiation, ideal for diagnosing brain, spine, and musculoskeletal conditions. Ultrasound, using sound waves, is suited for real-time imaging of soft tissues and is portable for bedside use. Endoscopy allows visualization

of internal cavities, enabling real-time diagnostics and interventions. Optical imaging, like fundus photography and OCT, is essential in ophthalmology for diagnosing retinal conditions[69].

Deep learning and traditional AI algorithm analysis

Traditional AI, encompasses a wide range of AI techniques and methods, including traditional machine learning algorithms (like decision trees, support vector machines), rule-based systems, and simple statistical models[70]. These techniques are applied to tasks where structured data, such as numerical or categorical data, can be analyzed to classify, predict, or detect patterns. Deep Learning, on the other hand, is a subset of machine learning that focuses on neural networks with multiple layers (also known as deep neural networks). Inspired by the human brain, these networks consist of interconnected layers of artificial neurons that learn to recognize features from raw data inputs through hierarchical abstraction[71]. Deep learning models excel in processing unstructured data, such as images, audio, and text, due to their ability to learn hierarchical representations of data features. Deep learning, with its ability to process large, complex, and unstructured datasets, is particularly suited for specialities including ophthalmology, and neurology, where medical imaging are complex.

Analysis of AI algorithm in certain medical specialty

We extract medical specialties information automatically by identifying the disease location from the generic names. According to the "Rules for Generic Names of Medical Devices" (Order No. 19) [72], characteristic terms describing specific attributes—such as the location of the disease where the software is used—are mandated to be included in the generic name. For example, a device named "Lung Nodule CT Image Auxiliary Detection Software" is classified under respiratory medicine, while "Chronic Glaucoma-like Optic Neuropathy Fundus Image Auxiliary Diagnosis Software" falls under ophthalmology.

We use this data-driven approach to achieve finer-grained classification of medical specialties compared to methods that categorize devices solely based on the underlying imaging or sensing technologies. For example, previous approaches often label most software that utilizes CT scans as radiology, regardless of the actual disease or organ system addressed, which limits downstream analyses of clinical applications [73].

Our results indicate that deep learning (DL) is widely used in specialties like respiratory, orthopedics, and neurology. CT scans, with their high resolution and structured data, are ideal for convolutional neural networks (CNNs) [74]. CNNs excel at feature

extraction and segmentation, especially in 3D data, making them effective for detecting fractures, tumors, and lung abnormalities in chest, abdomen, and brain imaging.

In respiratory, DL is applied to CT scans and X-rays for tasks like lung cancer detection and pneumonia diagnosis [75], using supervised learning with labeled images. Orthopedics benefits from DL's ability to analyze X-rays, CTs, and MRIs for bone fractures and musculoskeletal conditions [76]. In neurology, DL analyzes brain MRIs to identify tumors, strokes, and Alzheimer's, using supervised learning with labeled data [77]. In gastroenterology, DL is applied to endoscopic images to detect polyps, while traditional AI handles structured data like blood tests [78]. Hepatology and endocrinology similarly rely on traditional AI for analyzing structured clinical data, such as liver function tests and hormone levels, aiding in disease monitoring and treatment planning [79, 80].

Figure 3.8 presents a visual representation using a human body diagram, with key medical specialties highlighted and labeled to indicate the specific AI methods and technologies used within each domain. The diagram places these technologies directly near the anatomical regions they impact, offering a more intuitive understanding for readers.

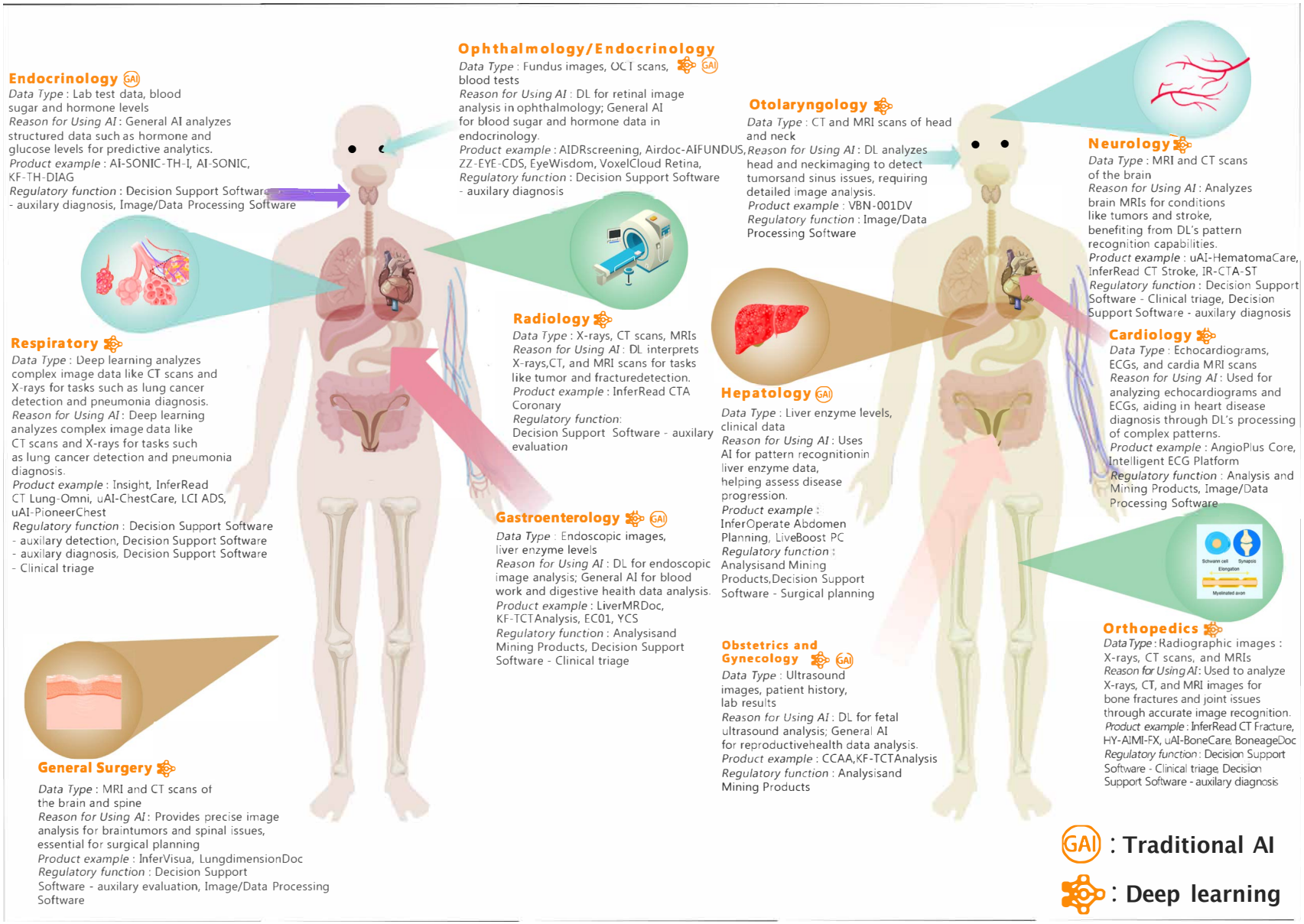


Figure 3.8: Applications of AI in Different Medical Specialties

Analysis of AI devices

From a functional standpoint, NMPA classifies AI-integrated software into specific functional categories, as outlined in Table 3.2. Of the 43 devices reviewed, 5 devices fall under the category of Image/Data Processing Software. For instance, INNERMED's VBN-001DV device falls into this category and is classified as Class II, despite incorporating deep learning components. This classification indicates that when the AI algorithm primarily enhances image quality or focuses on segmentation, the NMPA may assign a lower risk classification, reflecting the specific regulatory approach.

Thirty devices fall under the category of Decision Support Software, which is primarily designed to assist medical decision-making. This category can be further classified into subtypes based on functionality: auxiliary detection, auxiliary diagnosis, clinical triage, auxiliary evaluation, and surgical planning. For example, the "CT image-assisted triage software for intracranial hemorrhage" helps diagnose intracranial hemorrhage and supports triage decisions. The Insight, produced by Xunfel Healthcare, is one example. Although it does not utilize deep learning, it aids physicians by providing decision support for lung nodule detection in CT images, and the NMPA has classified it as Class III due to the function. Another device, the AI-SONIC-TH-I by Demetics Medical Tech, offers diagnostic support for thyroid nodule assessment using ultrasound images. Infervision's InferVisual Surgery Planning device provides surgical planning support, including the display, processing, measurement, and analysis of chest CT images. It can automatically identify pulmonary nodules of 4mm and larger, showing their spatial relationship to key anatomical structures—information that thoracic surgeons can use in planning lung surgeries. This device has been granted classification under the innovation pathway, reflecting its advanced functionality in surgical planning support.

Five of these devices are categorized as Analysis and data Mining Products, which are primarily utilized for data analysis and mining. These devices normally process structured data through AI technology to help identify potential risk factors or predict disease progression. The "chronic liver disease risk calculation software" is used to assess the risk of chronic liver disease, supporting early clinical intervention.

The fast-track pathway designation has been awarded to several devices, including uAI-HematomaCare and uAI-PioneerChest from United-Imaging in the areas of neurology and respiratory, respectively; AIDRscreening by Guiji in the field of ophthalmology/endocrinology; MY-YD-01 from Tencent, also in ophthalmology; InferVisual Surgery Planning by Infervision for general surgery; and Airdoc-AIFUNDUS from

Airdoc, again in ophthalmology/endocrinology. This trend may underscore the growing importance of ophthalmological AI technologies in China and reflects a global market demand for innovative ophthalmic solutions.

Table 3.2: Divide AIMD by software function from NMPA regulation

Category	Description	Number of Devices
Decision Support Software	Devices used for assisting medical decision-making that apply an AI algorithm, such as software for lesion identification and drug dosage calculation.	30
- auxilary detection	Assists in detecting medical conditions, often using imaging data or other diagnostic inputs.	1
- auxilary diagnosis	Supports diagnosis processes by providing probabilistic assessments or condition classification.	16
- Clinical triage	Aids in triaging patients based on severity, prioritizing care delivery.	10
- auxilary evaluation	Offers evaluation support, such as predicting treatment outcomes or assessing patient risk factors.	1
- Surgical planning	Provides guidance and planning for surgical procedures through image analysis or patient-specific modeling.	2
Image/Data Processing Software	Devices used for medical imaging and data processing using AI algorithms, enabling tasks like image segmentation and fusion.	5
Analysis and Data mining Products	Devices used to analyze and mine medical-related data, applicable in fields like drug development, medical research, and hospital information management.	5
Medical Assistant Products	Based on electronic health records or patient information, these devices use NLP and other technologies to make logical inferences about the patient's condition.	-

3.3.4 Observed Differences in International Regulatory Classification

Table 3.3 presents AI-enabled medical devices (AIMDs) imported into China, the majority of which originate from the United States and are classified as Class III, typically under the domain of radiology and regulated centrally by the National Office in Beijing.

Table 3.3: Foreign companies with software devices.

Company	Country	Classification	medical specialty
Simemens Healthcare GmbH	USA	III	Radiology
Nucletron B.V.	Netherlands	III	Radiology
Vavian Medical System	USA	III	Radiology
MIM Software	USA	III	Radiology

During our analysis, we noted several examples where the same AI-enabled device appeared to receive different classification levels across jurisdictions. For instance, the MIM Software system (Chinese license number National20153211878) is classified as Class III in China according to product code system (21-01-01), but is listed as Class II in the United States under 510(k) number K103576 [81]. These differences may reflect divergent interpretations of clinical risk or variation in regulatory frameworks, rather than inconsistencies in classification quality.

While our study does not aim to formally compare regulatory systems, these observations suggest that further research into regulatory divergence—especially for novel or AI-based medical technologies—could be valuable. The U.S. FDA permits Class II devices to be cleared via the 510(k) pathway by demonstrating substantial equivalence to existing approved devices. In contrast, NMPA frequently classifies decision-support AI systems as Class III and often requires clinical evaluations or trials, even for devices that are considered moderate-risk in other jurisdictions.

To offer additional context, we compiled Table 3.4, which outlines selected features of the FDA and NMPA regulatory frameworks. While the classification schemes are both risk-based and nominally similar, their implementation appears to vary in practice—particularly in the treatment of AI-enabled software as a medical device (SaMD).

A few illustrative cases underscore this variation: The Aidoc AI system for CT stroke detection is cleared as Class II via the 510(k) pathway in the U.S. [82], but registered as Class III in China; Viz.ai's stroke triage system, cleared under the De Novo pathway by the FDA [83], is also classified as Class III by the NMPA; In contrast, Zebra Medical Vision's radiology solutions are consistently treated as Class II in both jurisdictions [84]; For high-risk applications such as HeartFlow's AI-based CT analysis for coronary artery disease, both agencies assign a Class III designation [85].

These examples point to possible jurisdictional variation in perceived clinical risk and classification system. Many countries have faced challenges in consistently classifying AI-enabled medical devices [86], [87]. These discrepancies in classification often reflect deeper differences in legal accountability and regulatory oversight. For instance, in

the United States, liability for AI-assisted decisions is typically assessed through a tort-based framework, focusing on whether the standard of care was met. In contrast, other jurisdictions may apply stricter pre-market classification rules, driven by concerns over technological opacity or potential clinical risks [88].

Disclaimer: All observations in this section are descriptive and exploratory in nature. No inferential claims or normative judgments are made. These results are presented to support hypothesis generation and future research in regulatory science.

3.4 Discussion

The automated method and limitation

The automated search process demonstrated significant efficiency, particularly in processing a vast dataset of 4 million devices. Compared to previous studies [73], our method applies a structured filtering approach designed to improve reproducibility and consistency, with further validation to assess its accuracy across broader device categories. First, we systematically identified software-based devices and applied the 21-code rule to precisely extract Software as a Medical Device (SaMD)—a highly accurate step that is often overlooked by other methodologies without regulatory background author. This rule allowed us to confidently isolate the standalone SaMDs from the broader dataset. Subsequently, we further refined our search by implementing keyword-based filtering on the resulting medical software products, effectively pinpointing AI-enabled medical devices. This automated multi-step approach enabled us to reduce the dataset to a manageable and highly relevant subset while ensuring further rigorous selection process.

A key strength of this study is it is the first study in regulatory science about application of data science methods to map license numbers and automate the filtering of devices, offering a reusable process that contrasts with manual approaches or keyword-based methods used in other studies [55], [73]. using the code, reserachers could use their own keyword strategy and do the analysis at ease. While we do not claim that the current semi-quantitative automatic approach outperforms manual expert judgment, we believe it represents a necessary and bold step toward standardizing the classification process and generating reproducible insights. As regulatory data infrastructures evolve, combining quantitative tools with domain expertise may offer a promising path forward.

False positive results and regulatory term

In the final filtering stage, after applying keyword-based identification, we initially identified 75 devices. However, upon manual review, 32 devices were excluded(a false-positive rate of approximately 43%). The removal of these devices was not due to a fundamental flaw in our methodology but rather a result of the regulatory terminology "auxiliary diagnosis" and "auxiliary treatment," which are explicitly defined within NMPA regulations as inclusion criteria for medical devices. AI-enabled devices frequently incorporate this terminology in their official naming conventions [65]. According to regulatory guidelines, AI is presumed to be involved in auxiliary functions but is not necessarily the primary functional component of the device. It is crucial to distinguish that not all devices labeled with auxiliary functions are genuinely AI-enabled. This relationship follows a logical structure: $A \subseteq B$, but not necessarily $A = B$—meaning that while all AI-enabled devices for auxiliary diagnosis or treatment are captured, not all devices labeled for auxiliary functions necessarily utilize AI. Consequently, the primary source of false positives in our analysis arises from this regulatory language rather than a limitation in the automated filtering approach. Researchers employing alternative methodologies that do not rely on these specific regulatory terms would likely not encounter this issue. For example, one device manually excluded from our dataset was a red light therapy system. Its product description contained the term "auxiliary treatment," yet it did not incorporate any AI-based functionality. The description stated: "The red light therapy device consists of hardware and software components, typically including a light radiation source (such as LEDs), a control unit, and a support structure (which may include positioning components). It may also be equipped with light-guiding accessories. By utilizing red light wavelengths to irradiate gynecological areas (some devices may also include infrared wavelengths), it induces photochemical reactions and/or biological stimulation within human tissues, achieving an auxiliary treatment effect."

Limitations from data side and Why AI is currently absent in SiMDs

One key limitation of this study is that the dataset is derived from the UDI (Unique Device Identification) system, which does not provide a comprehensive representation of the entire medical device industry. The dataset is heavily influenced by how manufacturers choose to report their product information and how regulatory authorities define and enforce reporting requirements. This variability in data entry practices

affects the completeness and accuracy of AI-related classifications within regulatory databases. Through this study, we aim to highlight these limitations, providing insights for industry professionals, researchers, and regulatory bodies to enhance AI medical device classification and reporting standards.

Due to the absence of standardized algorithm classification in the UDI database, our classification relies on product descriptions as proxies, which may not always align with technical reality. This limitation is an inherent feature of the regulatory reporting system and highlights the need for improved transparency and labeling standards.

Our investigation into medical devices registered within the UDI system—including GE Vivid ultrasound, Canon MRI and CT scanners, and Edwards Lifesciences cardiac surgical monitors — revealed challenges in aligning known AI functionalities with information available in regulatory descriptions, when repurposing the database for identifying AI-enabled devices. Canon's MRI and CT scanners integrate Deep Learning Reconstruction (DLR) and AICE noise reduction algorithms, which enhance image quality and diagnostic accuracy [89]. GE's MRI systems feature AIR Recon DL, an AI-powered reconstruction technique designed to improve image sharpness while reducing scan times [90], GE Vivid ultrasound systems incorporate AI-based functionalities, including automated segmentation and measurement of left ventricular ejection fraction (LVEF), which enhance cardiac imaging and analysis [91]. Edwards Lifesciences' cardiac surgical monitors employ the Acumen Hypotension Prediction Index (HPI), which utilizes predictive analytics to forecast hemodynamic instability during surgery [92]. While these products are known to incorporate AI functionalities (e.g., deep learning-based reconstruction or segmentation), the corresponding UDI records do not consistently include explicit references to such terms. This reflects a limitation when repurposing the UDI database for secondary research tasks such as identifying AI-enabled features.

These findings highlight AI-enabled features are not systematically labeled in product descriptions, which presents a limitation when using the UDI database for secondary purposes such as identifying AI-enabled devices. As AI adoption in medical imaging and monitoring continues to grow, further refinement in regulatory reporting could improve transparency and facilitate AI-specific assessments within the UDI framework. This partially constitute why in this data analysis, AI is currently absent in SiMDs.

Descriptive Patterns of Algorithm Use in Medical Specialties

The classification of devices into deep learning and traditional AI was based on the terminology present in the medical device database and supplemented by expert review.

Devices were classified as *deep-learning-based* if their product descriptions explicitly mentioned terms such as "deep learning", "CNN" or "RNN". Devices were classified as *traditional AI* if they referred only to "artificial intelligence" or "machine learning" without specifying deep-learning techniques, or if they described statistical or rule-based models. This classification was not provided directly by the database, but was inferred from the textual descriptions and keywords found within the dataset.

From data, We observed that deep learning-based devices were more frequently categorized as Class III (high-risk) medical devices, particularly in imaging-related applications. Traditional AI devices, such as rule-based decision support systems, had a broader distribution across Class II and Class III categories. DL devices predominantly used for automated image analysis, segmentation, and real-time diagnostic assistance. Traditional AI devices were more frequently associated with predictive modeling, decision-support systems, and clinical workflow optimization.

From a clinical perspective, it remains an open question whether end-users—such as physicians—engage deeply with the technical architecture of AI-based tools. While regulatory agencies increasingly emphasize algorithmic explainability, it is unclear whether such transparency translates into meaningful clinical understanding or alters decision-making. Whether clinicians make a "second judgment" beyond the model's recommendation remains a critical area for future interdisciplinary research involving HCI, AI transparency, and medical education.

Descriptive Trends: Deep Learning and Classification Level

Our analysis of regulatory data reveals patterns in the classification of AI-enabled medical devices. While the assignment of Class II or Class III status is officially based on intended use and associated risk, we observe that certain types of AI algorithms, such as deep learning, appear more frequently among Class III devices. This may reflect their common use in higher-risk applications—including autonomous diagnosis, treatment planning, or real-time monitoring—though we emphasize that algorithm type is not the sole determinant of classification. Conversely, deep learning algorithms are also found in Class II devices, particularly where the application is lower-risk (e.g., workflow support or image enhancement). Similarly, some Class III devices do not use deep learning at all, indicating that traditional AI methods can still be associated with significant clinical risk when applied in complex or high-stakes scenarios.

We also observe that domestically manufactured Class III AI-enabled SaMDs outnumber imported ones, while the distribution of AI-enabled SiMDs appears more balanced.

This observation suggests differences in market strategy or regulatory positioning, but further research would be needed to assess the causes. Moreover, our analysis highlights that device records in the UDI system often lack consistent labeling regarding AI usage. This may pose challenges for data-driven research and underscores the importance of transparent reporting in regulatory documentation. Finally, we find that traditional AI techniques remain prevalent in domains such as gastroenterology and endocrinology, where clinical decisions rely more heavily on structured data and established protocols. In contrast, deep learning appears to play a growing role in more complex diagnostic specialties. These patterns merit further exploration in future work.

Deep learning (DL) models are often criticized for their "black-box" nature, as their complex architectures and high-dimensional feature spaces can make it difficult to interpret how decisions are made. In contrast, traditional AI techniques—such as decision trees, rule-based systems, or logistic regression—typically offer greater transparency and interpretability. These differences are not merely technical; they have direct implications for patient safety, clinical accountability, and regulatory oversight. For instance, when a model's decision cannot be readily explained, it becomes more challenging for healthcare providers to validate its recommendations or justify its use in clinical practice [93]. This opacity complicates legal attribution of responsibility in adverse events and often prompts regulators to impose stricter pre-market scrutiny or require post-market surveillance for DL-based systems. Conversely, the interpretability of traditional AI may facilitate clearer pathways for risk assessment and compliance.

The analyses of algorithm type and regulatory differences were exploratory in nature, and we do not draw generalizable conclusions. We acknowledge the limitations in data labeling, jurisdictional scope, and sample representativeness.

Interpretability Demands Across Clinical Specialties: Regulatory and Clinical Implications

Among the 43 AI-enabled medical devices analyzed, 32 incorporate deep learning algorithms. Of these, 21 are designed primarily for diagnostic purposes, as detailed in Appendix 2. Among the 32 deep learning-based devices identified, 21 were classified as diagnostic tools, primarily concentrated in radiology, ophthalmology, and respiratory medicine—fields that traditionally rely on visual pattern recognition and tolerate a certain degree of algorithmic opacity. Conversely, the 11 devices utilizing traditional AI (e.g., rule-based logic, statistical models) were more evenly distributed across specialties such as endocrinology, cardiology, gastroenterology, and hepatology. These fields

are characterized by structured clinical decision pathways, where physicians demand transparent justifications for each computational recommendation.

Clinically, specialties such as endocrinology and cardiology often require high levels of traceability and causality, especially in scenarios involving chronic disease management, dosage adjustment, and life-sustaining interventions. In these cases, the acceptance of opaque models such as deep neural networks remains limited. Regulatory authorities, such as China's NMPA, tend to classify AI systems deployed in such high-stakes clinical areas as Class III, regardless of whether deep learning or traditional AI is used. This reflects the elevated regulatory burden associated with low-interpretability systems in high-risk domains. Specialties like radiology, ophthalmology, and orthopedic imaging have shown a greater tolerance for black-box models, given that such tools function primarily as decision aids rather than autonomous decision-makers. Devices in these domains are more frequently classified as Class II, particularly when AI is used for pre-screening, image enhancement, or lesion detection with a physician-in-the-loop approach. Our findings suggest that interpretability requirements are not only a technical constraint but also a regulatory variable that influences classification, approval pathway, and post-market surveillance demands as shown in Table 3.5.

These observations align with broader regulatory trends in the United States and European Union, where agencies have begun to tailor evidence requirements based on algorithm type and clinical application. For instance, the U.S. FDA's Total Product Lifecycle (TPLC) framework incorporates transparency as a critical attribute in assessing AI/ML-enabled devices [94]. Similarly, the EU Medical Device Regulation (MDR) places increased emphasis on algorithm accountability in AI systems with diagnostic intent [95]. The findings suggest that the clinical demand for interpretability—and its translation into regulatory risk—is specialty-specific and context-sensitive. As AI tools become more complex and deeply integrated into healthcare workflows, developers must consider not only model accuracy but also the interpretability thresholds imposed by clinical end-users and regulators alike.

Fast-track approvals

Among the six fast-track approvals in China, three focus on ophthalmology, reflecting the demand for AI technologies like Guiji's AIDR screening device for diabetic retinopathy. The global market for such diagnostic software is projected to grow steadily, indicating strong market acceptance in China, in line with the New Generation Artificial Intelligence Development Plan [96].

3.5 Conclusions

Benjamens et al. [73] manually reviewed FDA records and identified 64 AI/ML-based medical devices, validating them through multiple sources. In contrast, Muehlematter et al. [97] reported approximately 500 AI medical devices across U.S. and European regulatory databases, although they did not clearly document their filtering or validation procedures. Benjamens et al. also highlighted the imprecision of keyword searches and emphasized the need for a dedicated AI labeling system to enable more effective device tracking.

To fully leverage data science for regulatory insights, a standardized labeling or coding system for AI functionalities is essential. While China launched an AI medical device registration platform in 2019 [98], no comprehensive, publicly accessible AI-focused medical device database currently exists. In contrast, the U.S. FDA has published a centralized and regularly updated list of approved AI/ML-enabled medical devices, which provides a more transparent starting point for research [99].

One limitation of our study stems not from the coding process itself but from the structure of the NMPA UDI database. Devices are submitted and published in batches, and some products may be missing from the dataset due to delays or incomplete reporting. This likely accounts for discrepancies between our findings and those of previous studies [55]. While the UDI dataset remains the most comprehensive registry of medical devices in China, our findings underscore the need for a more complete and AI-specific regulatory infrastructure [73].

As AI medical devices raise growing concerns around algorithmic transparency, bias, cybersecurity, and ethical use [100], the need for robust surveillance through dedicated regulatory databases becomes increasingly urgent. Questions of data privacy, fairness, and accountability—as well as the impact of black-box decision-making on the doctor-patient relationship—remain unresolved [101]. Public excitement around general-purpose AI tools like ChatGPT [102] has only intensified attention on AI's role in healthcare. In this context, the rapid proliferation of AI-enabled devices demands more coordinated regulatory oversight and public communication.

Our primary contribution lies in offering a scalable framework for identifying and classifying AI-related devices using real-world regulatory data. Rather than aiming to produce a definitive list of all AI devices in China, we demonstrate how structured filtering methods can surface meaningful patterns—and simultaneously reveal systemic

limitations in the current data infrastructure. This study represents an early but necessary step toward data-driven regulatory analysis at scale.

Our findings also suggest that even well-known AI-enabled devices may be missed under current conditions, simply because their product descriptions lack standardized terminology. This is not a flaw of the method, but a symptom of broader limitations in how AI capabilities are currently documented in regulatory systems. We therefore advocate for the enhancement of regulatory databases through dedicated AI labeling fields, consistent terminology, and clearer guidance for manufacturers.

Such improvements would not only enhance reproducibility and transparency for researchers but also support more rigorous oversight, safer deployment, and more informed policy. We hope this work will encourage collaboration among regulators, industry leaders, and researchers to build a more robust regulatory ecosystem—one capable of keeping pace with rapid advances in AI medical technology.

After exploring the policy context and conducting an empirical analysis of regulatory data, the subsequent Chapter 4, a systematic review research turns to the regulations themselves—not merely as administrative instruments, but as complex legal-linguistic texts. Through a systematic review, we examine how the language, structure, and framing of regulatory documents influence their interpretability and implementation.

Table 3.4: Comparison of Medical Device Regulatory Frameworks: FDA vs. NMPA

Regulatory Factor	FDA (United States)	NMPA (China)
Regulatory Framework	Medical devices are classified under 21 CFR Parts 800-898	Classified based on the Medical Device Classification Catalogue
Classification System	Three classes: I (low risk), II (moderate risk), III (high risk)	Three classes: I (low risk), II (moderate risk), III (high risk)
Risk-Based Approach	Risk assessment follows intended use and risk to patient safety	Similar risk-based approach, often more stringent for AI-driven decision-support devices
Class I Devices	Low risk: General controls (e.g., bandages, gloves)	Low risk: General controls (e.g., surgical masks)
Class II Devices	Moderate risk: General + special controls (e.g., X-ray systems)	Moderate risk: Requires technical review (e.g., diagnostic ultrasound systems)
Class III Devices	High risk: PMA required	High risk: Technical review + clinical trials
Approval Pathways	- **510(k)** for Class II (demonstrates substantial equivalence to a predicate device) - **De Novo** for novel moderate-risk devices - **PMA** (Pre-Market Approval) for Class III high-risk devices	- **Class II & III** require clinical evidence - **Class III** demands full clinical trials unless exemptions apply
Clinical Evidence Requirements	- **Class I**: Typically exempt - **Class II**: Bench testing, software validation, predicate comparison - **Class III**: Clinical trial data, human factors validation, risk mitigation plans	- **Class I**: Exempt - **Class II**: Regulatory testing, clinical evaluation - **Class III**: Mandatory clinical trials unless explicitly waived
SaMD Regulation	Assessed per device classification and IMDRF SaMD framework	Specific guidance for AI/software under NMPA rules
Foreign Device Registration	Requires approval in origin country or extensive documentation	Requires origin approval + local clinical trials (in some cases)
UDI Implementation	Mandatory UDI for all classes	Gradual UDI adoption; regional variations exist
Harmonization Efforts	Member of IMDRF (International Medical Device Regulators Forum), moving towards global alignment	Member of IMDRF, yet retains national requirements

Table 3.5: Interpretability Demands Across Clinical Domains

A. Clinical domains with high interpretability requirements

These domains typically involve high-stakes diagnostic or therapeutic decisions, elevated legal liability, and require physicians to clearly understand and trace system-generated outputs.

Clinical Domain / Scenario	**Rationale**	**Example**
Endocrinology (e.g., diabetes management)	Requires precise control of physiological variables (e.g., glucose, insulin); physicians rely on rule-based models	Medication dosage adjustment software
Cardiology	High traceability required for arrhythmia detection and treatment plans	ECG analysis, $CHADS_2$ risk scoring systems
Emergency / ICU	Demands rapid responses with clear intervention justifications	Shock index scoring, hospital readmission prediction
Clinical pathway management	Follows guidelines and decision trees for treatment standardization	Insulin injection decision-support systems
Pathology (non-imaging)	Emphasizes structured classification based on clinical logic	Interpretative systems for pathology reports

B. Clinical domains with lower interpretability requirements

These domains emphasize visual pattern recognition or large-scale empirical learning. Clinicians are generally receptive to AI-assisted outputs based on prior experience with diagnostic imaging tools.

Clinical Domain / Scenario	**Rationale**	**Example**
Radiology	Complex imaging features; clinicians already accustomed to AI as assistive tools	AI-assisted CT / MRI lung nodule detection
Ophthalmology	Highly structured visual input (e.g., retina); facilitates standardized model training	Diabetic retinopathy screening algorithms
Dermatology	Distinctive visual features; AI outputs easily understood and accepted	Melanoma detection tools
Preoperative Surgical Planning	Used primarily for visualization; not directly responsible for clinical decision-making	3D visualization of pulmonary nodules for surgical planning
Orthopedic Imaging	Structural changes (e.g., fractures) are easily quantifiable	Bone age estimation tools using radiographs

4

Overview of Quantitative Measurements of Linguistic Complexity for Regulations

This chapter is based on: Yu Han; Aaron Ceross; Jeroen Bergmann, Linguistic Measurements for understanding legal Text - A Systematic Literature Review, Accepted by 3rd International Conference on Political Sciences [103]

Author contributions (CRediT): Yu Han led the work (Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Visualization, and Writing – original draft); the co-authors contributed to Conceptualization, Writing – review & editing, and Supervision.

4.1 Introduction

The previous chapters (Chapters 2 and 3) established the empirical foundation of this thesis by examining the regulatory environment and analysing real-world data on AI-enabled medical device software. These pilot investigations addressed RQ1 and revealed the practical manifestations of regulatory complexity within the Chinese medical device system. With this contextual groundwork in place, the present chapter turns to the theoretical and methodological literature. Rather than positioning the literature review at the beginning of the thesis, it is placed here deliberately. Given the niche nature of AI-enabled medical device regulation, it is important for readers to first develop an understanding of the regulatory landscape before engaging with the broader scholarly work on legal and regulatory complexity. Chapter 4 therefore provides a systematic review of how complexity in legal and regulatory texts has been conceptualised and measured in prior research. This synthesis serves two purposes: it clarifies the analytical

tools available for studying regulatory complexity, and it assesses whether existing approaches—largely developed for general legal corpora—can be adapted or transferred to the domain of medical device regulation. This review directly informs the development of the methodological framework used in subsequent chapters to address RQ2 and to support later tool-building efforts for RQ3.

Law and regulation as a social function

Law is a framework for social control, comprising an interconnected body of rules that prescribe and determine the contours of acceptable actions that may be undertaken within society[104]. Furthermore, the law has a wide range of mechanisms of enforcement, which include injunction, imposing penalties or imprisonment[105]. Laws demand conformity and adherence to the stated duties and obligations; such conformity is not left up to the determination of the individual [106]. The rules contained within the law are often found in legislation, texts which are promulgated by a law-making body. These texts are then applied and interpreted by courts to produce determined outcomes.

Why the complexity of legislation is an important issue

Law and legislation texts are notoriously difficult to read. The increasing number and complexity of legal documents contribute to the high cost of knowledge acceptance by the general public [107]. In addition, overly complex regulations can lead to unnecessary economic costs to society [6]. The complexity of regulation may stem from various factors, including the design of regulations, the methods used to develop them, and the entities they regulate, making it difficult for people from all walks of life to engage with this information and resulting in different interpretations between readers [3]. The issue of complexity in legal or law texts is crucial because it can lead to confusion, loss of direction, and even malicious interpretation of the text by readers [4]. Regulations or laws should provide guidance to all stakeholders and create a platform for dialogue and development while implementing the principles of law and bringing legislation into effect. Complexity issues in regulation have been researched in various fields, including general legal and law [6], clinical settings [7–9], business areas [10], financial regulation [11], and tax law [12]. According to Hakenes and Schnabel [5], the inability of regulation to incentivize proper behavior may have contributed to the 2008 financial crisis, or at least the regulation did not mitigate its consequences.

The interpretation of legislation is often challenging due to the use of archaic language and an unusual format that differs from other types of documents. These texts contain headers and sub-sections and make references not only to other parts of the same document, but also to other legislative manuscripts, resulting in structures that are difficult to navigate and understand for non-specialists [108]. As natural language processing (NLP) technology has developed, AI algorithms are increasingly being used to analyze and highlight relevant information in legal and regulatory texts. AI machines can find relevant or potentially problematic clauses in millions of documents, improving efficiency and accuracy in legal analysis [109]. The way in which the original text information is constructed and conveyed is also crucial for better interaction with AI machines and producing better results in the long run. Reducing the complexity of legal texts is thus beneficial not only for people, but also for machines.

The necessity of measuring readability of legislation and regulations

Lawyers tend to use verbose prepositional phrases in place of simpler expressions—for example, "at this point in time" instead of "now," or "in the event that" instead of "if"" and legal workers may not realize it as "stylistic choice is merely a matter of habit, because this is how lawyers traditionally speak or write"[110]. As a result, the Plain English Movement emerged in the 1980s, which was campaigning to e.g. replace archaic words, avoid redundancy words, reduce sentence length and the use of nominalization [111]. From this initiative, and others such as the Good Law Initiative, it became apparent that there was a need for plainer language in legal text to allow for better engagement with the law, whether it being professionals or non-professionals since most people in society will come in contact with it at some stage of their life [112]. The Plain English movement reinforced the importance of plain words, favoring active-voice verbs, reducing cross-references, and making legal language clearer. It highlighted the importance of using simple language to convey ideas and replacing complex expressions in legal texts. However, there are also criticism around this movement. Some argue that the plain language heritage is opposing traditional decent legal writing, signifying a new attitude and a fundamental change from previously well-established practices. Others argue that it essential choosing between either accuracy or clarity [113].

The language used in the legal field is often criticized for its complexity, including the use of archaic language, lengthy sentences, and rare expressions that are not common in other industries, as well as the inclusion of jargon [111]. People have come to realize

the importance of using understandable language in the legal field to communicate effectively with both professionals and non-professionals [112]. In his book "The Laws of Simplicity", Maeda [114] suggests that the simplest way to achieve simplicity is through thoughtful reduction and design. However, what is considered simple and useful regulation in the legal field is debatable. Readability metrics aim to quantify how easy it is to understand a written text. However, it should be noted that the most common methods for assessing text complexity originates from techniques that are interested in determining the expected level of education required to read the text [15].

At the moment there are no studies that have systematically explored the readability metrics that have been applied to legal text. We want to determine what the most widely used methods are and in which domains they are applied. This systematic review will provide useful insights into where the field is at the moment.

4.2 Methods

Search Strategy

We used the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [115] to structure the process of filtering the literature that was gathered. The literature search was conducted using three databases: Scopus, Web of Science, and IEEEXplore. We selected eight keywords and combined them with Boolean operators to obtain papers related to legal readability or linguistic complexity. All papers published before February 2023 were included in this systematic search. Two independent reviewers read each identified abstract and thoroughly discussed their inclusion or exclusion. A third reviewer was used in case of disagreement. The search strategy included the following keyword combination: (complex* OR metric* OR measur*) AND (readability OR linguistic) AND (regulat* OR law OR legislation).

Study Selection

The inclusion criteria consisted of: (i) it must be written in the English language; (ii) It must contain readability or linguistics measurements or methodology; and (iii) It must related to the legal, law or regulatory field.

The characteristics and data extracted from each study included: (i) the year of publication; (ii) the readability measurement method used; and (iii) the domain or field of study. All eligible studies that were not accessible through library services were attempted to be obtained by contacting the corresponding author. The data

extraction process was conducted independently by two reviewers to ensure accuracy and consistency. Any disagreements were resolved through discussion and consensus. A third reviewer would be used in case of any disagreement.

4.3 Results

A total of 3,860 records were identified by searching the aforementioned databases. After removing 294 duplicates, the titles and abstracts of the remaining 3,566 papers were screened by the reviewers. Papers that were not related to the legal or regulation field were excluded from the review. Only research related to legal or law readability or complexity were kept, resulting in a total of 117 articles. Two reviewers then read the full text of each article and only kept those that contained complexity or readability metrics. After this process, 34 studies met the inclusion criteria and were further investigated. A PRISMA flowchart detailing the selection process can be found in Figure 4.1.

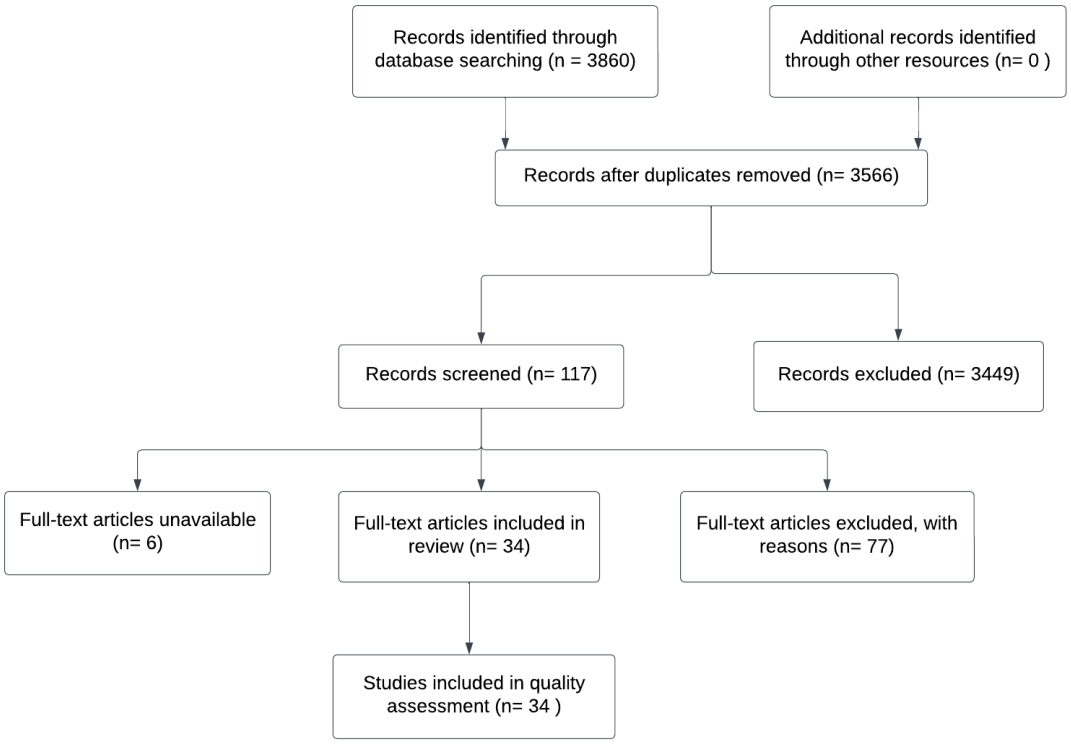


Figure 4.1: PRISMA flowchart of systematic review. The process consists of identification, duplicate removing, screening and inclusion of relevant papers.

Quality assessment
1. Does the study address a clearly focused question?
2. Do the authors discuss how they decided which method to use?
3. Is there sufficient detail regarding the methods used?
4. Was the analysis performed by more than one researcher?
5. Are the explanations for the results plausible and coherent?
6. Are there any potential confidentiality issues in relation to data collection?
7. Is any sponsorship/conflict of interest reported?
8. Did the authors identify any limitations?

Table 4.1: The items from the modified sure critical appraisal checklist for systematic review.

Quality Assessment

A modified checklist was used to evaluate the quality of each paper that was included in the final review. The quality assessment of each paper followed the questions of the adapted Specialist Unit for Review Evidence (SURE) to assist with the critical appraisal of the studies [116]. Each paper included in this systematic review was checked against the items/questions on the checklist. For each item, a positive answer would lead to a score of "1", otherwise a score of "0" was assigned. The modified checklist consisted of eight questions as shown in Table 7.13.

The list of the 34 papers that met the inclusion criteria are listed in Table 4.1. Each paper's title, linguistic measurement used, and domain of study are included in the table.

All included papers were assessed against the items on the paper quality checklist, which can be found in Table 4.2. This table also provides information on the method used in each study. A full list of the literature quality assessment is provided in Appendix 1 of Han et al. [103].

An overview of the linguistic measurement methods used in the studies and their descriptions are presented in Table 4.3.

No.	Title	Method	Domain	Quality score
1	Voluntary disclosure and complexity of reporting in Egypt: the roles of profitability and earnings management[10]	LIX index, FOG index	Business	6
2	Views of clinical trial participants on the readability and their understanding of informed consent documents[117]	Certain key word number percentage of whole	Medical ICF	7

3	Variation among Consent Forms for Clinical Whole Exome Sequencing[118]	Flesch-Kincaid Grade Level	Medical ICF	7
4	To understand or not to understand: This is the problem[119]	The Cloze Procedure, Cetinkaya and Uzun's formula, Flesch's readability formula, Atesman's formula	Medical ICF	7
5	Tax law improvement in Australia and the UK: The need for a strategy for simplification[120]	Consideration of rules in general terms, covering predictability, proportionality, consistency, compliance, administration, coordination and expression. No computational method was described.	Tax	1
6	Tax Law Complexity: The Impact of Style [121]	Flesch Readability Index and the Kincaid-Flesch Index, Reading Complexity Evaluation Index (RCE), two standard readability indices (Flesch and the Kincaid-Flesch readability formula)	Tax	5
7	A rural community's involvement in the design and usability testing of a computer-based informed consent process for the Personalized Medicine Research Project[122]	Flesch–Kincaid Grade Level Readability Formula	Medical ICF	7
8	Readability standards for informed-consent forms as compared with actual readability [9]	Flesch–Kincaid scale	Medical ICF	7
9	The readability of informed consent forms for research studies conducted in South Africa [123]	Flesch Reading Ease test, Flesch-Kincaid Reading Grade score, with the Simple Measure of Gobbledygook (SMOG) index	Medical ICF	7

10	Readability of foot and ankle consent forms in Queensland [124]	(i) Flesch–Kincaid grade level, (ii) the SMOG (Simple Measure of Gobbledygook), (iii) Coleman–Liau index, (iv) automated readability index and the (v) Linsear Wriste formula	Medical ICF	7
11	Readability of endoscopy information leaflets: Implications for informed consent [8]	Flesch reading ease, Flesch-Kincaid grade (F-K grade) and Simple Measure of Gobbledygook (SMOG)	Medical ICF	7
12	The Readability of Australia's Taxation Laws and Supplementary Materials: An Empirical Investigation[12]	Flesch Readability Index, Flesch Human Interest Index, Flesch–Kincaid Grade Level (F-KGL) Index	Tax law	5
13	Readability assessment of online ophthalmic patient information [125]	Flesch Reading Ease Score (FRES), Flesch-Kincaid Grade Level (FKGL), Simple Measure of Gobbledygook (SMOG), and Gunning Fog Index (GFOG).	Medical ICF	7
14	Readability assessment of Nigerian company income tax act [126]	FLESCH READABILITY EASE SCORE (FRES), FLESCH KINCAID GRADE LEVEL (F-KGL)	Tax Act	5
15	Readability and understandability of clinical research patient information leaflets and consent forms in Ireland and the UK: a retrospective quantitative analysis [127]	Flesch Reading Ease, Flesh Kincaid Grade Level, Simplified Measure of Gobbledegook, Gunning Fog, Fry, Raygor and New Dale Chall), (SMOG was developed specifically for assessing healthcare materials	Medical ICF	6

16	Readability and content of patient information leaflets for endoscopic procedures [128]	Flesch Reading Ease and the Flesch–Kincaid Grade Level scores	Medical ICF	7
17	Readability and Content Assessment of Informed Consent Forms for Medical Procedures in Croatia [129]	Simple Measure of Gobbledygook (SMOG) formula	Medical ICF	7
18	A randomized trial comparing concise and standard consent forms in the START trial [130]	Flesch Kincaid Grade Level (FKGL)	Medical ICF	7
19	Orthodontic treatment consent forms: A readability analysis[131]	Simple Measure of Gobbledegook (SMOG), Flesch Kincaid Grade level (FKGL), Flesch Reading Ease (FRE)	Medical ICF	7
20	A hybrid model of complexity estimation: Evidence from Russian legal texts [132]	Adapted Flesch–Kincaid formula, adapted Simple Measure of Gobbledygook (SMOG) formula, adapted formula for calculating the automated readability index ARI, Dale–Chale formula, Coleman–Liau index formula	legal	6
21	Evaluation of length and complexity of written consent forms in English and Hebrew for participation in clinical trials authorized in one medical institution in Israel: A descriptive study [133]	Flesch Reading Ease score and theFlesch-Kincaid readability test	Medical ICF	7
22	Evaluating the Readability of Informed Consent Forms Available Before Anaesthesia: A Comparative Study [134]	Gunning Fog index, Flesch-Kincaid Readability Formula and Ateşman Readability Formula	Medical ICF	6
23	Evaluating the readability of informed consent forms used in contraceptive clinical trials[135]	Fry Formula, Dale-Chall Formula, SMOG formula	Medical ICF	6
24	Emergency medicine research consent form readability assessment[136]	Flesch-Kincaid, Flesch, and Fog formulas	Medical ICF	6

25	Consent form readability and educational levels of potential participants in mental health research[137]	The Flesch Reading Ease Score (FRES), the Flesch-Kincaid Grade Level Index (FKGL), the Fog Index, and the Fry Graph	Medical ICF	7
26	Consent form heterogeneity in cancer trials: the cooperative group and institutional review board gap [138]	The Flesch Reading Ease Score (FRES) and the Flesch–Kincaid Grade Level (FKGL)	Medical ICF	6
27	Consent documents for oncology trials: does anybody read these things? [139]	The Flesch Reading Ease Score and a grade-level readability estimate using the Gunning Fog Index	Medical ICF	4
28	The Complexity of Medical Device Regulations Has Increased, as Assessed through Data-Driven Techniques [15]	Dale–Chall Readability score, Automated Readability Index (ARI), Coleman Liau Index, Gunning Fog, Flesch Kincade Grade and Bog Index	Medical Regulations	7
29	Assessing the Accuracy and Readability of Online Health Information for Patients With Pancreatic Cancer [140]	The Coleman-Liau Index, the Flesch-Kincaid Grade Level, the FORCAST Readability Formula, the Fry Readability Graph, the Gunning Fog Index, the New Dale-Chall Readability Formula, the New Fog Count, the Raygor Readability Estimate, and the Simple Measure of Gobbledygook Readability Formula	Medical ICF	7

30	Assessing readability and comprehension of informed consent materials for medical device research: A survey of informed consents from FDA's Center for Devices and Radiological Health [141]	SMOG readability grade level formula, the Flesch-Kincaid Index Grade Level Readability Formula, the Flesch Reading Ease test, and the Dale-Chall readability	Medical ICF	7
31	Are informed consent forms for organ transplantation and donation too difficult to read?[142]	Lexile Measure, Flesch–Kincaid Grade Level, and the Gunning Fog Index	Medical ICF	7
32	An analysis of the readability of patient information and consent forms used in research studies in anesthesia in Australia and New Zealand [143]	Simple Measure of Gobbledygook and Flesch-Kincaid formulas	Medical ICF	5
33	Linguistic complexity of public legal information texts for young persons [144]	Grammatical intricacy and lexical density	Legal	5
34	Common Contexts of Meaning in the European Legal setting opening Pandoras box [145]	Multilingual meaning problem	Legal	6

Table 4.2: The title, applied method, research domain and sum of the quality score is given for each of the included studies. ICF = Informed Consent Forms

Table 4.3: Linguistic complexity measurement methods and results of the systematic review.

c = character count; s = syllable count; L_w = word length; w_1 = word count; w_d = difficult word count; L_{wa} = average length of words with more than six letters; L_{sea} = average number of words per sentence; L_{se} = sentence length; se = sentence count; w_2 = count of words with more than two syllables.

Methods	Short Description	Papers
Flesch–Kincaid formulas (F-KGL)	The grade level formula is given by $0.39\frac{w_1}{se} + 11.8\frac{s}{w_1} - 15.59$	[124] [118] [8] [123] [12] [121] [119] [125] [126] [127] [128] [130] [131] [132] [133] [136] [137] [138] [140] [141] [143] [9] [142]
Flesch Reading Ease Score (FRES)	The score is computed as $206.835 - 1.015\frac{w_1}{s} - 84.6\frac{s}{w_1}$	[133] [121] [137] [138] [139] [15] [122] [123] [8] [12] [125] [126] [127] [128] [131] [134] [140] [141]

Methods	Short Description	Papers
Simple Measure of Gobbledygook (SMOG) index	An assessment tool that counts words with more than two syllables across sampled sentences. The following equation is applied: $(1.043\sqrt{w_2 \cdot \frac{30}{s}} + 3.1291)$, which can be extended to a grade level (G-SMOG).	[123] [124] [8] [125] [127] [129] [131] [132] [135] [141] [7] [140] [143]
LIX	$LIX = L_{wa} + L_{sea}$	[10]
Coleman–Liau index	$5.89(c/w_1) - 0.3(se/w_1) - 15.8$	[132] [15] [124] [140]
Linsear Write formula	The Linsear Write metric operates on a 100-word sample. Words with two syllables or fewer receive one point, while words with three or more syllables receive three points. The total is divided by the number of sentences.	[124]
Raygor Readability Estimate	Counts the average number of sentences and letters per 100 words.	[140] [127]
New Dale–Chall	$RawScore = 0.1579 \cdot PDW + 0.0496 \cdot ASL$, where $PDW = \frac{w_d}{w_1}$ and $ASL = \frac{w_1}{se}$.	[127] [135] [132] [15] [140]
Automated Readability Index (ARI)	$4.71(c/w_1) + 0.5(w_1/se) - 21.43$	[15] [124] [132]
Gunning Fog Index	$0.4[(w_1/se) + 100(\text{complex words}/w_1)]$	[125] [136] [137] [140] [139] [15]
Fry Readability Graph	A graph-based method using sentence length and syllable counts to estimate reading grade levels.	[127] [135] [137] [140]
Bog Index	Computed as Sentence Bog + Word Bog – Pep.	[15]
FORCAST Readability Formula	Grade level $= 20 - (N/10)$, where N is the number of single-syllable words in a 150-word sample.	[140]
Lexile Measure	Measures both reader ability and text complexity on a common developmental scale from 200L to 1200L.	[142]
Grammatical intricacy and lexical density	Grammatical intricacy measures clause complexity, while lexical density captures the ratio of lexical items per clause.	[144]

Table 4.4 shows the number of papers for each domain. The majority of papers included used readability metrics in the medical field, with specific focus on consent

forms. The next largest group of studies covered the tax domain, followed by the legal domain, whilst finance and medical regulations only had one paper each.

Domain	*n*	*pct*
Medical - Informed Consent Forms	25	73.5%
Tax	4	11.8%
Legal	2	5.9%
Finance	1	2.9%
Medical - Regulations	1	2.9%

Table 4.4: Number and percentage of papers by legal and regulatory domain

The most used readability method

A more complete description of the most used readability method is given below. This provides some further context with regards to the three most popular metrics that were found.

Flesch-Kincaid formulas (F-KGL)

The Flesch reading formulas comprise two related but distinct measures, which are often confused. The *Flesch Reading Ease* (FRE) score is reported on a 0–100 scale on which a *higher* score indicates an *easier* text: scores of 90–100 correspond to text that can be understood by an 11-year-old, whereas scores below about 30 indicate text that is very difficult and typically requires a university-level education. The *Flesch–Kincaid Grade Level* (F-KGL), by contrast, does *not* produce a 0–100 score; instead it expresses readability as a U.S. school grade level, reflecting the number of years of education generally required to understand the text. Both metrics were designed for the English language.

The Flesch reading formulas were originally developed for the U.S. Navy, to help military personnel better understand technical training material—a long-standing problem [146]—and were later adopted in a United States Military Standard. Both measures use the same two surface features, namely the average number of words per sentence and the average number of syllables per word, but combine them differently. In this thesis we use the *Flesch–Kincaid Grade Level*, which is computed as follows and yields a U.S. grade level (rather than a 0–100 score):

$$0.39\left(\frac{\text{total words}}{\text{total sentences}}\right)+11.8\left(\frac{\text{total syllables}}{\text{total words}}\right)-15.59$$

4.4 Discussion

The most widely applied method to determine readability was the Flesch- Kincaid and most of the studies focused on the medical domain (more specifically on the informed consent forms). In general, most studies scored high on the critical appraisal checklist. None of the papers had the analysis conducted by more than one researcher. Most papers used a computational approach to determine the metrics, so this additional check should carry less weight in the final appraisal of the research.

Readability assessment in other languages need to be carefully considered as most metrics are developed upon the English language. El-Din et al. [10] introduce an updated measure of the readability to allow it to be used for Arabic-written texts. Ashkenazi et al. [133] took a different approach and compared English and Hebrew versions in terms of basic readability metrics. They found that consent forms translated into English tended to be longer compared to their Hebrew counterparts. Differences between languages need to be accounted for in order to create metrics that work well across regions and more research needs to be done to better understand how create such metrics.

It should be noted that, each linguistic measurement has its own advantages and limitations. For instance, the Flesch-Kincaid calculates the final score based on the number of words per sentence and the number syllables per word, whilst SMOG specifically counts polysyllabic words consisting of 3 or more syllables. This means that SMOG creates a "threshold" value, which is not present in the Flesch-Kincaid approach. This idea of identifying more than 3 syllables is formed for the English language and care should be taken on how representative this method is for specific domains. The New Dale Chall suffers from a similar limitation, as it uses the percentage of difficult or non-familiar words. What is considered difficult might again differ per domain.

All of the current readability methods that were found lack inclusion of essential information that is likely to influence the complexity. Aspects such as the referencing between legal documents, the structure of the legal text, as well as the professional words applied all should be considered. Since readability formulas were developed for assessment from a young age, they are less useful to capture difference between text for a "typical professional" reader. They ignore background differences among readers and the influence that different content, layout, and retrieval aids might have [147]. Other researchers have proposed other ways to measure complexity linguistically, by using

for example the number of words [148] or pages [149], or legislative duration [150], or even a text's word entropy [151], which uses unigram tokens to measure the variety of words in a text, as it reflects conceptual variety [152]. Since law and regulation have a special structural format, researchers have also considered to measure the structural size and element depth as an attribute of complexity [152]. However, this review found that the approaches that are applied currently on "legal" documents remain relatively simple. Moreover, it shows that certain assessments only consider a single metric to draw conclusions from. Applying a wider range of metrics can provide a better reflection on the readability of a document and should be encouraged.

It should be noted that concepts such as cross-referencing is very common in legal and law documents. Nonetheless, current metrics do not take this into account as part of the assessment of the complexity. It further highlights the limitation of the metrics that were found in this review. Ideally, the metrics should cover the full experience of readers. Reading metrics could benefit from a more multi-modal approach that aims to cover several aspects at the same time. This holistic suggestion might be better in capturing the true experience of a reader. The inclusion of linguistics metrics within such a multi-modal can be very useful as well.

Linguistic measurements have not been used much in the legal or law contexts, with only a few papers exploring this area [144][145][15] . In other fields of study, several linguistic analysis techniques are frequently used and they include e.g. Zipf's law, Halstead measures or entropy. Not all of these are equally suitable for legal text, but some of them might provide unique further insights into the complexity of the text.

Halstead measures, originated from a need to explore the complexity of software programs and were introduced by Maurice Howard Halstead in 1977. Jean-Edouard and Co-Pierre leveraged the idea of algorithmic complexity in computer science and applied this method to finance regulation[153], by calculating norms of operators and operands, they generated a new way for evaluating complexity of regulation text. The Halstead method was also applied to medical device regulations [15] . This study also introduced time spend on a particular part of the text, as a possible marker for readability. This metric would of course not provide a single value for a given text, but generated a distribution of times for such a text. It was shown that the Halstead-based complexity metrics did not correlate strongly to the time that people spent on the text, further, highlighting the need to explore complexity across a range of metrics in order to increase our understanding on which aspects might make it difficult to read legal content.

From Table 4.2, we can see that several readability metrics have been applied in the medical domain, particularly in relation to Informed Consent Forms (ICFs). There is a legal duty of (medical) professionals to ensure that they obtain informed consents from patients before any trial, assessment or treatment can be carried out. The same applies if they would like to obtain personal data from the patient for a clinical trial. Thus, signing the informed consent form means that the patient agrees to what has been proposed in the (patient) information sheet. For consent to be informed, participants must understand what they are agreeing to. Concerns about the complexity of the consent document have been raised and this has probably led to more interest in applying quantitative methods [139]. It is clear that this field is at the forefront of applying readability metrics and that other fields can benefit from a similar increase in studies around the subject of readability. However, the suitability of available readability metrics is already greater in the research domain of "Informed Consent Forms", as they need to cater to all people including those with limited understanding of the languages. Most other fields might be more interested in the ease of understanding of legal text for (new) professionals or professionals in other disciplines.

Readability metrics are not a silver bullet to fix all issues with regards to reducing the complexity of legal content. As mentioned, readability metrics have many limitations, as most readability metrics were developed for assessing the education of young children and are not fit for purpose in more technical documentation [147]. Yet, these metrics might still provide a good first step to create a more robust discussion on how we can decrease the complexity of legal documents.

Overall, there is a lack of research exploring the readability of legal texts, with "informed consent forms" being the only domain that has received a sufficient amount of attention compared to others. Furthermore, the metrics used in these studies are not always suitable for the purpose of measuring complexity. It is likely that specific professional domains require their own metrics that cover a wider range of aspects beyond just linguistics. Although there are linguistic methods available, such as Halstead Measures or Entropy Complexity, they have not been widely used in the legal and regulatory context. Currently, there is a dearth of research on the readability and complexity of legal and regulatory content. This systematic review emphasizes the need for the development of new metrics that take into account the unique features of legal and regulatory domains. More research is necessary to better understand how legal texts can be made more comprehensible.

Given these limitations, it is clear that regulatory complexity cannot be inferred solely from readability metrics, and we also found that existing metrics are difficult to apply effectively in the context of regulatory affairs. In the next chapter, Chapter 5, the focus shifts from textual analysis to qualitative inquiry. Recognizing that regulatory complexity cannot be fully understood through readability metrics alone, we conduct an in-depth exploration of how regulatory challenges are perceived and addressed in practice. Drawing on interviews with experts and stakeholders, the next chapter investigates how different dimensions of complexity manifest within regulatory affairs. Through thematic coding and qualitative analysis, we aim to uncover the institutional, procedural, and human dimensions of regulatory compliance in real-world contexts.

5

Qualitative and quantitative methods to explore complexity in Regulatory Affairs

This chapter is based on: Yu Han; Aaron Ceross; Jeroen Bergmann, More than Red Tape: Exploring Complexity in Medical Device Regulatory Affairs, Frontiers in Medicine, Published in 2024 [154]

Author contributions (CRediT, as published in Frontiers in Medicine*): Yu Han — Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing; Aaron Ceross — Conceptualization, Methodology, Supervision, Writing – review & editing; Jeroen Bergmann — Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing – review & editing.*

As mentioned in the last chapter, this study investigates the complexity of regulatory affairs in the medical device industry, a critical factor influencing market access and patient care. Through qualitative research, expert insights were gathered to illuminate the factors contributing to this complexity. The study involved semi-structured interviews with 28 professionals from medical device companies, specializing in various aspects of regulatory affairs. These interviews were analyzed using a mix of qualitative coding and natural language processing (NLP) techniques. The findings reveal key sources of complexity within the regulatory landscape, divided into five domains: 1) regulatory language complexity, 2) intricacies within the regulatory process, 3) global-level complexities, 4) database-related considerations, and 5) product-level issues. The participants highlighted the need for strategies to streamline regulatory compliance, enhance interactions between regulatory bodies and industry players, and

develop adaptable frameworks for rapid technological advancements. Emphasizing interdisciplinary collaboration and increased transparency, the study concludes that these elements are vital for establishing coherent and effective regulatory procedures in the medical device sector. Regulatory affairs, crucial in industries such as healthcare, involves understanding and adhering to regulations to ensure product safety, efficacy, and quality requirements are met while promoting innovation [155]. Understanding the complexity of this work is essential due to legal implications, its impact on public health, and the evolving regulatory landscape. Moreover, regulatory affairs play a vital role in safeguarding public health, monitoring post-market performance, and adapting to evolving guidelines. Constant awareness of regulatory changes is crucial for professionals and policymakers, fostering evidence-based policy-making. Investigating regulatory affairs is indispensable for legal compliance, public health impact, and adapting to evolving industry needs [156].

In the process of bringing a medical device product to the market, it is necessary to obtain approval from the Medical Device Regulatory Authority (MDRA). The MDRA rigorously reviews the device dossier to determine its compliance with safety and efficacy requirements outlined in regulations and guidance. Regulations, as noted by Sakuma [157], can act as a double-edged sword; when applied effectively, they foster device innovation and the development of rational evaluation methods for safety and efficacy in medical devices. However, they can also form barriers for innovation. [158] Guerra emphasizes the need for a more streamlined regulatory evaluation to encourage innovation, while Onur [159] finds that shorter regulatory review times can lead to increased innovation in the medical device market. Stern [160] express concerns regarding the potential hindrance that regulations may pose to innovation in the medical device industry. Other authors raise concerns about the FDA approval system's ability to ensure the safety and effectiveness of complex medical devices, suggesting that regulations may slow down innovation in some cases [161]. Furthermore, there are debates about implementing an even more transparent and evidenced-based system of regulation [14]. Further research and interdisciplinary collaboration are required to comprehensively evaluate the impact of regulations on medical device innovation and to develop effective policies that foster both patient safety and technological advancement.

There is still a lack of research investigating the complexity in the regulatory affairs field. Some work has been published on the linguistic complexity of EU regulations [15], but this is only part of the overall complexity experienced by stakeholders. This chapter investigates the complex intricacies of regulatory affairs in the medical device

industry, aiming to gain a comprehensive understanding of the topic. This research will involve conducting semi-structured interviews with regulatory affairs professionals working in the field. The aim is to obtain insights into the complexities and challenges that are faced in the regulatory field. The findings from these interviews will contribute to enhancing knowledge and identifying potential strategies for navigating any issues of regulatory affairs in the medical device industry.

5.1 Research design and methods

The application of qualitative research methods in system engineering has been recognized for its ability to provide nuanced insights into complex social phenomena that intersect with engineering practices [162]. Qualitative research methods, such as interviews, focus groups, and ethnographic observations, offer valuable tools to explore the social, cultural, and behavioral aspects that influence a system of systems. These methods enable researchers to uncover contextual factors, understand stakeholder perspectives, and delve into the dynamics of human interactions within a system. In the medical field, several qualitative studies have been conducted to investigate key aspects of clinical and ethical implications. For example, Dhruva et al. [163] sought expert opinions on the consequences of expedited development and regulatory review pathways for new drugs and devices. Polisena et al. [164] explored the use of real-world data and evidence for medical devices through interviews. Additionally, Zhang et al. [165] focused on identifying factors influencing patients' preferences for primary healthcare institutions, employing qualitative methods to uncover the complexities of patient decision-making. These examples show the value of integrating qualitative methodologies within other domains.

5.1.1 Research Ethics

Approval for this study was granted by the Ethics Committee of University of Oxford. (Central University Research Ethics Committee (CUREC) Approval Reference: R71265/RE003). The study procedures, along with the associated risks and benefits, were explained to all participants before data was collected. The interview was only conducted after informed consent was obtained.

5.1.2 Semi-structured interviews

The interview consisted of posing several key questions to identify the underlying factors that contributed to the complexity within this field. By further exploring emerging themes, a deeper understanding can be gained regarding the challenges that were encountered. The interviews were conducted in 2023. These were held online using Zoom. This study utilised a snowball recruitment strategy, in which initial participants were asked to identify further eligible individuals within their professional networks. Snowball sampling is commonly used in qualitative research to access specialised or hard-to-reach expert groups [166]. In total, 28 volunteers participated this interview.

Inclusion/exclusion criteria

Inclusion criteria for selecting interview participants are as follows:

1. **Position Relevance:** Interviewees should occupy positions directly related to regulatory affairs within the medical device industry. Suitable positions include:

   - Vice President (VP) of Regulatory Affairs
   - Chief Executive Officer (CEO) of a medical device company
   - Regulatory Affairs Manager or Specialist within an organization

   These positions ensure that the interviewee possesses relevant expertise and insights into regulatory matters.

2. **Company Registration:** If the participant is employed by a manufacturer, the participant's company must be officially registered in accordance with the national laws of the country where the company is based. This criterion ensures that the participant represents a legitimate entity within the medical device industry.

3. **Engagement in Registration Activities:** The participant should be actively involved in medical device registration activities. This involvement ensures that the participant has practical experience and knowledge of the regulatory processes.

Additionally, an overview will be provided regarding the occupation of each volunteer, the relevant local authority they engage with, and the professional group that best describes their domain. This overview enhances the understanding of the participants' backgrounds and expertise.

Table 5.1: Characteristics of interview participants. These include the job the volunteer was holding at the time of the interview, the regulatory authority they were engaged/involved with and the work domain that best described the participant.

No.	Occupation	Authority	Groups
1	MNC (Multinational Corporation) regulatory affairs senior manager	FDA	Manufacturer
2	Regulatory affairs consultancy company CEO	EU, FDA	Consultancy
3	NMPA consultant	NMPA	Regulator
4	RA manager for global medical device company	FDA	Manufacturer
5	VP of regulatory affairs	FDA, EU	Manufacturer
6	Independent consultant of regulatory affairs	FDA, EU	Consultancy
7	Regulatory affairs engineering	NMPA	Consultancy
8	Regulatory affairs specialist	NMPA, FDA	Manufacturer
9	Regulatory affairs specialist	EU	Manufacturer
10	Regulatory affairs consultancy company CEO	FDA	Consultancy
11	MNC regulatory affairs senior manager	NMPA	Manufacturer
12	International regulatory affairs specialist	FDA, NMPA	Manufacturer
13	Regulatory affairs consultancy company CEO	FDA, NMPA, EU	Consultancy
14	VP of regulatory affairs	FDA, NMPA	Manufacturer
15	Independent consultant of RA	FDA	Consultancy
16	Quality and regulatory manager	FDA	Manufacturer
17	Regulatory company CEO	EU	Consultancy
18	Regulatory affairs specialist	NMPA, FDA	Manufacturer
19	Regulatory affairs senior specialist	NMPA	Manufacturer
20	Regulatory affairs manager	FDA	Manufacturer
21	Regulatory affairs senior manager	EU and FDA	Manufacturer
22	Regulatory affairs VP	EU	Manufacturer
23	Quality and regulatory manager	FDA and EU	Manufacturer
24	Quality and regulatory manager	NMPA	Manufacturer
25	International regulatory manager	EU, FDA	Manufacturer
26	MNC regulatory affairs senior manager	EU	Manufacturer
27	International regulatory specialist	FDA, EU	Manufacturer
28	Regulatory specialist	FDA	Manufacturer

5.1.3 Data Collection

We collected data via semi-structured interviews. Our interview guide was carefully developed based on existing literature and aimed to explore participants' perceptions and experiences regarding regulatory affairs within the medical device industry. This structured approach ensured thorough data collection and analysis, providing valuable insights into the intricacies of regulatory affairs.

Once a participant confirmed their participation, an interview was conducted using the online platform Zoom. Only volunteers who met the inclusion criteria, expressed willingness to participate, and provided signed informed consent were ultimately chosen as interviewees.

Each interview lasted between 45 to 60 minutes. Each interview was conducted by two researchers, both knowledgeable about regulatory affairs and medical device industry, and experienced in qualitative research methods. One researcher conducted the interview, while the other documented the responses and ensured the smooth flow and integrity of the interview process. The majority of the interviews were conducted in English. However, for a few interviews conducted in Chinese, the questions were translated on the spot from English to Chinese. For these responses, they were translated first into English and then independently back translated into Chinese. This ensures that the true meaning of the responses were accurately translated.

The interviews were recorded using the Zoom software. Afterwards, the recordings were sent to a professional transcription service, where they were converted into MS Word files. The qualitative analysis data were then cleaned up and organized in the MS Word files. To facilitate our analysis, we employed Nvivo (Microsoft 12 Version), a qualitative data analysis software, to code and categorize themes derived from the collected data. The software allowed us to systematically organize and analyze the data by assigning relevant analysis codes to specific themes.

5.1.4 Data Analysis

We analysed the data using two complementary approaches. The first approached utilised a qualitative approach, open coding. The second approach used NLP to attempt to find any emergent patterns in language which human annotators are unable to immediately detect.

Open Coding

Several recurring themes and ideas were identified during the data analysis phase. These were then summarized in a codebook using concise descriptors, referred to as nodes in NVivo. These codes were subsequently applied to code the transcripts from the interviews.

In our qualitative analysis, the coding structure was meticulously designed to facilitate a nuanced exploration of thematic elements. We initiated this process by establishing a “parent node", representative of a broader thematic concept, such as “complexity from legal language." This node then served as a foundational point for branching out into one or more “daughter nodes", such as “lack of legal education" or “difficult interpretation." These subsidiary nodes provided a granular and detailed elucidation of the overarching parent node theme. Our methodological approach was deeply rooted in interpretive phenomenology, guiding our open coding strategy. This approach was instrumental in interpreting and understanding the lived experiences of our respondents. By employing open coding, we deliberately refrained from imposing preconceived theoretical frameworks on our data. Instead, we sought to immerse ourselves in the raw, unfiltered narratives of our participants. This approach was pivotal in uncovering and appreciating the intricate complexities embedded within the regulatory landscape. Through this meticulous process, we were able to cultivate a profound and empathetic understanding of the nuanced realities faced by those navigating this domain.

As the project progressed, we conducted a meticulous review of the codebook and the results obtained from the NVivo analysis to uncover overarching concepts. These concepts form the basis of the themes identified through thematic analysis [167], which will be presented in a subsequent section. To ensure the validity and reliability of the coding process, two independent reviewers carefully examined the codes and engaged in thorough discussions to share their perspectives. In cases where differences of opinion arose, a third reviewer was consulted to facilitate resolution and achieve consensus.

NLP analysis

In our study, NLP techniques were leveraged to delve deeper into the dataset. Specifically, we applied NLP methodologies to the transcripts derived from our interviewees, who were categorized into three distinct groups: Manufacturers, Regulators, and Consultancy.

To this end, n-gram analysis was employed on the transcripts from each of these groups, aiming to uncover and analyze underlying linguistic patterns.

Our analytical framework was augmented by the integration of machine-based methods alongside manual coding, particularly for the purpose of topic modeling. A significant aspect of this computational analysis involved the application of the Latent Dirichlet Allocation (LDA) method. Renowned for its efficacy in both natural language processing and machine learning domains, LDA serves as a probabilistic model adept at uncovering concealed thematic structures within extensive textual datasets. This approach's effectiveness in revealing hidden topics within textual data has been underscored in the literature, notably by Bhat et al.[168]. Additionally, we utilized N Gram Analysis, recognized for its utility in exposing distinctive language patterns in NLP applications, as delineated in studies such as Vatanen et al. [169].

5.2 Results

Twenty eight RA professionals volunteered and were included in the study. Notably, a number of participants showcased extensive experience across multiple regulatory domains, enhancing the study's depth. Specifically, 2 experts possessed insights into both NMPA and FDA practices, 1 expert demonstrated expertise in both NMPA and the EU, while 5 experts navigated the areas of FDA and EU collaboration. Furthermore, 1 interviewee had a background of engaging closely with NMPA, FDA, and the EU. The subsequent analysis delves into the nuanced insights shared by participants, offering a holistic view of the regulatory landscape across different jurisdictions, see Table 5.1.

5.2.1 Results of Open Coding

The interviews mainly generated five key themes about contributing factors of complexity in regulatory affairs: (1) Complexity in Legal Language, (2) Complexity in Registration Process, (3) Complexity in Available Database, (4) Complexity in Product and (5) Complexity in Global Level.

Figure 5.1 presents a visualization of all the nodes in the project. This figure provides a concise overview of both the parent themes and the corresponding child themes that collectively capture the intricacies within regulatory affairs. We found that these child nodes were inherently connected, as one was leading to another. For example, a lack of database, which will cause trouble in mother node "registration process", also led

to ambiguity in regulatory interpretation. A more detailed description of some of the most frequently mentioned nodes will be provided later.

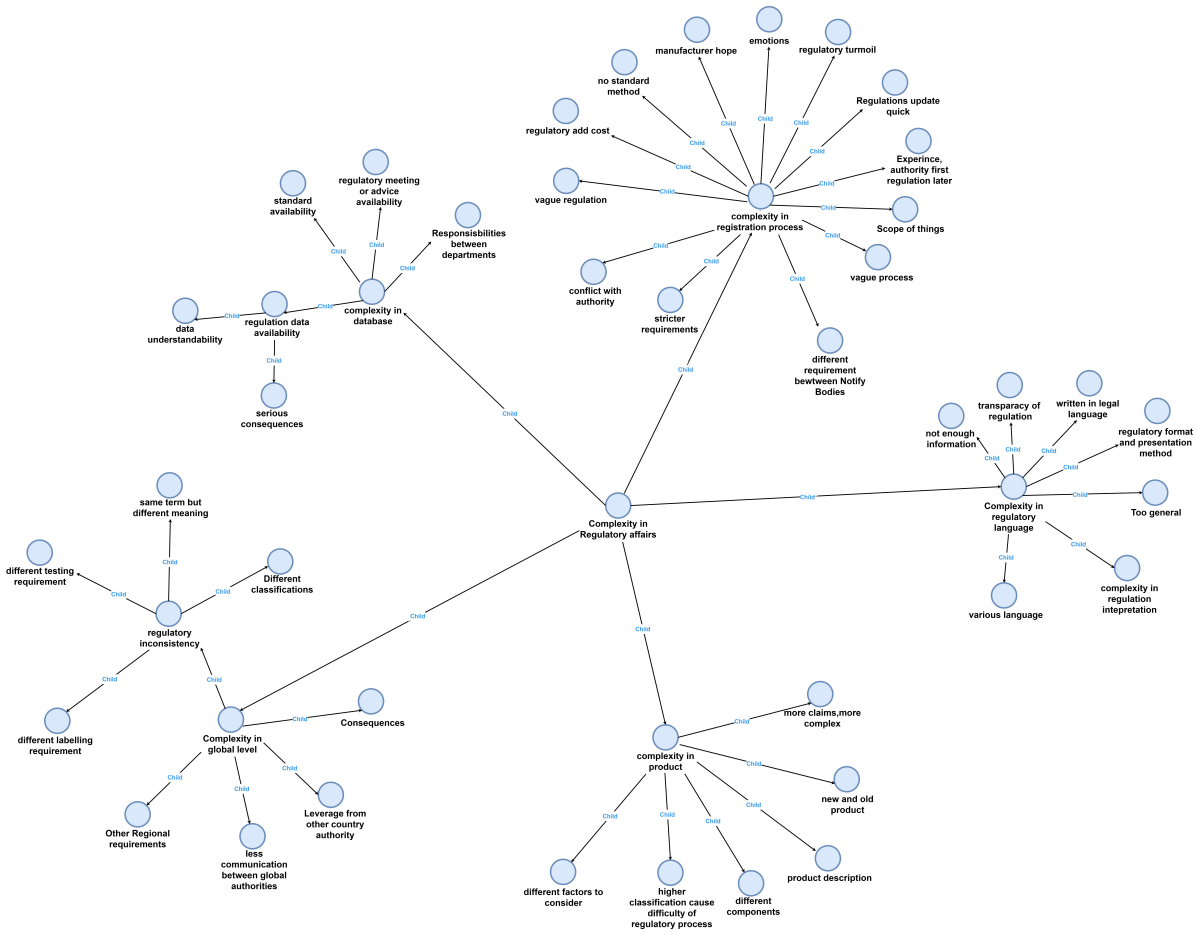


Figure 5.1: Navigating Complexity in Regulatory Affairs with Nvivo

5.2.2 Interview Theme 1 : Complexity in Regulatory Language

During the interviews, a prevalent concern expressed by the interviewees was the perceived generality of legal language within regulatory frameworks. This generalization often resulted in challenges for manufacturers in adapting their specific devices to conform to overarching regulations. Consequently, a sense of ambiguity arose regarding the necessity of compliance with certain regulations.

> "Everyone feels that no matter where the laws and regulations are, they are very obscure. We want to know whether regulations can be written more easily so we know how to implement it," said a medical device innovator.

The general feelings about different country-specific regulatory environments were also different. According to interviewees, the language used in FDA regulations are

relatively easier to read compared to those in the EU. As depicted in Figure 5.1, the child themes emanating from this mother theme comprised the following aspects.

Obscurity of legal language

During interviews, a prominent theme emerged regarding the intricate complexity of legal language within regulatory texts. The utilization of legal terminology frequently spawns intricacies and ambiguities, presenting formidable challenges for individuals lacking a legal background.

This issue is exacerbated by the necessity to align legal mandates with technical specifications, a task that demands a deep understanding of both domains. The use of terms such as “shall", “must" and “should" was cited as a prime example of the nuanced language that necessitates clear comprehension. Additionally, the distinction between legal norms and technical concepts further contributes to a sense of opacity, rendering interpretation an exercise that requires interdisciplinary collaboration and contextual awareness. The legalistic style of language often leads to a communication dissonance, particularly for individuals of a more scientific background. These professionals find it challenging to discern the underlying technical requirements obscured by the legal prose, contributing to a perceived lack of clarity.

This issue holds significance as clear and precise legal language is paramount for industry functioning. The inability to grasp and interpret legal texts not only hinders individual comprehension but also poses a substantial threat to the efficient execution of regulatory affairs activities. Such disruptions in regulatory processes have the potential to create obstacles for all operations within the entire industry, ultimately affecting the industry’s overall functionality and productivity.

Presentation of regulations

Issues stemming from the format and presentation of regulations, hindering effective information delivery and comprehension. The presentation format of regulations can vary considerably between different regulatory bodies. In the case of the FDA, regulations are structured with a clear hierarchy, starting from the Code of Federal Regulations (CFR), followed by regulations and guidelines. Changes are made within a coherent framework, continuously refining the system. This clarity aids accessibility and understanding. An interviewee highlighted the relative straightforwardness of navigating FDA regulations compared to European counterparts:

> "It [the FDA] seems a bit more direct ... It's a bit easier to go through ... It seemed just easier..." — Interviewee 2

Conversely, European regulations, such as MDR (Medical Device Regulation), can be extensive and complex. The MDR's comprehensive coverage necessitates section-by-section analysis, potentially hindering comprehension, according to interviewees. The interlinked nature of FDA regulations streamlines navigation and offers a more intuitive experience, while the European counterpart may require more effort to extract relevant information. Additionally, the language used in FDA regulations appears comparatively straightforward, fostering ease of comprehension, especially for non-native English speakers. This layout reduces the need for constant cross-referencing between visuals and text, as seen in the EU regulations. This clarity benefits those not fluent in English and those interacting indirectly with health authorities. In summary, the FDA's structured and coherent format, along with its language simplicity, enhances accessibility and comprehension for a broader audience compared to the more intricate European regulatory documents.

Difficulty in accessing sufficient information

An issue of noteworthy concern voiced by interviewees pertains to the challenge of acquiring comprehensive and timely information. It was evident that the landscape of regulatory information, especially within the context of medical devices, is subject to frequent updates and revisions. This rapid pace of change can render it difficult for stakeholders to keep abreast of the latest guidelines, leading to potential gaps in knowledge. Given the criticality of aligning devices with regulatory mandates, interviewees highlighted the necessity of dedicated efforts to stay informed through continuous engagement with regulatory agencies, industry peers, and information dissemination channels. The perceived scarcity of consolidated and readily accessible information amplifies the intricacy of compliance efforts.

Various languages of regulations

In examining the influence of language diversity on the comprehension of regulations, it is essential to recognise that European regulatory texts, although available in multiple languages, often pose challenges for individuals with a scientific background. The use of legal jargon, combined with the intricacies of translation, can obstruct understanding.

> "The European system has multiple languages. The regulation is drafted in English, but then it's translated into the various languages. The translation to other languages creates a lot of additional trouble, because sometimes you translate sentences in a different way, and this is perceived by the readers in a different way," explained Interviewee 9.

This statement illuminates the heart of the issue: the transformation of regulatory language through translation does not merely change words; it can alter meaning. Such alterations can lead to variations in interpretation, potentially skewing the original intent of the regulation. The challenge is akin to untangling the syntax of a foreign language, underscoring the critical need for precision in translation to preserve clarity and the intended message across different linguistic landscapes.

This complexity does not only signify the technical difficulty of accurate translation but also points to a deeper issue within the regulatory framework. It illustrates how linguistic nuances can cause significant shifts in understanding, thereby affecting regulatory compliance. The risk of distortion in regulatory intent highlights the need for a more nuanced approach to translation—one that goes beyond literal equivalence to consider the contextual and cultural dimensions of language. These insights shed light on the intricate relationship between regulatory language, translation practices, and cross-linguistic comprehension. They emphasize the significant challenge of ensuring consistent interpretation and adherence to regulations across the diverse linguistic tapestry of Europe.

Regulation interpretation

The problem of interpreting regulatory mandates emerged as a focal point within the interviews. A prevalent observation was the inherent complexity and ambiguity present within regulations, leading to divergent interpretations among professionals in the field. This multifaceted issue is compounded by the balance between regulatory precision and the need for regulatory frameworks to accommodate a broad spectrum of devices, each with its unique characteristics. As a result, manufacturers, regulators, and legal advisors are often confronted with the challenge of deciphering regulations in a manner that aligns with the intended objectives while accounting for the nuanced contexts of various devices.

Various countries have their own specific legal interpretation problem. In the EU, Interviewee 17 provided a critical view on the MDR:

> "I think it is a complex regulation, and I can tell you that nobody can understand this regulation. If anybody say they can understand. I would

> love to see that individual. Because it's still complex. I don't think that they did due diligence when they released the MDR."

Similarly, a perspective from China was shared by Interviewee 19, who commented on the NMPA regulation:

> "For us, when it comes to interpreting regulations or reading regulations, the common situation is, we think we can understand all the words, but we need to further consider the intended meaning behind them. We have had very detailed communication with the regulatory drafters and some evaluators. From the perspective of the regulatory drafters, they may feel that their expressions are very straightforward. However, from our perspective, we often do not know how to proceed to the next step."

This issue was rather common across all regions. People found that the language used in medical device regulations presents various challenges, hindering smooth compliance and implementation. Regulations sometimes remain too general, lacking the necessary specificity to address the complexities of the medical device industry. This ambiguity can lead to confusion in interpreting requirements and varying understandings among stakeholders. Moreover, the complex legal language used in regulations can make it difficult for manufacturers, regulatory authorities, and healthcare professionals to precisely interpret the obligations and standards. Multiple language versions of regulations, along with diverse formats, further complicates the compliance efforts. Lack of transparency and insufficient information in some regulations can also hinder stakeholders from making well-informed decisions. Addressing these issues is crucial for promoting clarity and consistency for regulatory compliance.

5.2.3 Interview Theme 2: Complexity in Registration Process

The registration process constitutes a comprehensive continuum that spans the entirety of the medical device life cycle, encompassing phases from research and development (R&D) through market listing, and extending into the post-market phase. This procedural framework fundamentally shapes the daily operations of professionals within this domain. Its purview extends beyond scientific considerations, incorporating facets such as device classification, laboratory testing, regulatory authority assessments, clinical trials design, animal experimentation, post-market surveillance, and renewal activities [170]. Moreover, within the realm of regulatory affairs activities, the registration process provides essential administrative and commercial insights crucial for entities involved in regulatory compliance. This encompasses considerations pertinent to Marketing

Authorization Holders (MAH), hiring Contract Research Organizations (CRO), as well as Good Manufacturing Practice (GMP) activities. Notably, the commercial dimension extends to strategic decisions. These should also conform to regulations regarding the in-house management of regulatory activities or potential outsourcing to external entities[171]. Below, we list the child nodes associated with this parent node. We have separated and categorized each child node that can be derived from the parent node concerning the registration process, including the following subsections:

Regulatory Quagmire

The interviewees express a sense of being mired in regulatory complexities due to the overarching regulations that are not sufficiently designed. This situation results in some complementary measures or regulations, as well as smaller components, not being effectively integrated into the entire framework. Summarizing from these interview narratives, there are specific issues with the regulatory framework of the NMPA. The process of formulating regulations may lack complete logic due to overarching regulations, which leads to difficulties in integrating complementary measures, regulations, or smaller components effectively. Interviewee 19 highlights a unique situation in China:

> "In China, there is a localized regulatory model where the provincial authority oversees products registered in that province. However, when surpassing this model, products produced overseas may only fall under the jurisdiction of the central NMPA. In this way, certain industrial-level demands currently remain unmet. Additionally, there are challenges with the production of products overseas under a domestic license, potentially creating a sense of territorial affiliation for the NMPA. Conversely, full imports enjoy considerable flexibility, and the design under the MAH system faces certain constraints. Registration requirements may seem somewhat illogical at times, reflecting issues in system settings. However, there has been gradual improvement in regulations and technical documents in recent years, with various aspects gradually falling into place."

The sense of a regulatory quagmire is a common sentiment across many countries. Interviewee 13 describes the complexity within the European system:

> "Because in Europe, a device is not registered with an authority. You need to go to a notified body, and the notified body would need to involve an authority, and both of them need to get an approval for a device, which is very complicated."

Echoing this complexity, Interviewee 26 points out the fragmentation within the European regulatory landscape:

> "So the other problem with Europe is that, in the FDA, it's a single agency for the entire United States, and in Europe there's 38 or so Notified Bodies now, it's like having 38 different possible clearance or approval organizations. So that causes a lot of disconnect."

These insights from the interviews underline the multifaceted challenges in navigating the regulatory environment, emphasizing the need for coherent and logical frameworks to streamline regulatory processes. This child theme encapsulates the prevailing sense of instability and unpredictability that industry professionals face while attempting to navigate the maze of regulatory frameworks. A central concern that emerged from the interviews is the ongoing ambiguity surrounding regulations such as the European Medical Device Regulation (MDR) and In Vitro Diagnostic Regulation (IVDR). The shifting deadlines, evolving requirements, and the dearth of clear guidance collectively create an environment of regulatory flux. This not only impedes companies in their efforts to comply but also raises concerns about potential compromises in patient safety and the timely availability of essential medical products. As a result, regulatory professionals find themselves in a perpetual struggle to adapt and realign strategies in response to this ever-changing landscape.

Rapid Regulatory Dynamics

The speed at which regulations are updated can be overwhelming for industry professionals. The continuous changes may result in confusion and uncertainty, making it difficult for manufacturers and practitioners to adapt swiftly. Industry experts shed light on the challenges posed by the swift and intricate updates to regulatory frameworks. Interviewee 17 offers a poignant observation on the fluidity of regulations in the EU and UK:

> "In EU and UK, you could see some regulation changes week to week, or sometimes even within the same day. It's just hard to get a kind of a straight and a consistent answer."

The European MDR and IVDR serve as primary examples, with their stipulations demanding a comprehensive understanding and swift adaptation. The interviews underscored the multifaceted nature of these regulations, which often leaves regulatory professionals grappling with the complexities that arise from this. The dynamic nature of these rules, coupled with the perpetual release of updated information, creates a pressing need for professionals to stay informed and continuously adjust their compliance strategies.

Authority Supersedes Regulation

In some instances, the expertise and authority of individuals involved may play a more dominant role in decision-making than strict adherence to regulations. This can lead to variations in interpretation and application of regulations across different entities. Interviews unveiled the pivotal role that seasoned expert and authoritative networks play in shaping regulatory compliance strategies. This transcends the influence of written regulations, as professionals often draw upon accumulated insights, historical conventions, and the guidance of respected figures in the field. The theme illuminates instances where a rigid adherence to regulatory text takes a backseat to the practical wisdom offered by those deeply entrenched in the regulatory sphere. Collaborations with experts such as medical writers, statisticians, and clinical professionals emerge as indispensable tactics to bridge the gap between regulatory mandates and practical implementation. However, concerns are raised, particularly for newcomers and startups who may face barriers in building the necessary networks and historical insights to effectively navigate this nuanced landscape.

Interpretation Discrepancies

The absence of a standardized method for interpreting regulations can lead to inconsistencies in compliance efforts, introducing varying interpretations among different stakeholders and complicating the compliance process. This lack of uniformity across diverse regulatory bodies and geographical regions emerges as a salient theme, highlighting inconsistencies in viewpoints and approaches taken by regulatory agencies. Notified bodies, responsible for assessing the conformity of medical devices in EU [172], may apply different interpretations and criteria during the certification process.

Reflecting on the challenges of international registration, Interview 13 from a company's international registration department noted:

> "In the context of international registrations, there can be disagreements and misunderstandings due to varying understanding of regulations. This can be influenced by factors such as cultural differences, as communication with counterparts from countries like the United States may be more trust-based, while interactions with peers from Germany may involve a meticulous examination of regulations. In the case of Chinese regulations, which often have multiple layers, conflicts and misunderstandings can arise if one party focuses on a layer that hasn't been clearly communicated. The key factors contributing to these conflicts are individual perspectives, cultural nuances, and the presence of biases and confidence issues. The conflicts often

> stem from different interpretations based on distinct levels of regulatory understanding between parties."

This lack of harmonization can lead to varying certification outcomes, adding complexity to the regulatory landscape. The theme underscores that this lack of alignment causes confusion and frustration among companies striving for compliance, emphasizing the necessity for more cohesive communication between regulatory bodies and industry stakeholders. A collective effort to bridge these disparities is pivotal for creating a streamlined approach to regulatory adherence. Additionally, manufacturers often acknowledge that cost considerations influence their decision-making process, leading them to opt for approaches that are less expensive.

Addressing the critical aspect of clinical studies, Interviewee 2 shared an insightful observation:

> "Manufacturers sometimes present their [clinical] study by stating they want to set the trial size at N. They don't perform a sample size calculation. So, they just pick a number of patients to cut costs. In such cases, when they present their study to the regulatory authority, the authority might question them on the lack of a sample size calculation because they didn't follow a statistical approach and provide scientific evidence. Consequently, they could face challenges. For instance, if they claim their device can perform certain functions, but their clinical study doesn't support this, they will face scrutiny on the study design and it may not be accepted."

The regulatory process for medical devices may be characterized by ambiguity and lack of clarity, which in turn presents challenges for understanding step-by-step compliance procedures and resulting in delays and potential misinterpretations of requirements.

Ambiguities in the Scope of Regulatory Requirements

The clarity of regulatory requirements for medical devices often remains ambiguous, causing confusion about applicable regulations for specific products and complicating their proper classification and evaluation. At this juncture, Interviewee 5 sheds light on the inherent challenges within the regulatory landscape:

> "Regulations never provide you with the recipe; they always give you a framework within which you need to operate. However, they don't explicitly detail what you are required to do based on the regulation. I believe it's like a decision tree determining whether something should enter the process or not. That's where companies often face difficulties,"

This insight highlights the gap between the regulatory framework and the detailed guidance needed for compliance. The challenge lies in navigating this framework, where a more defined decision-making process could alleviate the difficulties companies face, streamlining the path to product approval and ensuring safety standards are met efficiently.

5.2.4 Interview Theme 3: Complexity in Available Database

The complexity inherent in navigating available databases for regulatory information in the medical product sector presents a significant challenge, as highlighted by our interviewees. This complexity not only affects the efficiency of compliance efforts but also impacts the ability of manufacturers to remain informed about relevant regulatory requirements. The diversity and dispersion of regulatory information across various databases underline the need for a more streamlined approach to information management and accessibility.

Regulatory Database Availability

The issues around regulatory database availability have unveiled significant hurdles within the medical product sector. Interviewees pinpointed the fragmentation of regulatory information across multiple databases as a primary source of inefficiency in compliance efforts. This dispersion is further complicated by regional differences in regulatory frameworks and databases, adding layers of complexity to an already challenging landscape. Concerns were raised about the absence of a centralized, comprehensive database that consolidates all pertinent regulations, which would facilitate easier access to accurate and up-to-date information. The necessity for a harmonized and accessible regulatory database, encompassing various departments and jurisdictions, was emphasized as a means to simplify compliance processes and enable more effective navigation of regulatory requirements.

Highlighting a pivotal challenge, Interviewee 4 remarked:

> "[It] is sometimes more difficult to find the information. It's not that the information is not there. One of the issues is, in order to find something you have to know what you're looking for. So if you don't have your intended use correctly defined, you won't know what you're looking for. And then you have to know where to look, and how ... This is a systematic issue."

Similarly, Interviewee 19 addressed the difficulties in disseminating and utilizing information:

> "As a consulting firm, we're still getting a lot of very basic questions that the Commission should have answered, the IMDRF should have made a bigger effort... a lot of companies like us still need to answer the paying clients' questions, but they don't really help with the information sharing that's actually required industry-wide. They're only very specific to our clients, and that's it. So I think there's a big deficit in how people are getting the information, and how people are sharing the information."

These insights underscore the critical need for systemic improvements in how regulatory information is organized, accessed, and shared. Enhancing database availability and usability could significantly reduce the barriers to compliance, thereby supporting manufacturers in navigating the regulatory environment more efficiently.

Standard Availability

The theme of standard availability emerged as a critical aspect of the regulatory landscape for medical devices and diagnostics, emphasized by the diversity in standards across different regions and regulatory bodies. This diversity poses a significant challenge for manufacturers in ensuring compliance and understanding regulatory requirements uniformly. The interviews shed light on the value of existing standards and the gaps that necessitate continuous development and harmonization. The role of technical experts in contributing to the standardization efforts was stressed, highlighting the need for a collaborative approach to create comprehensive standards applicable to a broad spectrum of devices.

In this context, Interviewee 1 underscored the pivotal role of technical expertise in enhancing standardization efforts:

> "... technical experts should be contributing much more to standardization ... Standards are really valuable for a lot of devices, and without a standard, everyone would be lost."

This dialogue points to the essential need for a more unified approach to standards development and implementation. The emphasis on collaboration and harmonization reflects a broader consensus on the necessity to bridge the gaps in the current regulatory framework, thereby enhancing the regulatory landscape for all stakeholders involved.

5.2.5 Interview Theme 4: Complexity with Product

Numerous interviewees have noted that various products exhibit distinct levels of complexity. This theme encapsulates multiple factors influencing the regulatory compliance landscape, encompassing child nodes such as product classification, which

relates higher classification with increased compliance complexity, since products with a higher classification may entail a larger set of features, components, or functions, translating into a greater volume of technical documentation and data that must be thoroughly analyzed and validated during the registration process[173]. Furthermore, it is customary for a clinical trial to be mandated in the case of higher classification devices. In contrast, for devices with lower classifications, the requisites are less stringent, with only clinical evaluation being necessary[174].

The child node concerning the availability and duration of relevant guidance plays a pivotal role, as evidenced by Stern [160], who highlights that the presence of pertinent guidance significantly influences the approval timeline for medical devices. Interviewees within the industry emphasized that all stakeholders require a period to familiarize themselves with any new regulations, particularly if they represent a novel paradigm. Interviews also shed light on the temporal considerations of regulators, where the time a specific type of device has been on the market influences regulatory decisions.

The child nodes related to the claims made by a product, along with the number of components it includes, further shows how compliance can be affected by multiple factors. This is also a notion that is supported by prior research on product complexity and the interactions of components [175].

Highlighting the direct impact of product complexity on regulatory compliance, Interviewee 4 provided a clear perspective:

> "The more complex a design is, the harder it is to comply with the regulation, because it means that you're going to have to run more tests ... if you claim that your device does something very simple and not critical, then it's much easier to follow the steps to certify it."

This interviewee underscores an intricate web of connectivity binding these nodes. The narrative underscores that the complexity of a product does not merely influence the regulatory pathway in isolation but is part of an interconnected system where multiple elements dynamically interact. This system, shaped by both the tangible characteristics of the products and the regulatory frameworks they must navigate, highlights the importance of understanding and addressing these complexities in a holistic manner. Through this lens, the journey toward compliance is seen not as a linear pathway but as a nuanced exploration of the interplay between product features, regulatory expectations, and the broader context.

5.2.6 Interview Theme 5: Complexity at a Global Level

This facet of complexity is predominantly pertinent to global manufacturers aiming to expand market reach across different regions. The expensive nature of medical technology development also often requires manufacturers to consider multiple markets. As shown in Figure 5.1, particular emphasis is placed on the following child nodes below.

The node of “Regulatory Inconsistency", describes divergent labeling or testing requirements, among other variations. This leads to discrepancies in regulatory decisions and information supplied to medical practitioners and patients[176], which can pose risks to safety and effectiveness[177]. Regulatory inconsistency were mentioned by multiple interviewees and it has been a major concern if they wanted to move between markets. Highlighting the challenges within the EU, Interviewee 3 observed:

> "Unfortunately, the further we get into MDR for the EU, the less harmonization there is even amongst the platforms in the EU ..."

This highlights the issues that even occur within a specific region, let alone the differences that exists between e.g. the US, Japan, China, EU, and other countries. Similar occurrences are noted in other geographical contexts as well. Detailing the confusion arising from differing regulatory standards, a Regulatory Affairs Specialist, employed in the Brazil branch of a multinational company serving both Brazil and the USA, expressed:

> "So who should I follow? If they’re different, ... for example, in Brazil, if I have an ISO 13485 certification, it does not exempt me from needing a GMP from ANVISA, since they’re not the same. The ISO certificate is not a requirement for ANVISA, so should I follow what ANVISA dictates, or should I go with ISO? It’s quite confusing for us."

According to this professional’s comments, it is evident that there are various certifications required across different regions. Some of these certifications might appear to serve similar functions, yet Regulatory Affairs specialists are compelled to obtain them separately. It becomes a source of frustration when certifications with ostensibly similar functions entail different details.

5.2.7 Code analysis with NLP

Besides manual coding, NLP techniques has been deployed to uncover and decode intricate patterns embedded within textual data, focusing on the discourse among stakeholders in the medical device regulatory arena. Through a multifaceted analytical

approach, incorporating n-gram analysis, *tf-idf* (Term Frequency-Inverse Document Frequency) evaluation, and Latent Dirichlet Allocation (LDA) modeling, we meticulously analyze the nuanced language and thematic elements characteristic of this field. The use of n-gram analysis permits an in-depth exploration of the prevalent lexical choices, revealing the foundational language patterns across different stakeholder groups[178]. LDA modeling enables the discovery of underlying thematic clusters, providing a structured overview of the topics that dominate the discourse[179].

n-Gram Analysis

We conducted n-gram analysis on transcripts to discern potential indicators related to the three stakeholder perspectives. We opt for bigrams because they are less sparse than trigrams and higher n-grams, which means that they occur more frequently in the corpus and can provide more reliable statistical information with less data. Our examination involved categorizing interviewees into three distinct groups and then completing the analysis for each group individually: Regulators, Manufacturers and Consultants, as shown in Table 5.2. Certain words emerge with high frequency across the groups, such as “regulatory affairs", “medical devices", “clinical trials" and “guidance documents". For the top n-grams for regulators, the focus was on terms such as “regulatory affairs" , “clinical trials", and “guidance documents". These terms underscore a regulatory focus, with an emphasis on compliance. Additionally, terms like “research development", “technical requirements", and “technology transfer" suggest a strong orientation towards research and technological aspects within the regulatory framework. The n-gram analysis for manufacturers unveils a prominent emphasis on terms like “regulatory affairs", “medical device" and “clinical trial." This emphasizes a potential dual commitment to regulatory compliance and product development. Noteworthy is the recurrence of terms like “make sure", “notified bodies" and “quality assurance," indicating a stronger focus on ensuring product quality and compliance with regulatory standards. As for consultants, frequent terms like “medical device", “regulatory affairs" and “United States" . Notably, terms such as “United States", “class III" and “registration process" suggest a specific focus on regulatory classifications and registration procedures.

In examining the linguistic patterns among regulators, manufacturers, and consultants within the regulatory affairs domain, a distinct clustering of bi-gram frequencies is evident, which reveals the nuanced priorities and focal points of each stakeholder group. The term "regulatory affairs" consistently registers high frequency across all groups, underscoring its overarching importance in the industry’s discourse. The prominence of

"clinical trials" among regulators and consultants, with a strong but lesser emphasis by manufacturers, suggests a shared interest in the intricate processes of medical product testing and validation. Conversely, manufacturers demonstrate a particular focus on the term "medical device," reflective of their direct involvement in product development and commercialization. The presence of terms such as "notified bodies" in the manufacturers' and consultants' lexicons, as opposed to their absence among regulators, highlights the practical aspects of market access and the operational role these entities play in product certification. The inclusion of geographically and product-class specific terms such as "united states" and "class III" by consultants further indicates the specialized nature of consultancy work, delineating its advisory scope. Furthermore, the use of more subjective terms like "make sure" and "feel like" by manufacturers may suggest a discourse characterized by the intent to assure product quality and compliance, thus reflecting a more pragmatic approach. Collectively, these linguistic patterns not only mirror the interrelated yet distinct responsibilities of each group in the regulatory ecosystem but also accentuate their collective commitment to ensuring compliance and safeguarding public health.

Table 5.2: Top 10 Bi-grams and their frequencies for Regulators, Manufacturers, and Consultants

Regulators		**Manufacturers**		**Consultants**	
Bi-gram	**Freq.**	**Bi-gram**	**Freq.**	**Bi-gram**	**Freq.**
regulatory affairs	10	regulatory affairs	47	medical device	51
clinical trials	10	medical device	44	regulatory affairs	37
guidance documents	6	clinical trial	31	united states	35
research development	6	guidance documents	24	clinical trials	32
technical requirements	6	make sure	20	conduct clinical	30
conduct clinical	6	medical devices	19	notified body	29
language used	6	things like	19	medical device	27
technology transfer	5	notified bodies	17	class III	21
regulatory authorities	5	feel like	17	registration process	20
medical products	4	quality assurance	17	notified bodies	20

Latent Dirichlet Allocation Model

In our study, we utilized the Latent Dirichlet Allocation (LDA) model [180], a powerful technique for categorizing textual content into distinct topics. Applying this model to our transcript data revealed recurrent keywords such as “product,” “regulatory,” “device,” “FDA,” and “regulation” across the identified thematic groups.

To address the variation in the number of experts among three groups—regulators, manufacturers, and consultants—we incorporated TF-IDF (Term Frequency-Inverse Document Frequency) analysis into our LDA modeling process. We compiled a collection of documents for each stakeholder group and extracted bi-grams from these documents. We then calculated TF-IDF scores for these bi-grams to assess their importance within each group, taking into account their frequency across all documents. The TF-IDF scores were visualized using a heatmap, where the color intensity indicates the normalized TF-IDF scores, allowing for an easy comparison of keyword importance across groups. This approach underscored the unique and shared linguistic patterns among the groups.

Determining the optimal number of topics (k) for our analysis required a meticulous iterative process. We developed and evaluated numerous Latent Dirichlet Allocation (LDA) models, each configured with a different number of topics. Through this rigorous assessment, guided by the coherence scoring methodology established by Hasan et al. (2021) [181], we ultimately determined that a model configuration with exactly k = 7 well-defined topics best achieved the desired balance of coherence. This configuration allowed us to identify seven distinct topics within our dataset, marking a significant finding in our analysis.

Utilizing the capabilities of the Gensim Python library [182], we crafted an Inter-Topic Distance Map, and we note down each topic top terms and tokens frequency. Delving deeper into this, we notice the close relational ties between certain topics, notably between Topic One and Topics Three and Four. Topic One, for instance, is characterized by a frequent amalgamation of terms such as "product," "device," "FDA," "regulatory," and others, echoing its primary focus. Our analysis progressed as we extracted and analyzed the most frequently occurring terms from each topic cluster, meticulously curating a set of terms that succinctly encapsulate the thematic core of each topic. This led to the assignment of descriptive titles to each topic cluster, effectively summarizing their thematic concentration. These designations, along with their respective key terms, are methodically detailed in Table 5.3, offering a structured overview of our thematic findings.

The spatial dynamics between each topic cluster on the map act as a nuanced indicator of thematic affinity or disparity as shown in Figure 5.2. Topics in close proximity suggest a thematic overlap or shared focus, potentially highlighting similar concerns or areas of regulatory discourse. For example, the proximity of Topics 1 (Product Compliance and Regulatory Oversight) and 2 (Diverse Product Regulation and Compliance) on the map underscores a shared focus on navigating the complexities of FDA regulations and product compliance. This proximity not only sheds light on the interconnectedness of regulatory themes but also illustrates how these thematic elements weave together to form the complex tapestry of medical device regulation.

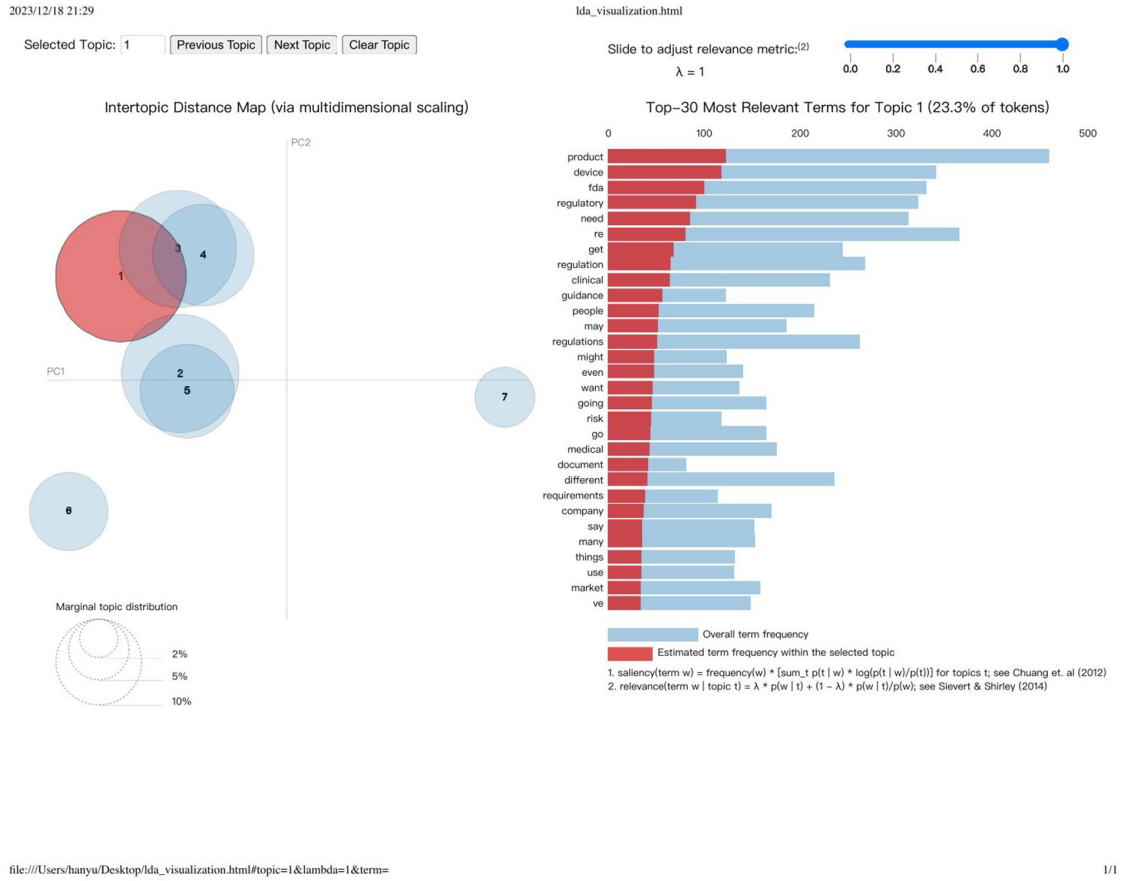


Figure 5.2: Inter-topic distance map (produced with pyLDAvis). Each circle is a topic; the distance between circles approximates topic similarity via multidimensional scaling of the topic–term distributions. The expression at the foot of the figure is the pyLDAvis term-*relevance* metric, $r(w, t \mid \lambda) = \lambda\, p(w \mid t) + (1 - \lambda)\, p(w \mid t)/p(w)$, in which $\lambda \in [0, 1]$ trades off a term's probability within a topic against its overall corpus frequency when ranking the terms most characteristic of each topic.

In contrast, topics that find themselves situated at greater distances from one another denote a thematic divergence, spanning a wide array of regulatory discussions from clinical data analysis and device requirements to manufacturing standards and

compliance. Such variations underscore the diverse nature of conversations within the regulatory landscape, highlighting the extensive range of challenges and focus areas within the domain. Through this refined analytical lens, our study navigates the intricate maze of medical device regulation, delivering insights that are both deep and broad.

Topic	**Top terms**	**Topic proportion (%)**
Topic 1: Product Compliance and Regulatory Oversight	product, device, FDA, regulatory, need	23.3%
Topic 2: Diverse Product Regulation and Compliance	product, different, FDA, right, regulations	18.9%
Topic 3: Clinical Data Analysis and Device Requirements	people, get, data, FDA, regulatory	18.7%
Topic 4: Regulatory Landscape for Medical Devices	product, get, people, regulations, device	14%
Topic 5: Broad Regulatory Framework in Clinical Context	product, regulation, devices, clinical, need	11.9%
Topic 6: Regulatory and Clinical Product Dynamics	regulatory, product, need, device, clinical	8.4%
Topic 7: Standards and Compliance in Device Manufacturing	device, FDA, need, standard, product	4.9%

Table 5.3: Results of topic modelling showing the top terms and proportion of each topic. The percentage indicates the proportion of the corpus attributed to each topic by the LDA model.

5.3 Conclusions

In this section we discuss the results of our work and contextualise the findings within larger issues of medical device regulation. We also identify limiations of our approach as well as possibilities for future work.

5.3.1 The wider context of the interview findings

Regulations can affect various aspects of product approval and compliance. They are an essential component of the medical device industry. Due to various reasons, experts have found that regulations are becoming more complex, as shown by previous research [15]. More importantly, complicated regulation increases the cost for industry and weakens the innovation process [6]. Understanding complexity is therefore essential, as it could aid decision making during the product approval process. Complexity in a regulatory environments has been previously investigated in other domains, such as financial regulation Easley and O'Hara [183], where they looked at the ambiguity

of regulations. Muchmore [184] developed an analytic framework for understanding the role of uncertainty in regulatory design. More recent ideas suggest the existence of multiple dimensions of complexity within the medical device regulatory landscape. These complexities, if left unaddressed, could hinder the development, approval, and market access of medical devices [185]. Understanding and mitigating these complexities are essential for creating a conducive environment for innovation and ensuring timely and safe access to medical technologies for patients. Additionally, the understanding of complexity can be leveraged to design easy-to-use tools which can expedite decision making processes. For example, Ceross and Bergmann [186] developed a highly accurate text classification model which reduced the complexity of navigating Australian medical device regulation. Our interviews show that tools of this type are needed within the medical device regulation domain, although there are few available.

Our investigation revealed five primary origins of complexities that warrant consideration for mitigation: legal language complexity, registration process intricacies, database complexities, product-level intricacies, and complexities on a global scale. The coding structure involved a hierarchy, with child nodes extending from parent nodes. For instance, a lack of an appropriate database can catalyze complexity in the registration process and impedes progress due to insufficient information. This underscores that the challenges in medical device regulatory affairs constitute a system of systems problem, rather than a singular issue [187]. The escalating complexity of modern systems further exacerbates these challenges, necessitating additional research and the development of innovative approaches to effectively address these intricate issues.

The n-gram analysis unveiled nuanced distinctions in language usage across various stakeholders. Moreover, this analytical method could offer actionable insights that can help facilitate communication between different stakeholder and foster collaborative solutions tailored to the unique challenges and opportunities within each of the areas of interest.

In the realm of NLP and open coding, particularly when examining LDA-generated themes (topic names), we find notable parallels with themes unearthed by the manual coding approach. Take, for example, Topic 2, “Diverse Product Regulation and Compliance," which mirrors the theme of product complexity identified with the manual coding. In a similar vein, Topic 3, “Clinical Data Analysis and Device Requirements," echoes the theme of registration process complexity. Moving to Topic 6, “Regulatory and Clinical Product Dynamics", this resonates with the rapidly evolving dynamics we identified in open coding themes. Yet, when relying solely on LDA outcomes, it is challenging to

trace these results back to open coding themes, as the latter tend to be more nuanced and explanatory. Whilst NLP findings provide valuable insights and guideposts for additional manual coding, it is crucial to recognize that, with the current state of technology, human-led manual coding remains more accurate and indispensable [188].

The European Union's medical devices framework strives to establish a harmonized and robust system that encourages innovation while prioritizing public health. This framework outlines a comprehensive set of requirements and guidelines for manufacturers to adhere to, ensuring the quality and performance of medical devices [189]. Within the EU, the regulatory landscape for medical devices is governed by specific directives and regulations. The engagement of notified bodies, independent third-party organizations designated by regulatory authorities, is crucial in the conformity assessment process [190]. However, insights garnered during the transcript process reveals notable concerns and issues. The primary problem raised is the inconsistency among the procedures and requirements for each notified body [191]. Interviewees also express apprehension about the extended waiting times for scheduled testing with testing labs in the EU, indicating that, currently, it takes years to secure a testing slot.

5.3.2 Limitation

The inherent nature of qualitative research often involves working with smaller sample sizes [192]. The results derived from our study are intrinsically linked to the specific contexts and experiences of our participants. While our participant cohort encompasses a geographically diverse array from the United States, European Union, and various Asian countries, providing a broad cultural perspective, the insights and themes extrapolated from our interviews reflect the viewpoints and experiences of this particular group, and caution should be exercised when attempting to extrapolate these findings to a wider population. This limitation underscores the necessity for additional research, potentially incorporating a more extensive and varied participant base, to further validate and expand upon our findings.

Additionally, our exploration is confined to employing two natural language processing (NLP) methods, which are the n-gram and LDA modeling. Divergent results may emerge when employing alternative NLP methodologies, underscoring the importance of recognizing potential variations in analytical outcomes among different researchers. While Bidirectional Encoder Representations from Transformers (BERT) stands out for its efficacy on larger datasets[193], our study, constrained by the available data, exclusively utilizes n-gram and LDA modeling to better fit the size of our dataset. A

prospective avenue for research involves the consideration of BERT, particularly with an expanded dataset in the future.

To our knowledge, this is the first qualitative study that explores factors to regulatory affairs complexity. This study has some limitations that can be addressed in future studies. Foremost, the research primarily focus on US, EU and China, potentially limiting the generalizability of findings across a broader spectrum of regulatory contexts. Thus, future studies can be extended to other areas. Subsequently, the qualitative nature of this study allows for a comprehensive exploration of regulatory complexities, yet quantifying the prevalence and impact of specific challenges may require a complementary quantitative approach. Utilizing quantitative surveys or metrics could offer a more precise understanding of the relative significance of identified complexities. Therefore, future research can use quantitative design to explore the influencing degree of these factors on complexity of regulatory affairs. Lastly, the study focuses on the present state of regulatory affairs and the challenges they entail. However, regulatory environments are subject to evolution and reform, potentially influencing the nature and intensity of complexities over time. To capture the dynamic nature of regulatory affairs, longitudinal studies could offer insights into trends and changes within this intricate landscape.

5.3.3 Future work

The indispensable role of regulations in the medical device sector underscores a complex landscape where development and distribution of products are intricately governed. Despite the critical importance of these regulations, a research gap persists in fully understanding the specific challenges and opportunities they present. Our investigation has illuminated various sources of complexity within the regulatory framework, pinpointing obstacles such as regulatory language, procedural intricacies, global challenges, database management issues, and product-specific nuances. These complexities affect a wide range of stakeholders, including manufacturers, regulatory agencies, and healthcare professionals, highlighting the need for an in-depth understanding and strategic approaches to navigate these challenges efficiently.

To address these complexities, our study employs a dual-methodological approach, integrating manual coding with Natural Language Processing (NLP). Manual coding allows for a nuanced understanding of the regulatory discourse, enabling researchers to capture the subtleties and complexities of regulatory language and processes. It provides a deep dive into the qualitative aspects of regulatory challenges, offering rich, detailed insights that automated methods might overlook. On the other hand, NLP

offers the ability to analyze large datasets, uncovering patterns and trends at scale that manual coding alone cannot feasibly identify. By employing NLP, we can systematically assess the frequency and context of specific regulatory challenges across a broad corpus, providing a comprehensive overview of the regulatory landscape.

The synergy between manual coding and NLP methodologies enriches our analysis, ensuring both depth and breadth in understanding the regulatory challenges within the medical device sector. This combination allows for a more holistic view, leveraging the strengths of both qualitative and quantitative analyses to uncover the multifaceted nature of regulatory complexities.

Looking forward, the integration of regulatory affairs with data science offers a promising avenue for simplifying regulatory processes [194]. The advent of Artificial Intelligence in Regulatory Affairs (AIRA) and the potential application of Large Language Models (LLMs) [195] signal a transformative shift towards leveraging AI to enhance regulatory frameworks. Such technological advancements promise to streamline regulatory procedures, improving efficiency and ensuring that regulations remain adaptive to the fast-paced evolution of medical technologies.

By utilizing both manual coding and NLP, we comprehensively examine the text in order to develop informed insights into the complexity of regulatory frameworks that effectively address stakeholder needs. This comprehensive approach facilitates a deeper understanding of existing regulatory challenges faced by stakeholders. As we venture into this new era of regulatory affairs, the insights gained from our dual-methodological analysis will be instrumental in fostering an environment conducive to innovation while upholding the highest standards of safety and efficacy.

Furthermore, the advent of Large Language Models (LLMs) [195] opens new frontiers for regulatory affairs, with potential in harnessing AI for evaluating medical products and policies. In our own interviews, the topic of LLMs did not appear and there have been no other studies to date as to how medical device regulation stakeholders view this. This topic will likely need monitoring as the ease of use and availability of LLMs becomes more ubiquitous.

Overall, it is imperative to clearly understand the complexities present in the regulatory affairs landscape. Recognizing these challenges provides the foreground from which to develop tools and methods by which to navigate and mitigate these complexities. This approach not only promises to streamline regulatory processes but also ensures that the regulation of medical devices remains robust, responsive, and conducive to innovation.

Having examined the regulatory challenges from experts words, we now shift our focus to the functional capabilities of emerging AI tools. In next Chapter 6, we evaluate whether large language models—currently among the most transformative technologies—can be effectively applied to downstream tasks in regulatory affairs, with a particular emphasis on medical device classification. This investigation serves as a proof of concept to assess the feasibility of leveraging such models to support and potentially streamline compliance workflows. If foundation models demonstrate strong performance in regulatory tasks, they may hold significant potential for future integration into authoritative regulatory practices.

6

Evaluating Precision: The Role of Large Language Models in Medical Device Classification

This chapter is based on: Yu Han; Aaron Ceross; Florence Bourgeois; Paulo Savaget; Jeroen H. M. Bergmann, Evaluation of large language models for the classification of medical device software, Bio-Design and Manufacturing, 7:819–822, 2024 [196]

Author contributions (CRediT, as published in Bio-Design and Manufacturing*): Yu Han — Methodology, Formal analysis, Data curation, Visualization, Writing – original draft, Writing – review & editing; Aaron Ceross — Data curation, Writing – review & editing; Florence Bourgeois — Project administration, Writing – review & editing; Paulo Savaget — Supervision, Writing – review & editing; Jeroen H. M. Bergmann — Conceptualization, Methodology, Data curation, Project administration, Supervision, Writing – original draft, Writing – review & editing.*

6.1 Introduction

As highlighted in the previous chapter, the complexity of regulatory implementation cannot be fully understood through policy or textual analysis alone—it also involves practical challenges faced by the regulatory affairs professionals interviewed in Chapter 5.3.3, namely manufacturers, regulatory consultants and regulators. These challenges include interpreting ambiguous or obscure regulatory language, determining the correct risk classification of a device, and reconciling inconsistent requirements across jurisdictions. In this chapter, we build on that insight by investigating whether emerging

AI technologies—specifically large language models (LLMs)—can help address one of these challenges directly, namely medical device classification.

The rise of LLMs signifies a transformative leap in artificial intelligence, heralding new opportunities across diverse sectors. From day-to-day applications to fields like healthcare that demand strict regulatory compliance, LLMs bring unparalleled prowess in understanding, generating, and navigating through human language[197]. Their ability to sift through and analyze extensive textual data makes them ideally suited for intricate tasks, such as the vital role of classifying medical devices. This skill is especially crucial within the complex regulatory framework of the healthcare industry, where precise device classification can greatly enhance the efficiency and reliability of regulatory processes.

LLMs are sophisticated tools trained on massive datasets, including billions of words sourced from the internet, books, and various texts. They have acquired an intricate understanding of language patterns, context, and semantics. Operating on the principle of predicting the next word in a sequence based on the preceding context, LLMs utilize probabilities learned during their training. This seemingly simple process is underpinned by advanced neural network architectures and significant computational power, allowing LLMs to tackle a broad spectrum of language tasks.

In the healthcare sector, regulatory affairs play a pivotal role, overseeing the lifecycle of medical devices, pharmaceuticals, and other healthcare products from development and testing to manufacturing and market introduction. This domain is essential for ensuring that these products adhere to the rigorous standards set by global health authorities, including the FDA (United States), NMPA (China), and EU (Europe). Regulatory affairs professionals are tasked with navigating this intricate regulatory environment, guaranteeing that products fulfill the strict safety and efficacy criteria before market entry. Given the increasing complexity of regulatory affairs, highlighted in recent research [15] [198], our study aims to investigate the effectiveness of LLMs—the forefront of contemporary AI tools—in addressing the challenges within this crucial field.

The classification of medical devices is a critical first step in the regulatory oversight process, establishing the degree of scrutiny a device will undergo before it can be approved for market. This classification hinges on the device's intended use, associated risks, and potential impact on patients and users. It is a process that deeply concerns manufacturers, regulatory bodies, and healthcare providers alike because it directly influences the approval timeline, post-market monitoring, and, most importantly, patient safety. Ensuring that devices are subjected to the appropriate level of testing and

evaluation is vital for mitigating risks and protecting public health. The rapid pace of technological advancements in the medical field, along with the increasing complexity and variety of devices, highlights the need for accurate and adaptable classification systems. Both regulatory authorities and Regulatory Affairs (RA) professionals working for companies pay close attention to classification. Misclassification can have significant repercussions. For instance, if a device is mistakenly classified at a higher level than appropriate, the manufacturer may face unnecessary testing and trials, elevating costs and potentially placing them at a disadvantage compared to competitors who can bring their devices to market more economically. Conversely, a lower-than-appropriate classification can lead to rejection by regulatory authorities, preventing the device from reaching the market at all. Additionally, for manufacturers aiming to enter global markets, navigating the varied classification systems of different regulatory bodies can significantly affect costs. Thus, accurate classification is not just a regulatory formality; it is a strategic consideration that impacts the financial and operational aspects of bringing medical devices to the global market.

In this study, we focus on evaluating LLM performance on classification of AI medical devices. Our emphasis on Artificial Intelligence (AI) devices specifically stems from their complexity and the regulatory challenges they present. The dynamic nature of AI technology, often described as a "black box" due to its opaque internal workings, poses unique challenges for Regulatory Affairs (RA) authorities [199]. The rapid evolution of AI in medical devices complicates the establishment of consistent regulatory standards. This ambiguity raises questions about how well current regulatory frameworks can adapt to and effectively manage the oversight of AI devices. Given this backdrop, our study seeks to understand the potential role of Large Language Models (LLMs) in aiding the classification and regulation of these technologically advanced devices. We are particularly interested in exploring how LLMs could contribute to demystifying the "black box" of AI devices for regulatory authorities and ensuring that these innovative technologies are safely and effectively integrated into healthcare.

6.2 Methods

Our study embarked on a systematic examination of medical devices across key global markets, including China, the United States, and Europe. The selection of devices for this study was strategic, aiming to cover a broad spectrum of device types and regulatory environments. For instance, the focus on software-based devices in China

was prompted by the country's recent push towards digital health solutions. The device list for the United States was extracted from Benjamens et al. (2020) [73], which provided an extensive overview conducive to a focused investigation within American regulatory databases. European device selection was guided by the listings in Van Dijck et al. (2021) [200], chosen for their relevance in the current European regulatory scenario. Our methodology involved querying several large language models (LLMs) for device classifications, comparing these with the official classifications from respective regulatory databases—each verified by authoritative regulatory professionals.

In our selection of LLMs, we applied criteria aimed at ensuring accessibility and usability for a broad audience, including non-computer science professionals:

1. Availability of an interactive website, eliminating the need for deployment, special setups, or coding, and

2. Cost-free usage, excluding models like GPT-4 due to its usage fees.

Consequently, we selected seven widely used publicly accessible LLMs for our study: GPT-3.5, Gemini, Qianwen, ERNIE Bot, Baichuan, ChatGLM, and Mistral Large. These models, hailing from China, the United States, and the European Union, represent the pinnacle of AI innovation within these significant regions. This diverse selection aims to provide an extensive insight into the global advancements of LLM technology. We meticulously document the unique attributes of each model in Table 6.1 for a thorough comparative analysis.

Transformer model forms the foundational architecture for all Large Language Models (LLMs). For instance, OpenAI's GPT-3.5 symbolizes a significant leap in generative pre-trained transformer technology, showcasing its versatility across a broad spectrum of applications, from engaging conversational agents to generating complex content. According to the table, GPT-3.5 boasts 175 billion parameters and demonstrates a wide application potential, notably achieving a 70.0% accuracy in the Massive Multitask Language Understanding (MMLU) benchmark, reflecting its competence in solving math problems, answering sensitive questions, generating code, and visual reasoning. MMLU is a widely used benchmark for evaluating language models across 57 diverse tasks spanning subjects such as mathematics, history, law, and medicine. A higher MMLU score indicates stronger general knowledge and reasoning capabilities. This score reflects GPT-3.5's competence in solving math problems, answering sensitive questions, generating code, and visual reasoning [201]. The Gemini model [202], with its 32.5 billion parameters and training on 3.3 trillion words, achieves a remarkable

90.0% accuracy in MMLU after undergoing Instructional Fine-tuning. This highlights Gemini's dual-purpose design for both generative and understanding tasks, offering researchers and developers a versatile AI tool. Models like Qianwen[203], Baichuan[204], and ChatGLM[205] , specifically designed to cater to the subtleties of the Chinese language and culture, showcase parameter sizes of 720 billion, 130 billion, and 60 billion, respectively. These models represent the focused effort to adapt LLMs to the linguistic and cultural context of China. ERNIE Bot[206], developed by Baidu, with its 2600 billion parameters and a complex training regimen that includes supervised fine-tuning and enhancements in knowledge, retrieval, and dialogue, shows exceptional performance in tasks such as text classification and summary generation. Mistral Large [207] focuses on efficiency and scalability, featuring 320 billion parameters and a zero-shot learning approach, achieving an 81.2% accuracy in MMLU. This underscores the move towards more sustainable and broadly accessible AI technologies.

In our methodology, we first aimed to establish a robust dataset of medical devices by sourcing detailed information and regulatory classifications for 130 distinct products from the databases of both the National Medical Products Administration (NMPA) of China and the U.S. Food and Drug Administration (FDA), and EU database. This dataset served as the ground truth against which we could evaluate the performance and reliability of various large language models (LLMs), whole dataset is provided in Appendix 1 of Han et al. [196].

Prompting Process

In our study, we initiated the inquiry by prompting the Language Model (LLM) : "Assume you are the role of a Regulatory Affairs (RA) professional, what is the classification for product XXX?" Initially, GPT tends to decline to provide a direct response, citing limitations in accessing real-time external databases. However, when pressed further, it generally provides insights. For instance, when asked about software designed to assist in the evaluation of brain ischemia images, GPT indicated that such software could typically be categorized as medical devices, specifically medical diagnostic software intended to aid physicians in diagnosing brain ischemia based on imaging data. It pointed out that in many jurisdictions, such software falls under medical device regulations, subjecting it to rigorous safety, efficacy, and quality standards. Nevertheless, GPT underscored the importance of recognizing potential variations in classification across

different regulatory bodies and jurisdictions, suggesting consultation with regulatory experts or local medical device regulatory authorities for precise classification.

When we directly asked GPT for classifications without implying that it assumes the role of a regulatory affairs professional, we noticed that GPT typically offered a response, as detailed in Table 6.1. Our initial hypothesis was that identifying ourselves as professionals might lead GPT to seek answers from professional websites, which often lack sufficient information for it to formulate a response. Alternatively, positioning GPT in the role of a professional might make it more cautious, possibly due to an implied responsibility associated with providing serious answers. This cautiousness could stem from GPT's programming to avoid making definitive statements in areas where it lacks direct access to up-to-date or specific regulatory information. The perceived cautiousness can also stem from the model's programming to recognize the limits of its knowledge or the complexity of certain queries. For example, in domains like legal, medical, or regulatory affairs, where incorrect information could have serious implications, GPT might be programmed to indicate uncertainty or recommend consulting a human professional [208], reflecting a design choice to minimize the dissemination of potentially inaccurate or harmful advice.

Interestingly, our investigation into classifications for the European Union (EU) and the United States (USA), employing both Chinese (CH) and English (EN) queries, yielded consistent results. We hypothesized that language choice did not significantly influence the response, indicating a degree of consistency across linguistic domains.

If an answer diverges from the established ground truth, we label it as incorrect. In particular, if the response is vague—for instance, suggesting that a product could fall into either Class II or Class III without precisely matching the ground truth—we also deem this as a misclassification. Consequently, we simplify the problem from determining specific classes to a straightforward binary determination: the answer is either correct or incorrect.

6.3 Results

The performance overview of these models are detailed through the comparative analysis in Table 6.1. Those parameter are searched on their website, which we will compare with our devices results with later. We can see that from model size (Parameters), ERNIE Bot (2600B) and GPT-3.5 (1750B) are among the largest models, suggesting they might be capable of understanding and generating more complex and contextually

Model Name	Number of parameters(billion)	Fine-Tuning Method	Training Data(TB)	Accuracy Reported on Website
GPT-3.5	1750	RLHF based on human feedback	Not disclosed (GPT3.0: 45TB)	• Solving math problems: 86.8% • Answering sensitive questions: 8.0% • Code generation: 2.0% • Visual reasoning: 12.2% • MMLU: 70.0%
Gemini	32.5	Instructional Fine-tuning	3.3 trillion words	• MMLU: 90.0%
Qwen	720	Supervised	3 trillion tokens	• MMLU: 77.4%
ERNIE Bot	2600	Supervised fine-tuning, human feedback reinforcement learning, prompts, knowledge enhancement, retrieval enhancement, dialogue enhancement	4TB	• Text classification: 88.9% • Summary generation: 82.9%
Baichuan	130	SFT; RLHF	2.6 trillion tokens	• MMLU: 59.17%
ChatGLM	60	Supervised	40	• MMLU: 47.86%
Mistral Large	320	Zero-shot	65	• MMLU: 81.2%

Table 6.1: Comparative Analysis of Language Models

appropriate text. However, their size also means they require more computational resources. Models like Gemini (32.5B), Mistral Large (320B), and others offer a balance between capability and computational efficiency. These models might be more suitable for users with moderate computational resources. For fine-tuning Method, ERNIE Bot uses a mix of supervised fine-tuning and reinforcement learning from human feedback,

potentially making it more adaptable to domain-specific tasks requiring fine-grained distinctions. In contrast, Mistral Large's zero-shot learning capability might make it suitable for applications where training data is limited or not available. For accuracy performance, Gemini stands out with 90.0% accuracy in Massive Multitask Language Understanding (MMLU), indicating strong general language understanding capability. ERNIE Bot shows high accuracy in text classification (88.9%) and summary generation (82.9%), suggesting strengths in these specific areas.

For each device accuracy result, details are shown in Figure 6.1. We have visualized it as a bar plot in Figure 6.2.

Figure 6.1: Comparative Analysis of Large Language Model Performance in Medical Device Classification: Ground Truth Versus Predictive Accuracy Across Diverse Models

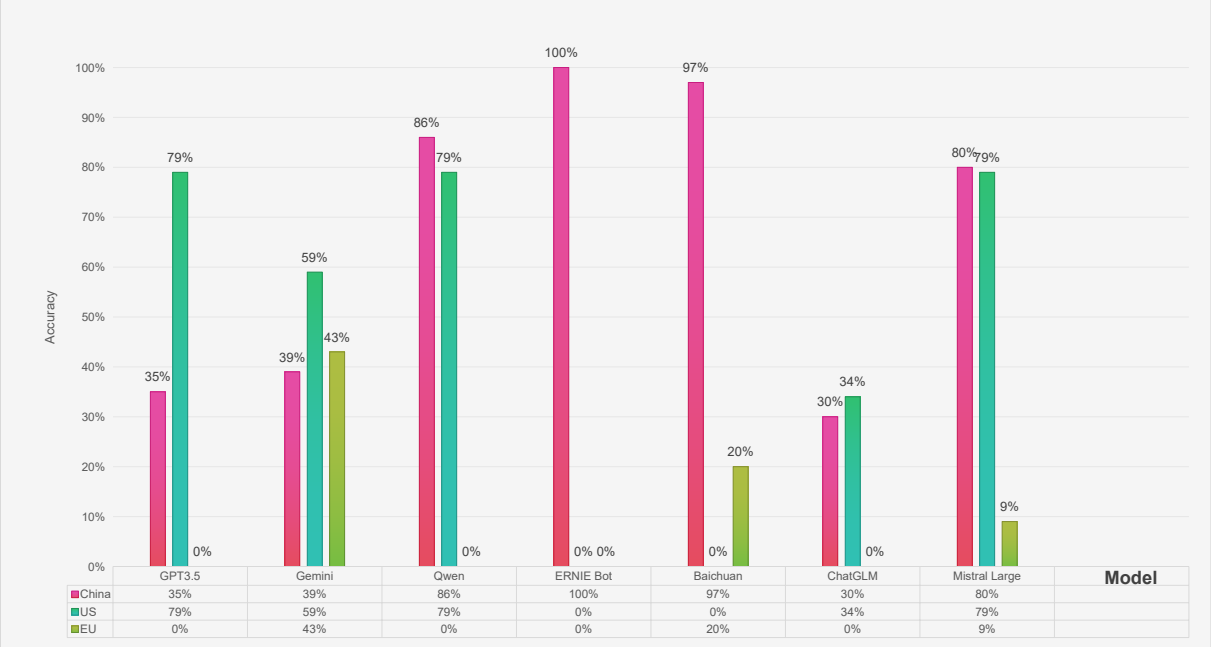


Figure 6.2: Accuracy Results Visualization

In our analysis of the performance of various Large Language Models (LLMs) across different regions, including China, the US, and the EU, a distinct pattern emerged, illustrating each model's capability to accurately classify medical devices. GPT-3.5 demonstrated a 79% accuracy rate in the US and 35% in China, but struggled in the EU with a 0% accuracy rate. This variance underscores the model's regional adaptability challenges, particularly in the European regulatory context. Gemini showcased a more consistent performance across the board, with a 39% accuracy rate in China and 43% in the EU and slightly higher in the US at 59%. Its relatively balanced performance highlights its versatility, though it indicates room for improvement in classification tasks. Qwen performed strongly in China with an 86% accuracy rate and in the US with 79%, demonstrating its effectiveness across different regulatory environments. ERNIE Bot stood out with an impressive 100% accuracy in China, suggesting an exceptional understanding or specialization in that region. However, it did not show any performance data for the US and the EU, which might point to limitations or a focused deployment strategy. Baichuan and ChatGLM displayed varied results, with Baichuan achieving a 97% accuracy in China and ChatGLM showing moderate performance with 30% and 34% accuracy in China and the US, respectively. Their performances indicate a strong regional focus and potential areas for growth in international adaptability. Mistral Large achieved 80% accuracy in China and 79% in the US, but showed a significant drop in the

EU with only a 9% accuracy rate. This pattern suggests Mistral Large's effectiveness may be influenced by the regulatory complexities of the European market.

From countries perspective, for China, ERNIE Bot achieved 100% accuracy on the Chinese device sample (n=66), suggesting strong performance in its primary linguistic domain. However, this result should be interpreted with caution given the limited sample size; Qwen also performed admirably with an 86% success rate. In the US, the performance was more varied, with GPT-3.5, QWen and Mistral Large sharing the lead at 79% accuracy. The performance of Gemini, at 59%, was respectable but highlighted room for improvement. Baichuan and ERNIE Bot's failure to classify any US devices correctly suggests a significant misalignment with the characteristics of US devices or possible gaps in their training data. The EU proved challenging for all models, where Gemini's top score of 43% contrasted starkly with the complete ineffectiveness of ERNIE Bot, Qwen, and ChatGLM, each at 0% accuracy. These results suggest that models trained on data from specific regions may not generalize well to others, underscoring the necessity for a diverse and representative dataset for model training.

6.4 Discussion

Gemini, which generally excels with a 90.0% accuracy in MMLU, indicating robust general language understanding, displayed variable performance in the regional trial. Its accuracy ranged from 39% in China and 43% in EU to 59% in the US. This contrast suggests that while Gemini possesses strong overall language capabilities, its effectiveness in medical device classification can vary significantly by region, possibly due to differences in linguistic nuances or regulatory terminology. ERNIE Bot's exceptional 100% accuracy in China aligns with its high general performance in text classification (88.9%) and summary generation (82.9%), highlighting its specialization and effectiveness in tasks requiring deep contextual understanding, especially within its primary linguistic domain. However, the lack of performance data for the US and the EU in our trial could indicate a narrower focus or the need for further adaptation to these markets.

GPT-3.5 shows a stark disparity between its regional performance and its expected capabilities. It achieved a 79% accuracy rate in the US and 35% in China, but failed entirely in the EU with 0% accuracy. This suggests limitations in its adaptability to European regulatory contexts, which may be attributable to the higher complexity of EU medical device regulations compared to other jurisdictions. The varied results of Baichuan and ChatGLM, with Baichuan achieving high accuracy in China and ChatGLM

showing modest performance, underline the importance of localized training data and the challenge of cross-regional applicability. This is particularly noteworthy when considering the potential linguistic and regulatory complexities that might influence model performance across different markets. Mistral Large achieved strong performance in China (80%) and the US (79%), but struggled significantly in the EU with only 9% accuracy. This pattern—where EU performance lags behind other regions—was consistent across most models tested, suggesting that the complexity of European regulatory standards poses a particular challenge for current LLMs.

This methodical approach allowed us to deploy LLMs for the evaluation of device classifications in line with specific regulatory frameworks, revealing insights into the models' effectiveness and limitations. Initial findings, offer a snapshot of LLM performance in this essential regulatory function. While preliminary, these results lay the groundwork for further analysis and discussion regarding the application of LLMs in regulatory affairs. This study not only sheds light on the potential of LLMs to enhance regulatory compliance procedures but also calls attention to the need for ongoing advancements in AI to meet the high standards of accuracy and reliability required in the healthcare regulatory domain. In essence, while general language understanding capabilities (as reflected in tasks like MMLU, text classification, and summary generation) provide a valuable baseline for assessing LLMs, our trial results emphasize the critical role of regional specificities in determining model effectiveness for specialized applications like medical device classification. This underscores the need for tailored approaches, including region-specific training and fine-tuning, to fully leverage LLMs in global regulatory contexts.

When selecting a Large Language Model (LLM), it is essential to align the model's capabilities with the specific requirements of the intended application. For example, GPT-3.5 is renowned for its proficiency in solving mathematical problems, making it an excellent choice for applications that demand high-level computational reasoning. Conversely, for projects requiring a broad understanding of language across diverse tasks, the Massive Multitask Language Understanding (MMLU) scores of models like Gemini, which exhibits exceptional performance, should be a key factor in your decision-making process.

Moreover, the computational demands of larger models such as ERNIE Bot, which boasts extensive capabilities, necessitate a careful assessment of your available computational resources. The adaptability of a model to specialized tasks, facilitated by advanced fine-tuning methods like those employed by ERNIE Bot, is another crucial

consideration, particularly if you possess the means to engage in custom training or feedback sessions. Additionally, for projects with a focus on specific linguistic or cultural contexts, models like Qwen, Baichuan, and ChatGLM, which have been meticulously trained on relevant data, may present significant advantages in understanding and generating language that aligns with cultural nuances.

Choosing the optimal LLM thus requires a comprehensive evaluation of these factors: the model's task-specific performance, computational resource demands, adaptability through fine-tuning, and cultural or linguistic specificity. This holistic approach ensures the selection of an LLM that not only meets your immediate project needs but also aligns with your long-term goals and resource constraints, enabling successful and efficient task execution.

6.5 Limitation

This study specifically targets AI devices due to practical constraints, notably the considerable expenses involved in analyzing a broader range of devices, which inevitably introduces certain limitations. Despite the breadth of our investigation, it is important to recognize its boundaries. The scope of our analysis is limited by the models and datasets available at the time of research, possibly omitting some of the latest advancements in Large Language Model (LLM) capabilities. Furthermore, the ever-evolving landscape of AI technology and regulatory standards highlights the need for continuous research to keep our findings up-to-date and relevant. Additionally, our focus on the accuracy of device classification within this study does not cover wider ethical, privacy, and security considerations, which are critical when applying LLMs in sensitive fields such as healthcare. Additionally, the "black box" nature of LLMs presents a significant limitation. Since the devices tested are already publicly registered, we cannot rule out the possibility that models retrieved correct classifications from their training data rather than applying genuine regulatory reasoning. This raises questions about how well these models would generalize to novel, unregistered devices.

These aspects represent vital areas for future, more comprehensive studies. By expanding our research focus to address these considerations in future studies, the academic and professional communities can progress toward fully harnessing AI's potential to a specific area.

6.6 Conclusion

This study marks an initial foray into understanding how large language models (LLMs) can be applied within the regulatory landscape, particularly focusing on the crucial task of medical device classification. This endeavor is vital for ensuring public health and safety. Our findings shed light on the considerable promise and existing limitations of LLMs in adhering to the rigorous standards of regulatory compliance in healthcare. The variation in model performance underlines the early phase of integrating LLMs into regulatory processes, pointing out the necessity for focused improvements in areas such as model training, embracing linguistic diversity, and enhancing contextual comprehension.

Our investigation offers an in-depth evaluation of how top-tier LLMs fare in the essential duty of categorizing medical devices across key global markets, revealing LLMs' significant potential to aid in healthcare regulatory compliance tasks. Yet, it also delineates essential areas requiring enhancement and further inquiry. By mapping out the strengths and weaknesses of current LLM technologies, this research not only enriches our grasp of AI's role in regulatory practices but also paves the way for future breakthroughs that could markedly boost the efficiency and efficacy of regulatory operations. As AI technologies continue to evolve, the insights from this study highlight the need for ongoing research and collaboration to effectively integrate LLMs into regulatory workflows.

Having observed the current limitations of large language models in regulatory downstream tasks, Chapter 7 shifts focus to the several algorithms including ML, DL, LLM. Using a larger dataset sourced from a real regulatory authority—the National Medical Products Administration (NMPA)—we systematically assess the classification performance of these models. In addition, we apply explainability techniques such as SHAP to better understand the factors influencing each model's performance, and to identify why certain methods succeed or fail in classifying regulatory texts accurately.

7

Text Analysis, Machine Learning and Requirement Engineering

This chapter is based on: Yu Han; Aaron Ceross; Jeroen Bergmann, Governing AI in Medical Device Regulation: A Framework for Risk-Stratified Accuracy, Explainability, and Lifecycle Feasibility, ACM Transactions on Computing for Healthcare, under review [209]

Author contributions (CRediT): Yu Han led the work (Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Visualization, and Writing – original draft); the co-authors contributed to Conceptualization, Writing – review & editing, and Supervision.

7.1 Introduction

Building on the insights gained from initial experiments with large language models in Chapter 6, this Chapter 7 expands the scope of investigation to include a broader set of algorithms—ranging from traditional machine learning to deep learning and foundation models—evaluated on a large-scale regulatory dataset.

In recent years, regulatory agencies have shown growing interest in leveraging artificial intelligence (AI) to enhance the efficiency and accuracy of their review processes. Notably, the U.S. Food and Drug Administration (FDA) completed its first AI-assisted scientific review pilot in early 2025 and announced plans for agency-wide deployment of generative AI tools across all centers by mid 2025 [210]. As part of this initiative, the FDA engaged in exploratory discussions with OpenAI on a project referred to as "CDERGPT," aiming to incorporate large language models (LLMs) into product evaluation

and regulatory workflows [211]. These developments underscore a broader global trend toward integrating AI into regulatory science, highlighting the urgent need for rigorous, context-sensitive evaluations of AI models—particularly in high-stakes domains such as medical device classification, where interpretability and legal compliance are paramount.

Regulatory affairs is a multidisciplinary field dedicated to ensuring that healthcare products—including medical devices, pharmaceuticals, and biologics—comply with the stringent requirements of government agencies and international health authorities. These agencies include the U.S. Food and Drug Administration (FDA), the European Union (EU), and the National Medical Products Administration (NMPA) in China, among others. Regulatory professionals are responsible for guiding products through a complex and evolving landscape of compliance, from early-stage development to clinical testing, market authorization, and post-market surveillance. Among the many responsibilities in this field, the classification of medical devices is foundational, as it determines the necessary level of regulatory scrutiny. Depending on the classification, the requirements for clinical evidence, labeling, safety documentation, and conformity assessments vary substantially [23, 212, 213].

Globally, the risk-based classification system remains the dominant framework for evaluating medical devices [198, 214]. In the U.S., the FDA mandates that Class I devices undergo general controls, while Class II devices typically require 510(k) clearance, and Class III devices—representing the highest risk—must pass rigorous premarket approval (PMA). Similarly, under the EU's Medical Device Regulation (MDR), and comparable systems in Australia and China, increasing classification levels correspond to increased regulatory requirements. Non-compliance with these systems can result in legal and financial consequences, including registration failure, recalls, or product bans.

A crucial first step in this regulatory workflow is the submission of detailed product descriptions. These descriptions are central to the classification decision, as regulators typically rely on textual claims made by manufacturers to assess intended use, operational features, and risk levels [215, 216]. Misleading or incomplete disclosures can not only compromise patient safety but also breach legal requirements. For instance, if a device incorporates AI functionality that is not clearly stated in the product description, regulators will not classify it accordingly, and subsequent marketing claims regarding AI use would be considered non-compliant [217]. Thus, product classification hinges on accurate and complete textual information, which becomes a critical input for any AI-driven regulatory analysis.

As discussed in Chapter 5, the increasing complexity and heterogeneity of medical device descriptions present substantial challenges for accurate risk classification. In this context, AI-driven approaches are becoming increasingly indispensable to support regulatory decision-making processes. Prior research has documented a steady rise in regulatory complexity, alongside an expansion in the volume of classification requirements and compliance assessments [15]. To address those, researchers have explored the application of machine learning (ML) and natural language processing (NLP) to automate text classification tasks. In clinical contexts, ML and deep learning models have proven effective for tasks such as diagnosis prediction, treatment recommendation, and report classification [218–220]. For regulatory affairs—where legal and scientific language intersects—the opportunity for automation is substantial but underexplored. Prior work has focused on specific use cases, such as AI-assisted dossier construction, data extraction from forms, and safety signal detection in post-market surveillance [221, 222].

Text processing methods for medical product classification is still underexplored. Text processing tasks can be broadly categorized into rule-based and machine learning-based approaches. Rule-based methods rely on manually predefined classification rules [223], which regulatory experts construct based on existing guidelines and domain knowledge. While such methods can achieve high accuracy in well-defined scenarios [220], they suffer from significant limitations, including high maintenance costs, lack of adaptability, and difficulty in handling complex, unstructured regulatory texts [224]. Each update to regulatory guidelines necessitates manual revisions to rule sets, making them impractical for large-scale classification tasks. Additionally, reliance on human consultants introduces cost inefficiencies and potential inconsistencies, as manual classification decisions are susceptible to variations in expert judgment [225]. These challenges highlight the need for scalable, automated approaches capable of adapting to evolving regulatory requirements while maintaining interpretability and reliability.

Machine learning-based methods could automatically learn classification patterns from historical regulatory data. Traditional algorithms such as Support Vector Machines (SVMs) and Naive Bayes classifiers have been widely used in text classification tasks. SVMs are particularly effective in high-dimensional spaces, leveraging hyperplane-based decision boundaries to separate different categories [226]. Naive Bayes classifiers, which assume feature independence, provide computational efficiency and robustness for large text datasets [227], making them well-suited for initial classification tasks in regulatory affairs. Additionally, semantic analysis techniques have been employed to extract meaningful patterns from medical texts, improving classification accuracy and

decision-making in regulatory applications [228, 229]. However, traditional machine learning models typically rely on manual feature engineering, limiting their adaptability to diverse regulatory texts and reducing their ability to generalize across different classification scenarios. Recent work by Ceross et al. has shown that machine learning models can classify medical devices based on structured and unstructured data extracted from regulatory databases, such as the Australian Register of Therapeutic Goods (ARTG) [186]. Deep learning and transformer-based models have emerged as promising tools for complex NLP tasks, including classification, named entity recognition, and entailment analysis. Large language models (LLMs) are capable of understanding context-rich input, capturing domain-specific language, and generating human-like interpretations [230]. These models hold potential for more nuanced classification decisions, particularly when regulatory guidance is ambiguous or device descriptions are detailed and complex. Nevertheless, their adoption in regulatory environments remains limited due to concerns about interpretability, computational overhead, and legal accountability. Artificial intelligence offers the potential to automate or assist in medical device classification, but regulatory applications demand more than just high accuracy. In this context, interpretability of AI models is crucial so that regulators can trust and understand the basis of a model's recommendation [231], [232]. For example, an algorithm that classifies a device as high-risk should be able to explain whether it was due to invasive use, critical function, or other factors recognizable to regulatory scientists. Additionally, computational cost becomes important for practical adoption: regulatory agencies may need to run these models at scale or on secure systems with limited resources, making efficiency and speed necessary considerations. Balancing accuracy with interpretability and efficiency is therefore a key challenge for deploying AI in this domain [233]. The balance between these factors is not fixed, and regulatory decisions should be based on the particular context.

In this study, we pick classification as an RA downstream task, seek to systematically evaluate multiple AI approaches—ranging from rule-based to traditional ML, deep learning, and LLMs—for the task of medical device classification based on regulatory product descriptions. We examine three core challenges:

1. **Accuracy:** How well can each model predict the correct regulatory class from free-text product descriptions?

2. **Interpretability:** Can we explain why a model made a decision, and do these explanations align with regulatory reasoning?

3. **Cost:** What is the computational cost of using high-performing models, particularly LLMs, and how does this impact scalability and sustainability?

7.2 Related Work

Text classification has long been a central task in natural language processing (NLP), and a wide array of computational methods have been proposed to solve it, ranging from rule-based systems to machine learning (ML), deep learning (DL), and large language models (LLMs). Each category has shown strengths and limitations in regulatory applications such as medical device classification, particularly regarding interpretability, scalability, and computational efficiency.

7.2.1 Rule-Based Systems

Rule-based systems rely on manually crafted patterns, ontologies, or decision trees to categorize text data. These approaches have traditionally dominated regulatory domains due to their interpretability and adherence to predefined logic [223, 225]. For instance, keyword-based systems can classify medical device descriptions based on phrases like "non-invasive" or "implantable," which directly map to regulatory risk categories. However, they lack scalability and adaptability to novel or nuanced language use [220, 224], requiring constant manual updates when regulatory guidelines evolve.

7.2.2 Traditional Machine Learning

Classic ML models such as Naive Bayes, Support Vector Machines (SVMs), Logistic Regression, and ensemble methods like Random Forest and XGBoost have been extensively used in text classification tasks [227, 234, 235]. These models rely on preprocessed and vectorized representations of text, such as TF-IDF (Term Frequency–Inverse Document Frequency) or word embeddings, and provide strong performance when combined with effective feature engineering. Naive Bayes and Logistic Regression offer interpretable decision-making via conditional probabilities and coefficients, respectively, while SVMs and tree-based models require post-hoc explanation methods [236].

XGBoost in particular has emerged as a highly effective model due to its ability to handle non-linear feature interactions, and is now a standard baseline in many structured and semi-structured classification tasks [235]. Recent studies in regulatory informatics have demonstrated the feasibility of using these models for classification based on structured product descriptions [186].

7.2.3 Deep Learning Models

Neural network-based models, especially Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have been adopted for regulatory text classification for their ability to automatically extract hierarchical and sequential features [237, 238]. CNNs are efficient at identifying local patterns (e.g., clinical usage phrases), while RNNs and their variants (e.g., DPCNN, RCNN) are better suited for capturing long-range dependencies within sentences [239, 240]. Although these models outperform traditional ML in accuracy, they are often considered black boxes and lack native interpretability, necessitating tools like Integrated Gradients or LIME (Local Interpretable Model-agnostic Explanations) [241, 242].

7.2.4 Transformer-Based Models and LLMs

The introduction of transformer architectures [243] revolutionized NLP by enabling parallel sequence modeling through self-attention mechanisms. Pretrained models such as BERT (Bidirectional Encoder Representations from Transformers) [244] and its Chinese variants [245] have been applied to regulatory tasks by generating rich contextual embeddings for downstream classifiers. More recently, LLMs like DeepSeek and LLaMA [246, 247] have shown potential for zero-shot classification through prompt-based methods. In *zero-shot classification*, a model assigns an input to one of a set of candidate categories described in natural language at inference time, without having been fine-tuned on any labelled examples for that specific task.

While LLMs offer unprecedented generalization capabilities, their applicability in regulatory domains remains limited by issues of computational cost, consistency, and lack of transparency [230]. *Post-hoc interpretability*—methods applied *after* a model has been trained to explain its predictions, rather than relying on an inherently transparent model—via techniques such as LIME or token-level attribution helps address this, but introduces challenges related to explanation stability and semantic alignment with human reasoning [248].

7.2.5 Model Interpretability in Regulatory Contexts and Computational Cost

Interpretability is a critical concern in regulated environments, where model decisions must be traceable and understandable [217, 249]. Tools such as SHAP (SHapley Additive exPlanations) [236], LIME [242], Anchors [250], and Integrated Gradients [241] have

become standard for explaining predictions of black-box models. SHAP in particular enables consistent feature attribution across model types and has been effectively applied in medical AI contexts [251]. TreeSHAP, a variant tailored for tree ensembles like XGBoost, is especially efficient and faithful [252]. However, explanation fidelity tends to degrade with model complexity, particularly for deep or transformer-based models.

Beyond accuracy and interpretability, computational cost is a practical bottleneck in real-world deployment [253]. Classical models like Logistic Regression and Naive Bayes offer fast inference and low memory usage, ideal for integration into constrained systems. In contrast, LLMs require significant resources for both inference and training, which can limit their accessibility in clinical and regulatory settings [254]. Efficient variants of BERT and inference-optimized architectures are being explored to mitigate these barriers [255].

7.2.6 Trade-off

Although AI has the potential to enhance regulatory decision-making, model selection requires careful consideration of trade-offs across accuracy, interpretability, and computational feasibility [212, 216], particularly in high-stakes domains such as healthcare and regulation. Recent work has highlighted that these performance dimensions are often in tension, requiring practitioners to make deliberate design choices based on contextual priorities [256]. For instance, while deep learning models tend to achieve superior predictive accuracy, they often do so at the expense of interpretability and computational cost, making them less ideal for settings requiring transparency and auditability. Simpler models like logistic regression or decision trees may provide clearer reasoning pathways but struggle with complex or nuanced inputs.

Zhang et al. proposed a Composite Interpretability (CI) score that quantifies this trade-off by integrating model transparency and explanation fidelity, observing that the relationship between interpretability and performance is not always monotonic and may depend on data domain and model architecture [256]. Similarly, studies in the fairness and accountability literature have emphasized that these trade-offs are not merely technical, but are shaped by broader socio-technical constraints, including the epistemic trust of regulators and the institutional requirements for interpretability [257]. Chen et al. further explore this space by showing how user-facing explanation systems must balance cognitive effort with informativeness, highlighting that interpretability is not a purely objective metric but one that varies with audience expertise and context [258].

Our study situates the model selection problem for medical device classification squarely within this trade-off space. By benchmarking traditional ML models, deep learning architectures, and LLMs across accuracy, interpretability, and computational cost, we provide empirically grounded, engineering-oriented guidance for selecting models that align with regulatory objectives—whether the priority is classification accuracy, decision transparency, or scalable deployment.

7.3 Theoretical Framework : Text classification methods

7.3.1 Rule-Based Classification

Rule-based classification relies on a predefined set of manually crafted linguistic rules to categorize text data. In regulatory affairs, professionals often employ this approach to determine the classification of medical devices by following explicit regulatory guidelines and domain knowledge. These rules typically leverage keyword searches, regular expressions, and ontological frameworks to identify relevant features in the text and assign classifications accordingly. In the context of medical device classification, rule-based systems can efficiently categorize devices by detecting specific terminology or "markers" outlined in regulatory documents. These systems operate based on a structured set of "if-then" rules, where classification decisions are explicitly defined by domain experts rather than learned from data. While rule-based approaches can provide high interpretability and consistency, they are inherently limited in scalability and adaptability, as they require continuous manual updates to accommodate changes in regulatory guidelines and evolving medical technologies.

The decisions made by rule-based classifiers are easy to understand and explain, making them particularly useful in domains where interpretability is crucial, such as in finance and medicine. These systems are straightforward to implement and modify, as they do not require any training process. Once the rules are defined, the classification decisions remain consistent unless the rules are changed. However, as the complexity of the data increases, managing and updating the rules can become cumbersome. Rule-based systems are not adaptive, which means they do not learn from new data. They may not perform well if the underlying data distribution changes over time.

7.3.2 Traditional Machine Learning Models

Support Vector Machines (SVM)

Support Vector Machine (SVM) is a widely utilized classification algorithm known for its effectiveness in supervised learning tasks [234]. SVM operates by constructing an optimal decision boundary, called a hyperplane, that separates the data points belonging to different classes. Among the many boundaries that could separate the classes, SVM selects the one that maximises the margin — that is, the one that leaves the largest possible gap between the boundary and the nearest training points of each class. By sitting as far as possible from the closest points of either class, this boundary gives the classifier greater robustness and better generalisation to unseen data. This margin is defined as the distance between the nearest data points from each class, known as support vectors, which play a critical role in defining the optimal classification boundary. SVM is particularly advantageous in multi-class classification problems due to its ability to handle high-dimensional feature spaces and non-linearly separable data through the use of kernel functions. The dimensionality of the hyperplane depends on the feature space: in a two-dimensional space, the hyperplane is a line, whereas in a three-dimensional space, it forms a plane. More generally, for an N-dimensional feature space, the separating hyperplane is an $(N-1)$-dimensional subspace [259].

Optimization Problem

SVM aims to minimize the following objective function:

$$\min_{w,b} \frac{1}{2}\|w\|^2 \tag{7.1}$$

Subject to the constraint:

$$y_i(w^T x_i + b) \geq 1 \tag{7.2}$$

Here w is the weight vector normal to the separating hyperplane, b is the bias (offset) term, x_i is the feature vector of the i-th training example, and $y_i \in \{-1, +1\}$ is its class label; the constraint requires each example to lie on the correct side of the margin. This optimization problem (Equations 7.1 and 7.2) can be solved using techniques like the Lagrange multiplier method or quadratic programming.

Using a Radial Basis Function (RBF) Kernel

Since text data is not always linearly separable, a kernel function could be introduced to handle nonlinear relationships. One common choice is the Radial Basis Function

(RBF) kernel, which is defined as:

$$K(x, x') = \exp(-\gamma \|x - x'\|^2) \tag{7.3}$$

Where:

- x and x' are two input feature vectors (e.g., BERT embeddings of two device descriptions),
- $\|x - x'\|^2$ is the squared Euclidean distance between these two vectors,
- γ is a parameter that controls the kernel's spread (higher values make the model more sensitive to individual data points).

The kernel maps the input vectors into a higher-dimensional space where a linear separation might be possible.

In this study, we propose a novel approach for medical device classification by combining SVM with BERT embeddings to categorize devices based on their textual descriptions. While SVM and BERT are well-established techniques individually, their combined application to regulatory medical device classification represents an original contribution of this thesis. Device descriptions typically contain essential information regarding their intended use, and our method leverages this textual data for automated classification. Given a dataset of medical device descriptions, the objective is to predict the appropriate classification (e.g., Class I, Class II, or Class III) based on textual features. To implement this approach, medical device descriptions must first be converted into vectorized representations, enabling SVM to process and learn from the textual data. Modern natural language processing techniques, such as BERT embeddings, provide high-dimensional vector representations that capture semantic meaning and contextual relationships between words [260]. For instance, consider a vocabulary consisting of key regulatory terms such as monitoring, patient, surgery, heart, implant, and glucose. Each medical device description can then be represented as a numerical vector, where each dimension corresponds to the presence or relevance of a specific term. By leveraging these vectorized features, SVM constructs a hyperplane that separates the different device categories based on textual patterns. The effectiveness of this method depends on the quality of text representation, the choice of kernel functions, and the ability of SVM to handle high-dimensional sparse data. Although SVM offers strong generalization capabilities, its performance is highly sensitive to feature engineering

techniques, making the selection of appropriate text embeddings a crucial step in achieving optimal classification accuracy.

SVM can classify medical devices by transforming their descriptions into vector representations. By optimizing the decision boundary, SVM maximizes the margin between different categories, effectively distinguishing between them. In cases where the data is not linearly separable, SVM can leverage kernel functions like the RBF kernel to map the data into higher-dimensional spaces, where a linear separation is possible. To illustrate how SVM classifies a new device, consider a trained model that has learned decision boundaries from the training data. When a new device description such as "device for heart surgery" is converted to its vector representation, the SVM applies its learned decision function to determine the class. If the function output is positive, the device is classified as one class (e.g., Class II); if negative, it is classified as the other class (e.g., Class III). The specific decision boundaries are learned automatically from the training data rather than manually specified.

7.3.3 Deep Learning Models

Convolutional Neural Network

Convolutional Neural Networks (CNN) are widely used in image processing tasks but have also been extensively applied to Natural Language Processing (NLP) tasks, such as text classification. When dealing with text, CNN extracts local features (e.g., phrases or keywords) through convolutional layers, reduces the feature dimensionality via pooling layers, and finally classifies through fully connected layers. In this case, we will use CNN to predict the classification of medical device descriptions (e.g., Class II or Class III). CNN extracts local features from the input using convolution operations, making it particularly useful for capturing phrases and keywords in the text. The main components of a CNN are:

- **Convolutional Layer:** Uses filters (kernels) to slide across the input text and generate feature maps.
- **Pooling Layer:** Reduces the size of the feature maps by selecting the most important features (e.g., max pooling or average pooling).
- **Fully Connected Layer:** The output from the convolutional and pooling layers is *flattened*—that is, the multi-dimensional feature maps are reshaped into a single one-dimensional vector—and passed to fully connected layers for classification.

For a given input vector x and a convolution filter w, the convolution operation is defined as:

$$y_i = \sum_{k=1}^{m} w_k x_{i+k-1} + b \tag{7.4}$$

Where:

- x is the input vector,
- w is the filter (or kernel),
- b is the bias term, and
- m is the size of the filter.

This formula describes how the filter slides over the input, generating a feature map.

Convolutional Layer: The convolutional layer applies filters that slide over the input matrix to extract local features. Suppose the filter size is 2×2, and we apply the convolution to the input matrix.

For the device description "Used for heart implant surgery," represented over the six-term example vocabulary {monitoring, patient, surgery, heart, implant, glucose} by the vector $x = [0, 0, 1, 1, 1, 0]$ (in which the entries for *surgery*, *heart* and *implant* are set to one), we use a filter w of size 2×2:

$$w = \begin{bmatrix} 0.2 & 0.3 \\ 0.4 & 0.5 \end{bmatrix} \tag{7.5}$$

Assume part of the input matrix is:

$$\text{Input} = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \tag{7.6}$$

The convolution operation using Equation 7.4 would be:

$$y = 0.2 \times 0 + 0.3 \times 1 + 0.4 \times 1 + 0.5 \times 1 = 1.2 \tag{7.7}$$

The filter slides across the input matrix, generating multiple feature values.

Pooling Layer: After the convolution, the feature maps are passed through a pooling layer to reduce dimensionality. The most common method is max pooling, which selects the maximum value in each region:

$$\text{Max Pooling} = \max\{y_1, y_2, \dots\} \tag{7.8}$$

This layer reduces the size of the feature map while retaining the most important features.

Fully Connected Layer: The output from the pooling layer is flattened and passed to a fully connected layer for final classification. The output of the fully connected layer is passed through a Softmax function to compute the probability for each class:

$$P(y = k|x) = \frac{\exp(w_k^T x)}{\sum_{j=1}^{C} \exp(w_j^T x)} \tag{7.9}$$

Where:

- w_k is the weight vector for class k,
- C is the number of classes.

After applying convolution and pooling, suppose the pooled feature vector for "Used for heart surgery" is $z = [0.7, 0.5, 0.2]$; the length of z equals the number of convolutional filters (three in this example), not the number of words in the description. We use a fully connected layer to classify the device. Assume the weight vectors are:

$$w_{\text{Class II}} = [0.4, 0.3, 0.1], \quad w_{\text{Class III}} = [0.6, 0.2, 0.5] \tag{7.10}$$

To calculate the probability for each class using the Softmax function (Equation 7.9):

Score for Class II:

$$z_{\text{Class II}} = 0.7 \times 0.4 + 0.5 \times 0.3 + 0.2 \times 0.1 = 0.49 \tag{7.11}$$

Score for Class III:

$$z_{\text{Class III}} = 0.7 \times 0.6 + 0.5 \times 0.2 + 0.2 \times 0.5 = 0.67 \tag{7.12}$$

Softmax Calculation:

$$P(\text{Class II}|x) = \frac{\exp(0.49)}{\exp(0.49) + \exp(0.67)} \approx 0.46 \tag{7.13}$$

$$P(\text{Class III}|x) = \frac{\exp(0.67)}{\exp(0.49) + \exp(0.67)} \approx 0.54 \tag{7.14}$$

Therefore, we predict that this device belongs to Class III since it has the higher probability.

Large Language Models (LLMs)

When applying a Large Language Model (LLM) to a task where X represents medical device descriptions and Y represents the classification, the core goal of the model is to learn the mapping between input X (medical device descriptions) and output Y (classification, such as Class I, Class II, Class III).

The formula and derivation of LLMs are based on a probabilistic framework for conditional generation. Here, we can consider an LLM as a generative model that learns to maximize the probability $P(Y|X)$ for classification.

Objective: Maximize Conditional Probability

For each input medical device description X, the goal of the LLM is to learn the conditional probability distribution $P(Y|X)$, where Y represents possible classification categories (such as Class I, Class II, or Class III). The objective is to maximize the following formula:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \tag{7.15}$$

However, for classification tasks, the focus is on directly modeling $P(Y|X)$, which simplifies to:

$$P(Y|X) = \text{softmax}(W^T h(X)) \tag{7.16}$$

Where:

- $h(X)$ is the text embedding generated by the LLM (such as GPT or BERT), representing the contextual vector of the input description.
- W is the weight matrix of the classification layer, which acts on the model's output vector to generate the probability distribution over the classes.
- softmax is the normalization function used to convert the model output into a probability distribution.

Text Encoding

The input medical device description X is transformed into an implicit contextual representation $h(X)$ via the encoding layer of the model. LLMs use the self-attention mechanism from the Transformer architecture to generate this contextual representation. The formula for this process is:

$$h(X) = \text{Transformer}(X) \tag{7.17}$$

In the Transformer, self-attention is used to compute the relationship between each word in the input sequence and generate a contextual embedding for each word.

Self-Attention Mechanism

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{7.18}$$

Where:

- Q, K, and V are the query, key, and value vectors obtained from the input word embeddings via linear transformations.
- d_k is the dimension of the key vector used to scale the attention scores.

After multiple layers of the Transformer, each word in the input description X interacts with its context, generating a global semantic representation $h(X)$.

Classification Head: From Text Embedding to Class Probabilities

The final step involves passing the contextual representation $h(X)$ through a classification head (usually a fully connected layer followed by a softmax function) to convert the representation into a probability distribution for each class. The specific formula is:

$$P(Y = k|X) = \frac{\exp(W_k^T h(X))}{\sum_{j=1}^{C} \exp(W_j^T h(X))} \tag{7.19}$$

Where:

- W_k is the weight vector for class k,
- C is the total number of classes.

Derivation Process

Input Preprocessing: First, preprocess the input medical device description X through tokenization and adding special tokens. BERT-based models use [CLS] (classification token) at the beginning of the input to represent the entire sequence for classification tasks, and [SEP] (separator token) to mark the end of a sentence or separate multiple sentences. The preprocessed input is then fed into the Transformer model. **Contextual Representation Encoding:** The Transformer generates the contextual embedding

$h(X)$ for X, using self-attention to calculate the relationship between each word and its context. In BERT or GPT models, the embedding of the [CLS] token or the last token is typically used for downstream classification tasks.

Class Probability Prediction: Pass $h(X)$ through a fully connected layer, multiply it by the class weight matrix W, and apply the softmax function to obtain the probability of each class (as in Equation 7.19).

The model will ultimately choose the class with the highest probability as the predicted output. **Input Data:**

Device Description (Text)	Category
Used for patient heart monitoring	Class II
Device for implant surgery	Class III
Device for blood glucose monitoring	Class II
Device for heart stent surgery	Class III

Table 7.1: Medical device descriptions and their categories

Suppose we want to classify a new input description, "Used for heart surgery."

Input Processing: Tokenize and preprocess the input description. If using a GPT model, the input format might look like: `"Used for heart surgery"`.

Contextual Representation: The model generates a contextual representation $h(X)$ for the description. Suppose the final contextual vector is $h(X) = [0.7, 0.3, 0.5]$.

Classification Prediction: Pass $h(X)$ into the classification layer to compute the probability for each class:

Score for Class II:

$$z_{\text{Class II}} = W_{\text{Class II}}^T h(X) = 0.4 \times 0.7 + 0.2 \times 0.3 + 0.1 \times 0.5 = 0.43 \tag{7.20}$$

Score for Class III:

$$z_{\text{Class III}} = W_{\text{Class III}}^T h(X) = 0.5 \times 0.7 + 0.3 \times 0.3 + 0.4 \times 0.5 = 0.65 \tag{7.21}$$

Probability Calculation:

$$P(\text{Class II}|X) = \frac{\exp(0.43)}{\exp(0.43) + \exp(0.65)} \approx 0.44 \tag{7.22}$$

$$P(\text{Class III}|X) = \frac{\exp(0.65)}{\exp(0.43) + \exp(0.65)} \approx 0.56 \tag{7.23}$$

Therefore, the model predicts that the description belongs to Class III, as it has the higher probability. In the task of using an LLM (such as GPT or BERT) for medical device description classification, the core is learning the conditional probability distribution

between input text X and output class Y. The Transformer model generates a contextual representation $h(X)$ of the text through the self-attention mechanism, and this representation is mapped to the classification probability space via a classification head.

7.4 Experiments

In this study, we implement a comprehensive data preprocessing pipeline to prepare medical device descriptions for a classification task. The raw dataset comes from the National Medical Products Administration (NMPA) Unique Device Identification (UDI) database, containing records of medical devices with descriptions in Chinese. We outline the key preprocessing stages below – from data collection and cleaning, through text normalization, embedding generation, and finally splitting the processed data for model training. Each stage is designed to ensure the data is clean, consistent, and properly formatted for input into either traditional machine learning models or deep learning models, as shown in Figure 7.1.

7.4.1 Dataset and Preprocessing

We curated a real-world dataset from the China National Medical Products Administration (NMPA) to support the classification of medical devices by regulatory risk. The primary data source was the UDID FULL RELEASE 20240701 registry [261], which contains over 4 million entries. After deduplication and the removal of incomplete or placeholder records, we retained a high-quality subset of 42,000 labeled samples with complete fields for device description (cpms), product name, and device category.

Due to the significantly higher computational cost of large language model (LLM) inference, we further sampled 1,000 entries—again using stratified sampling, that is, drawing the subset so that the Class I/II/III proportions match those of the full dataset—for our zero-shot classification experiments. The goal of this reduced subset was not to match the predictive performance of fully trained models, but rather to explore the reasoning capabilities and interpretability of LLMs in a constrained yet representative setting.

We acknowledge that the use of different dataset sizes introduces a significant limitation in directly comparing models across experiments. The primary reason for this design choice was computational: LLM inference is orders of magnitude more expensive than traditional ML models—processing 42,000 samples with LLMs would require prohibitive computational resources and time (estimated at over 40 hours for

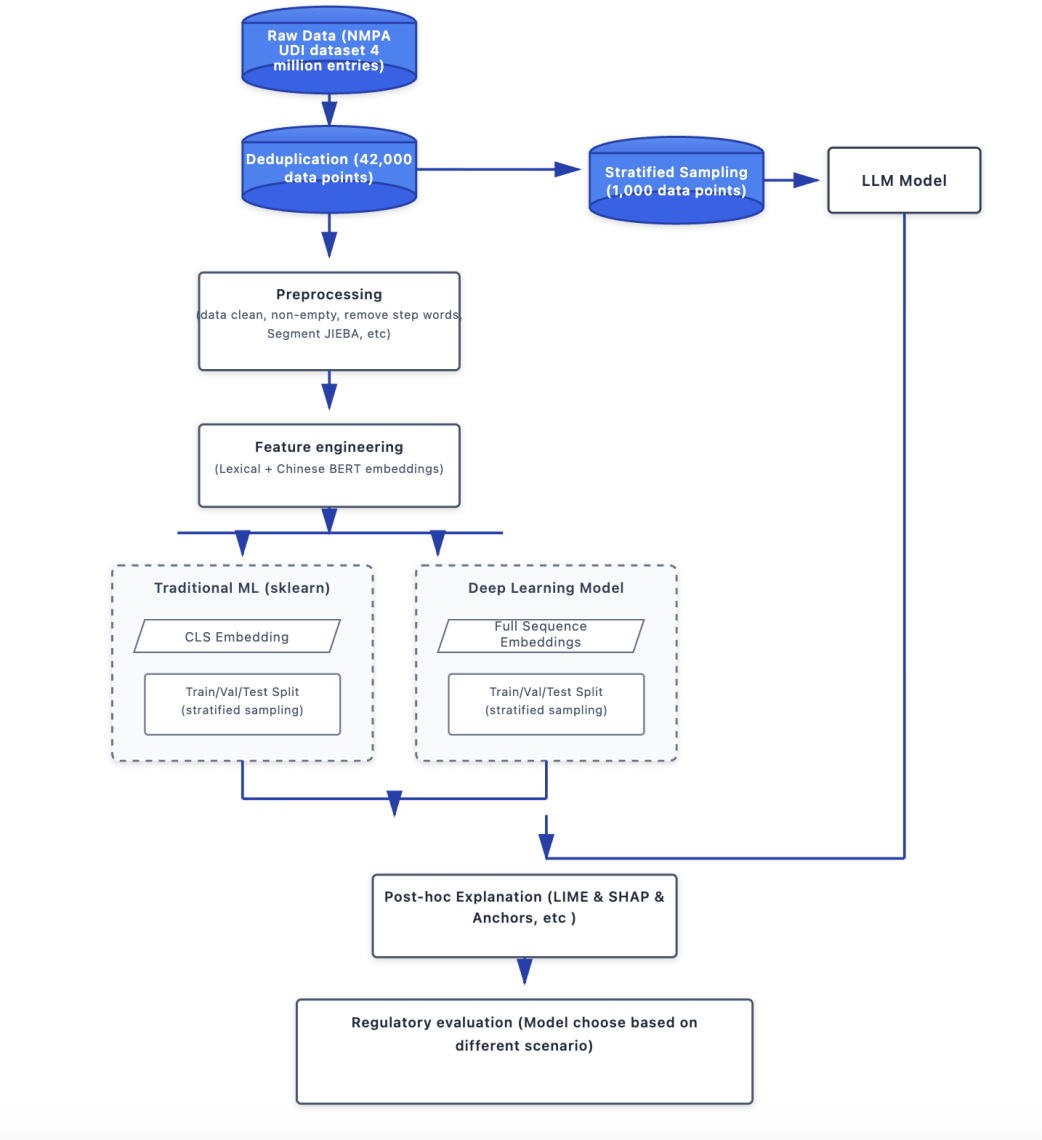


Figure 7.1: Text process, Model Training and Evaluation Pipeline

a single model). To address this limitation, we: (1) clearly distinguish between results from the 42,000-sample experiments (Tables 7.7–7.8) and those from the 1,000-sample LLM evaluation (Table 7.9); (2) use stratified sampling to ensure the 1,000-sample subset maintains the same class distribution as the full dataset; and (3) caution readers against making direct performance comparisons across these experiments. Future work with greater computational resources could enable a more direct comparison on identical dataset sizes.

For classification targets, we used the officially assigned NMPA risk labels—Class I, II, or III—as ground truth. In cases where labels were missing, we inferred the class from the unique device identifier field. In the NMPA database, this field (internally

named zczbhhzbapzbh, an abbreviation for "registration certificate number") contains encoded information about the device's risk classification. By parsing specific character positions within this identifier, we could reliably extract the regulatory class designation. To preserve domain-specific semantic nuances, device descriptions (cpms) were processed in Chinese using pretrained language models. We applied *jieba* for word segmentation, and input sequences were encoded using the *hfl/chinese-bert-wwm-ext* model from HuggingFace [245].

Standard preprocessing steps included punctuation removal, stopword filtering, and lexical normalization. Class labels were numerically encoded using *LabelEncoder* to support multi-class classification—here the three mutually exclusive risk classes (Class I, Class II and Class III), with each device assigned to exactly one class. The data was split into training (80%) and testing (20%) subsets. To ensure robust evaluation, we applied 10-fold stratified cross-validation to the training set and repeated all experiments across multiple random seeds to assess stability.

Table 7.2: Descriptive Statistics of Word Count by Risk Class. N indicates the number of devices in each class; Mean, Std Dev, Min, Max, and percentiles refer to the word count of device descriptions.

Risk Class	N (devices)	Mean (words)	Std Dev	Min	25%	Median	75%	Max
Class I	998	53.45	66.53	1	29	34	51	949
Class II	37,810	61.69	77.78	0	22	37	68	1321
Class III	27,664	57.23	70.43	1	16	32	75	1577

Each device has a textual description, which forms the semantic corpus used as model input; the distribution of word counts across the three risk classes is summarised in Table 7.2 and Figure 7.2. The distributions are strongly right-skewed—most descriptions are short while a small number are very long—so, under the standard $1.5\times$IQR whisker convention used in the box plot, the many points lying beyond the whiskers in Figure 7.2 are genuine long-tailed descriptions rather than plotting errors. To refine the data for model input, we identified and removed devices with unusually short or unusually long descriptions as outliers, using an adjusted interquartile range (IQR) method. This adjusted threshold helps ensure that only extreme outliers are excluded, thereby enhancing the accuracy of the model by focusing on a more representative dataset. For each risk class, we calculate the first quartile (Q1), the third quartile (Q3).

The outliers were then defined using a broader range to capture extreme values by setting the bounds as follows:

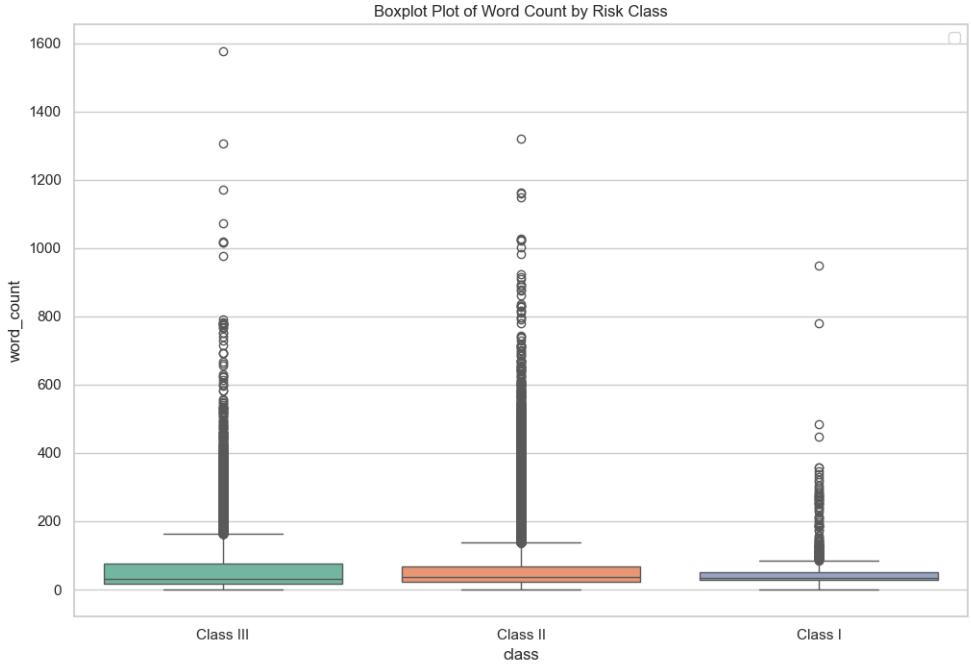


Figure 7.2: NMPA device Corpora Characteristics in Classification I, II and III Devices

$$\text{Lower Bound} = Q1 - 3 \times \text{IQR} \tag{7.24}$$

$$\text{Upper Bound} = Q3 + 3 \times \text{IQR} \tag{7.25}$$

Table 7.3 shows the quantiles (25th, 50th, and 75th percentiles) for each risk class. This provides insights into the distribution of word counts across different classifications. Following this, Table 7.4 summarizes the number of outliers identified for each risk class.

Table 7.3: Word Count Quartiles by Risk Class. Values represent the number of words in device descriptions at each percentile threshold.

Risk Class	25th Percentile (words)	Median (words)	75th Percentile (words)
Class I	29	34	51
Class II	22	37	68
Class III	16	32	75

Table 7.4: Outlier Summary by Risk Class

Risk Class	Number of Outliers
Class I	80
Class II	1,808
Class III	615

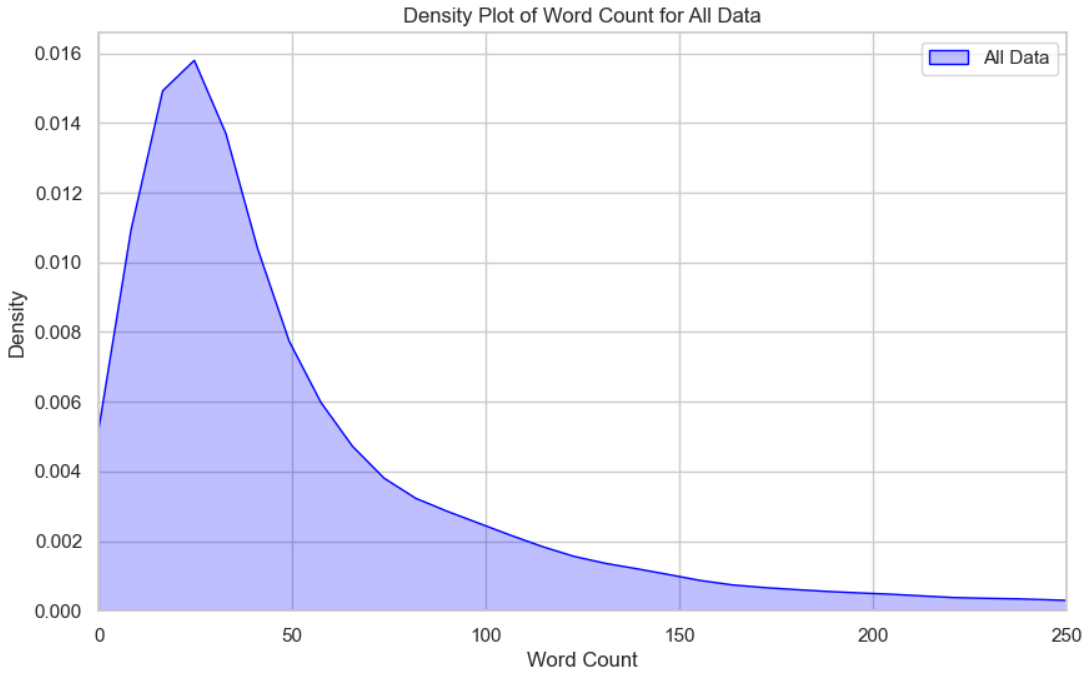


Figure 7.3: Kernel density estimate of the character-count distribution across the whole corpus. The horizontal axis is the number of characters in a device description; the vertical axis is the estimated probability *density*, normalised so that the area under the curve integrates to one (it is a relative frequency per unit character count, not a raw device count).

In the analysis of the character count distribution, the density plot is shown in Figure 7.3 and Figure 7.4. We can see the tokens amount across class I, II and III are very similar. We also checked the vocabulary overlap between those three sections and Table 7.5 presents the vocabulary overlap between Class I, II, and III, which is also visualised in Figure 7.5. The overlap is shown both in terms of the number of shared words and the percentage relative to the vocabulary size of the first class in each pair:

Table 7.5: Vocabulary Overlap Between Risk Classes

Risk Classes	Overlap Count	Overlap Percentage
Class I & II	2,470	83.00%
Class II & III	12,191	39.28%
Class I & III	2,294	7.75%

7.4.2 Text Representation

Since traditional machine learning models such as SVM cannot process raw text directly, we first convert text into numerical representations. We encode the tokenized text descriptions using the BERT model, which transforms each device description into a 768-dimensional vector that captures its semantic meaning. Specifically, we extract

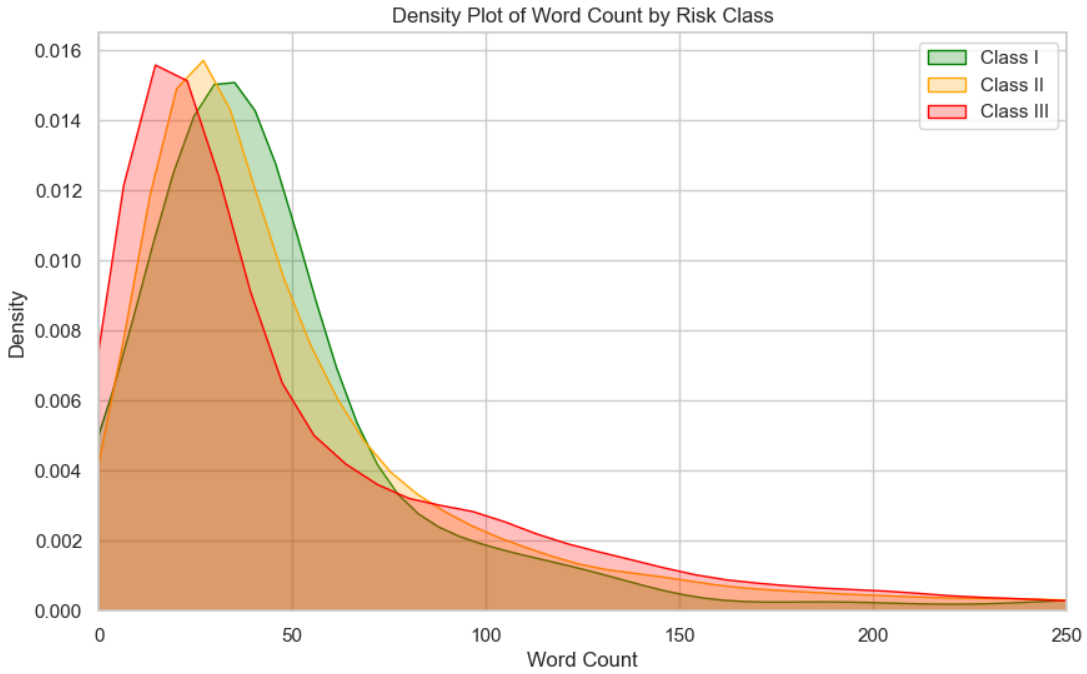


Figure 7.4: Density of Corpora by classification

the [CLS] token embedding as the sentence-level representation, resulting in a feature matrix X_bert. This transformation enables text-based classification using algorithms originally designed for numerical data: the SVM, for example, operates on these 768-dimensional vectors rather than on the raw text, finding optimal decision boundaries in this high-dimensional embedding space. This approach combines the contextual language understanding of BERT with the classification power of traditional machine learning algorithms.

Deep neural models such as BERT-LSTM and TextCNN utilize the full sequence output from BERT. Instead of summarizing each sentence into a single vector, they take the sequence of contextual token embeddings as input, allowing the model to learn from intra-sentence dependencies and local n-gram structures. Although the preprocessing and embedding stage is consistent across all models, this difference in representation usage allows deep models to capture more fine-grained semantic patterns. The corresponding class labels are encoded numerically with LabelEncoder.

7.4.3 Train–Validation–Test Split

The final step in data preprocessing is to split the prepared dataset into separate sets for model training, validation, and testing. We perform a stratified split to ensure that each subset has the same class distribution as the full dataset. In other words, the proportion

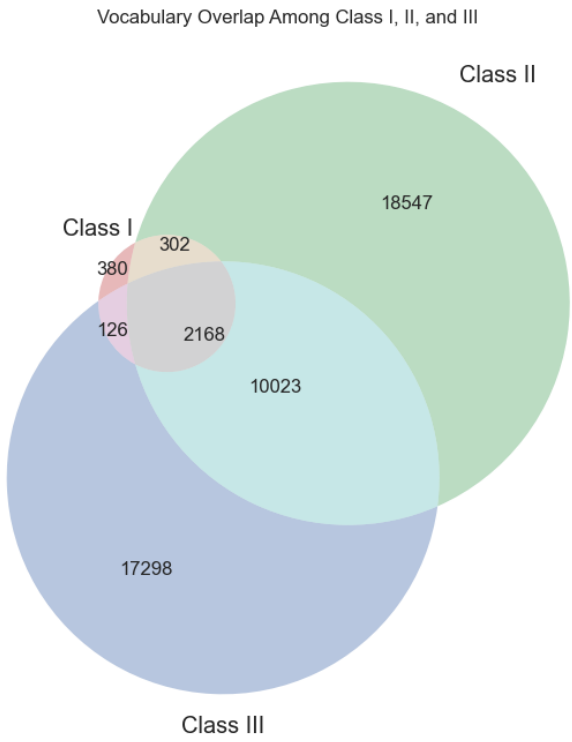


Figure 7.5: Vocabulary overlap between classifications. Values indicate the number of shared unique tokens between each pair of device risk classes.

of Class I, Class II, and Class III devices is preserved in the training set, the validation set, and the test set (each subset mirrors the overall distribution). This stratification is important for imbalanced or multi-class data to prevent, for example, the test set from having only high-risk devices or an uneven mix. Preserving class proportions across splits leads to a more reliable evaluation – the model's performance on the test set will reflect performance across all classes, since no class is under- or over-represented by chance [262]. Following standard machine learning practice, we adopt an 80/10/10 split: 80% of the records for training, 10% for validation, and 10% for testing [263]. The training set is used to train the models (either to fit the scikit-learn classifier or to train/fine-tune the deep neural network). The validation set is held out during training and used to tune hyperparameters and make model selection decisions – for instance, deciding when to stop training to avoid overfitting or which model architecture works best. Finally, the test set is kept completely independent of any training or tuning, and is used at the end to obtain an unbiased evaluation of the model's classification accuracy on new, unseen data. By using stratified sampling, we ensure that if (for example) Class III devices make up 20% of the whole dataset, they will be about 20% of each of the train/val/test sets as well, thus the model is trained and evaluated on representative data from all classes [262].

7.4.4 Classifiers and Training Setup

To evaluate the effectiveness of different modeling strategies for regulatory medical device classification, we implement and compare three families of classifiers: traditional machine learning models, deep neural network architectures, and large language models (LLMs). All models are trained and assessed using 10-fold stratified cross-validation, and performance is evaluated on a held-out test set using standard metrics including accuracy, precision, recall, and F1-score. We also log training and inference times to assess computational efficiency.

For input representation, we use the Chinese pre-trained BERT model *hfl/chinese-bert-wwm-ext* from HuggingFace [245]. Device descriptions are tokenized and encoded using this model, with the final [CLS] token embedding extracted to form a 768-dimensional fixed-length input vector.

Traditional Machine Learning Models. We evaluate five widely used classifiers: Logistic Regression (LR), Gaussian Naïve Bayes (GNB), Support Vector Machine (SVM), Random Forest (RF), and XGBoost (XGB). LR is configured with L2 regularization and class-balanced weights. SVM uses an RBF kernel with hyperparameters tuned via GridSearchCV over $C \in \{0.1, 1, 10\}$. RF includes 100 decision trees trained with class weights, while XGB is configured for multi-class classification using *mlogloss*, with tree depth $\in \{4, 6, 8\}$ and learning rate $\in \{0.05, 0.1\}$. All models use BERT [CLS] embeddings as input and are implemented in scikit-learn [264].

Deep Learning Models. We implement Convolutional Neural Networks (CNN), Deep Pyramid CNN (DPCNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) models using TensorFlow/Keras. All models take input shaped `(n, 768, 1)` from BERT embeddings. The CNN model consists of a 1D convolutional layer with 128 filters and kernel size 3 (*padding='same'*), followed by *GlobalMaxPooling1D*, a dropout layer (rate = 0.5), a dense layer with 64 ReLU units, and a softmax output layer. The LSTM model includes a single unidirectional LSTM with 128 units, followed by dropout (0.5), a dense layer (64 ReLU), and a softmax layer. Both models are compiled with the Adam optimizer (learning rate = 0.001), *sparse_categorical_crossentropy* loss, batch size of 32, 10 epochs, and a 20% validation split.

Large Language Models (LLMs). To explore zero-shot classification, we employ *DeepSeek-R1-Distill-Qwen-7B* via the Ollama local API. We use a 10-sample subset of

our dataset, each labeled with a regulatory class inferred from the NMPA identifier field (*zczbhhzbapzbh column*).

Each prompt simulates expert reasoning using NMPA risk classification:

> *You are a senior medical device regulatory affairs specialist familiar with the NMPA (National Medical Products Administration) classification system in China. Please classify the following medical device based on its description into one of the three NMPA regulatory risk classes:*
> **Class I** (Low Risk)
> **Class II** (Medium Risk)
> **Class III** (High Risk)
> **Device Description:** *{text}*
> **Your answer must be ONLY one of the following:** Class I, Class II, or Class III.

The LLM is asked to respond with a single class label based on its internal reasoning and the structured prompt. This setup enables evaluation of whether LLMs can simulate regulatory logic even in zero-shot settings, and how their interpretability compares to other models.

All models were evaluated for classification metrics as well as runtime and memory usage using the *psutil* library. These experiments allow for direct comparison across model types in terms of practical deployment within a regulatory workflow.

7.4.5 Interpretability

Interpretability is a foundational requirement in regulated domains such as medical device classification, where transparency, auditability, and trust are not optional but mandated by law. In this study, we systematically evaluate the interpretability of each model—traditional machine learning, deep neural networks, and large language models (LLMs)—using methods appropriate to their structural characteristics. Our goal is to compare models not only on their predictive performance but also on how clearly and efficiently their decisions can be understood by domain experts.

The interpretability strategy adopted for each model type and input format is summarised in Table 7.6. The models are categorized into two broad classes: intrinsically interpretable models (e.g., Logistic Regression, Naive Bayes), and black-box models requiring post-hoc explanation (e.g., SVM, XGBoost, CNNs, and LLMs) [249]. For intrinsically interpretable models, we leverage native explanations: Logistic Regression coefficients directly represent global feature importance across classes, while Naive Bayes allows inspection of class-conditional likelihoods $P(x_i \mid y)$. To enhance local interpretability, we supplement with LIME and SHAP visualizations [236, 242].

For SVMs, which lack transparency due to kernelization, we use LIME and KernelSHAP to approximate local feature attributions through perturbation-based sampling [236]. Permutation-based global importance is also computed to aid summary-level interpretation. For XGBoost, we adopt TreeSHAP [252], which provides exact and efficient Shapley values tailored for tree ensembles, enabling interpretable deployment in production settings.

For deep learning models such as CNNs, RNNs, and DPCNNs, we apply input-specific post-hoc techniques: Integrated Gradients [241] and LIME-Text for token-level attribution [242]. These methods reveal which parts of the input text contribute most to the model's decision.

For LLMs (e.g., DeepSeek, LLaMA), we employ token-level LIME. Although SHAP becomes infeasible at this scale, LIME offers a lightweight approximation of feature influence. We also visualize attention maps where relevant [248], though we acknowledge their limitations in reliably reflecting causal importance. These transformer-based architectures present unique interpretability challenges: LIME requires significant sampling, explanations are sensitive to perturbation, and attribution outputs (e.g., attention scores or token masks) may not align with human semantic understanding. These issues are critical in regulatory applications where explanations must be stable, auditable, and human-comprehensible.

We match input formats to model capabilities for optimal performance and interpretability. Traditional models like Logistic Regression and Naive Bayes operate effectively on vectorized representations (e.g., TF-IDF or BERT embeddings), which preserve global lexical patterns and are interpretable through coefficient-based weights [242]. Tree-based models (XGBoost, Random Forest) benefit from high-dimensional embeddings like BERT due to their ability to capture non-linear interactions without explicit sequence modeling [252]. For neural architectures (CNNs, RNNs, LLMs), raw or tokenized text inputs are preserved to leverage their capacity for sequential feature learning and context modeling. Particularly for SVMs and LLMs, using raw text enables token-level attribution via LIME-Text or attention-based methods, which are more faithful to the linguistic structure of the input [248].

Table 7.6: Interpretability Strategy by Model Type and Input Format

Model Type	Naïve Bayes
Input Format	TF-IDF
Primary Method	Direct inspection of *feature_log_prob_*
Supplementary Tools	LIME, SHAP
Model Type	Logistic Regression
Input Format	TF-IDF or BERT embeddings
Primary Method	Anchors, Coefficients
Supplementary Tools	LIME, SHAP
Model Type	SVM
Input Format	Raw text
Primary Method	LIME-Text
Supplementary Tools	Anchors, KernelSHAP, Permutation
Model Type	XGBoost / Random Forest
Input Format	BERT embeddings
Primary Method	TreeSHAP
Supplementary Tools	Permutation Importance
Model Type	CNN / RNN / DPCNN
Input Format	Raw tokenized text
Primary Method	Integrated Gradients
Supplementary Tools	LIME-Text

7.5 Experiments and Results

We evaluate all models across three dimensions: predictive accuracy, interpretability, and computational cost. These dimensions are essential in the regulatory context, where decisions must be both reliable and explainable, while also being efficient for deployment.

7.5.1 Model Accuracy

To assess classification performance, we compute the following metrics using 10-fold stratified cross-validation on the 42,000-device dataset. This protocol applies to the traditional machine-learning and deep-learning models; the LLMs were not evaluated under cross-validation but were instead assessed zero-shot on the 1,000-device stratified subset described above:

- **Weighted F1 Score**: Measures the harmonic mean of precision and recall, weighted by support.
- **Weighted Precision and Recall**: Captures the balance between false positives and false negatives across classes.

- **Area Under ROC Curve (AUROC)**: One-vs-Rest strategy for multi-class evaluation.
- **Accuracy**: The proportion of correctly predicted instances.

Based on overall accuracy as shown in Table 7.7 and Table 7.8, the CNN model outperformed all other approaches, achieving an accuracy of 88.28%, followed by XGBoost (86.00%) and Random Forest (84.00%). Among traditional machine learning models, XGBoost demonstrated the most balanced performance across metrics, with a Macro F1 of 0.69 and strong performance in high-risk Class III devices (F1 = 0.87). Logistic Regression remained the most computationally efficient (0.0208s) while maintaining reasonably competitive performance (Accuracy = 82.00%, Macro F1 = 0.62), making it attractive for scenarios where transparency and low latency are prioritized.

Within deep learning, CNN clearly outperformed RCNN, RNN, and DPCNN, achieving the highest overall F1 (0.6768) and superior Class III performance (F1 = 0.80). Other architectures struggled to generalize as well, particularly on Class I devices, with RNN and DPCNN showing Class I F1 scores below 0.15, reinforcing that CNNs are better suited for localized textual pattern recognition in regulatory descriptions.

In contrast, large language models (LLMs) performed substantially worse on accuracy (DeepSeek R1: 45.93%), despite showing moderate recall for Class II and III devices (around 60–70%) as shown in Table 7.9. This suggests that zero-shot LLMs can approximate regulatory heuristics, but without domain-specific tuning, they fall short in precision and struggle with low-risk class identification (Class I F1 < 0.15).

Table 7.7: Performance of Traditional Machine Learning Models. Best performance in each metric is highlighted in bold.

Model	Accuracy	Macro F1	Macro Precision	Macro Recall	Inference Time (s)
SVM	0.840	0.670	0.650	**0.690**	82.870
Logistic Regression	0.820	0.620	0.600	0.670	**0.020**
Random Forest	0.840	0.670	**0.760**	0.630	0.230
XGBoost	**0.860**	**0.690**	0.740	0.660	0.050
Naive Bayes	0.520	0.400	0.450	0.510	0.140

Model	Class I F1	Class II F1	Class III F1	Class I Recall	Class II Recall	Class III Recall
SVM	0.310	0.840	0.850	**0.390**	0.830	0.850
Logistic Regression	0.200	0.810	0.830	0.370	0.800	0.830
Random Forest	0.320	0.840	0.840	0.220	**0.860**	0.820
XGBoost	**0.340**	**0.850**	**0.870**	0.260	0.850	**0.870**
Naive Bayes	0.030	0.560	0.610	0.500	0.500	0.530

Table 7.8: Performance of Deep Learning Models. Best performance in each metric is highlighted in bold.

Model	Accuracy	Macro F1	Macro Precision	Macro Recall	Inference Time (s)
CNN	**0.880**	**0.680**	**0.710**	**0.660**	0.003
RCNN	0.760	0.500	0.520	0.510	0.003
DPCNN	0.660	0.440	0.500	0.440	**0.002**
RNN	0.740	0.490	0.500	0.500	0.003

Model	Class I F1	Class II F1	Class III F1	Class I Recall	Class II Recall	Class III Recall
CNN	**0.500**	**0.730**	0.800	**0.390**	**0.750**	0.830
RCNN	0.200	0.590	0.720	0.190	0.610	0.730
DPCNN	0.120	0.420	0.750	0.140	0.380	**0.800**
RNN	0.140	0.520	**0.760**	0.120	0.520	**0.800**

Table 7.9: Performance of Large Language Models (LLMs) on 1000 Samples. Best performance in each metric is highlighted in bold.

Model	Accuracy	Macro F1	Macro Recall	Inference Time (s)
DeepSeek R1 7B	**0.460**	**0.250**	**0.390**	3.990
LLaMA 3.1 8B	0.460	0.250	0.290	**0.490**

Model	Class I F1	Class II F1	Class III F1	Class I Recall	Class II Recall	Class III Recall
DeepSeek R1 7B	**0.150**	**0.580**	0.570	**0.140**	**0.600**	**0.720**
LLaMA 3.1 8B	0.130	0.560	0.570	0.120	0.550	0.700

7.5.2 Model Interpretability

In our interpretability experiments, we applied a variety of methods to evaluate different model types and their interpretability mechanisms. Intrinsic models (e.g., logistic regression, naive Bayes) provided stable and globally coherent explanations by design. Feature coefficients and class-conditional likelihoods were directly accessible and aligned with domain expectations, requiring minimal computational overhead.

For Naive Bayes, which uses TF-IDF as the input format, we employed direct inspection of feature log probabilities as the primary method. The results are in Table 7.10. The log probability tells us the likelihood of a feature appearing in a particular class. A more negative value indicates a less likely occurrence for that feature in that class. For example, "cutting" with a log probability of -7.7423 in Class I suggests that this term is relatively common for Class I devices, while features with less negative log probabilities are more closely tied to the class.

Logistic Regression, which utilizes either TF-IDF or BERT embeddings, was examined through Anchors as shown in Figure 7.7. For Class I, the anchor words identified were "instruments" and "used", with a high precision (96.8%) but low coverage (12.3%), suggesting that while these words are reliable indicators of low-risk instruments, they appear only in a limited subset of Class I samples. For Class II, the anchor word "version"

Class	Rank	Feature	Importance Score
Class I	1	cutting	100.0
Class I	2	parts	94.1
Class I	3	orthodontic	91.9
Class I	4	laminoscopes	90.3
Class I	5	staining	89.0
Class I	6	usually	85.5
Class I	7	transfer	85.5
Class I	8	linear	85.3
Class I	9	components	83.6
Class I	10	non	83.6
Class II	1	disposable	100.0
Class II	2	consists	89.2
Class II	3	product	89.2
Class II	4	used	84.8
Class II	5	tube	76.7
Class II	6	stapler	71.3
Class II	7	medical	70.2
Class II	8	needle	55.6
Class II	9	body	48.9
Class II	10	type	48.0
Class III	1	product	100.0
Class III	2	needle	68.3
Class III	3	catheter	62.6
Class III	4	screws	61.2
Class III	5	infusion	47.6
Class III	6	consists	45.7
Class III	7	used	42.5
Class III	8	bone	36.6
Class III	9	sterilized	36.2
Class III	10	tube	32.9

Table 7.10: Naive Bayes Model: Top 10 Most Predictive Features by Risk Class. Importance scores are normalized within each class (highest = 100), derived from class-conditional probabilities $P(\text{feature}|\text{class})$.

exhibited both high precision (98.7%) and substantial coverage (50.6%), indicating that references to software versions or firmware updates are not only distinctive but also prevalent among Class II devices. For Class III, the model consistently relied on the anchor word "box" (e.g., control box), with a precision of 97.5% and coverage of 51.1%, reflecting the common presence of control hardware in high-risk, complex systems.

This comparative analysis highlights how the model's interpretability varies by class: Class I relies on sparse but specific cues, while Class II and III benefit from broader, high-impact lexical anchors. These findings reinforce the importance of class-specific vocabulary in medical device classification and support the model's use in regulatory decision-making contexts.

For Support Vector Machines (SVM), we can work with just raw text input, focusing on LIME-Text as the primary interpretability tool. As shown in Figure 7.6, the text is analyzed by highlighting the most influential words, which are identified based on their contribution to the predicted class. In this specific example, the model predicts the probabilities for three possible classes: Class I, Class II, and Class III. The prediction probabilities suggest that Class III has the highest likelihood (0.57), followed by Class II (0.41), and Class I (0.02).

The words "resorbable", "fracture", "internal", "fixing", and "screw" are highlighted according to their significance in the decision-making process. In the context of Class II, words like "fixing" have higher weights, which influence the model's prediction that this sample belongs to Class II. However, the word "screw" is more strongly associated with Class III, indicating that it contributes to a higher probability for Class III. By using LIME-Text, we can visually explain how each word affects the model's output and interpret which parts of the input are driving the classification, helping to make the SVM's decision-making process more transparent and interpretable.

In the case of tree-based models like XGBoost and Random Forest, we used BERT embeddings and TreeSHAP as the primary method, as shown in Figure 7.9. When using BERT embeddings, we visualize the embedding vectors rather than specific words. This adds an abstraction layer to the explanation, as BERT embeddings represent a more complex, high-dimensional space that captures the semantic relationships between words. Therefore, when visualizing BERT embeddings, we interpret the relationships between these vector representations, which may not directly correspond to individual word-level contributions. This distinction impacts how we interpret the importance of different parts of the text, as we are not directly attributing the importance to individual words but rather to the abstracted embedding representations.

Explanations for black-box models such as SVM, XGBoost, and deep neural networks depended heavily on the quality of local approximations. Although LIME successfully generated sparse, human-readable explanations for individual predictions across all models, its stability varied. For example, SVM explanations were occasionally sensitive to sampling randomness due to decision boundary complexity. In the case of XGBoost, TreeSHAP yielded highly stable and faithful explanations, often outperforming LIME in both consistency and global insight.

Although CNNs and RNNs are considered "black-box" models, they still process raw tokenized text (i.e., inputting text directly into the model after tokenization), and Integrated Gradients is used to explain their predictions. Integrated Gradients helps to understand the importance of individual tokens (words) by attributing the prediction to each token based on how much it affects the output when moving from a baseline to the actual input. This method is particularly effective for deep learning models that process raw text because it works well with the continuous nature of deep models, offering insights into how each word influences the model's decision as shown in Figure 7.8.

Finally, LLMs exhibited the greatest interpretability for understanding why the model predicts that a certain device belongs to a particular classification. Due to their ability to process and generate natural language, LLMs provide intuitive, human-readable explanations, making the reasoning behind their decisions more transparent. They are able to explain why a device might belong to a certain class based on contextual understanding and domain knowledge, making it easier to trace the logic of their predictions as shown in Figure 7.10.

However, despite the superior interpretability of LLMs, the accuracy results indicate a different story. We observed that the accuracy rate of the LLM-based model was relatively low compared to other models like SVM or XGBoost as shown in Table 7.9. This suggests that while LLMs aim to provide insightful and explainable reasoning for their decisions, they do not always offer the correct legal reasoning.

While LLMs can generate human-readable rationales when prompted to explain their predictions, these explanations are self-reported justifications rather than mechanistic insights into the model's internal computations. We obtain these rationales by explicitly asking the LLM to provide reasoning alongside its classification output; however, research has shown that such prompted explanations may not faithfully reflect the actual decision process [265]. The model may confabulate plausible-sounding reasoning that does not correspond to the features it actually relied upon. Therefore, LLM

rationales should be viewed as auditable narratives for human review rather than accurate accounts of model behavior.

Given these limitations, their performance trade-offs remain substantial. In practical deployments, models such as XGBoost and CNN currently offer the best balance between accuracy, class-level reliability, interpretability (via SHAP or Integrated Gradients), and computational efficiency.

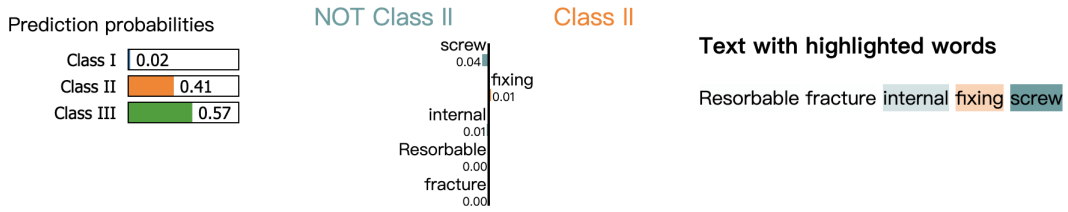


Figure 7.6: SVM LIME Explanation: Text is analyzed by highlighting the most influential words based on their contribution to the predicted class.

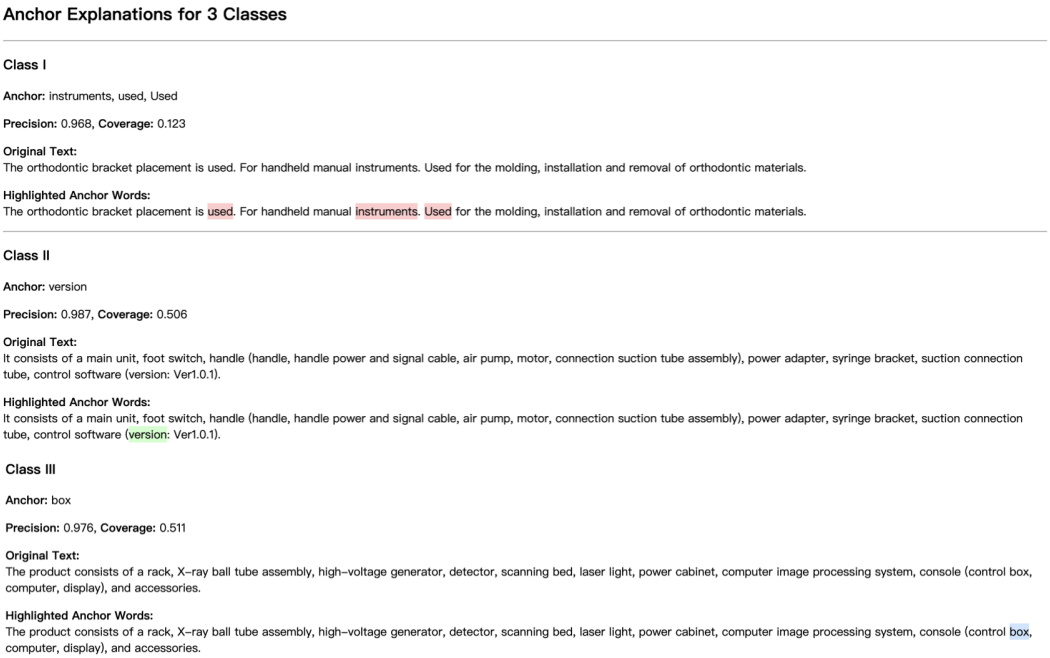


Figure 7.7: Logistic Regression Anchor Explanation: Anchor words identified for each class with precision and coverage metrics.

While interpretability is often cited as a key requirement in medical regulatory AI systems, there exists no universally accepted benchmark or ground truth to evaluate the quality of explanations. We benchmarked SHAP, LIME, Anchors, and Integrated Gradients across multiple models (Logistic Regression, SVM, XGBoost, etc.), yet acknowledge that explanation quality is inherently model- and context-dependent. In regulatory

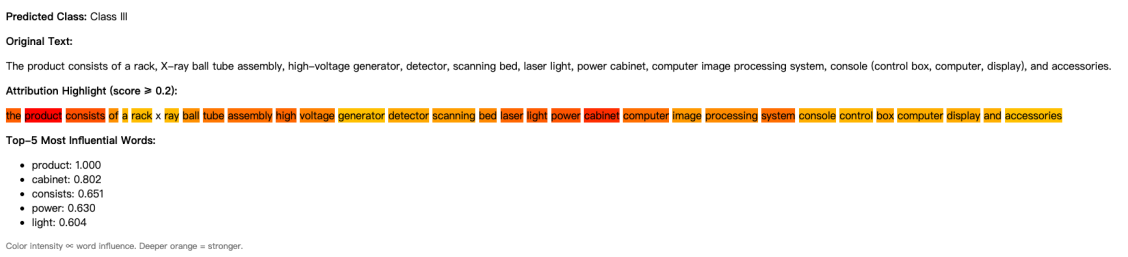
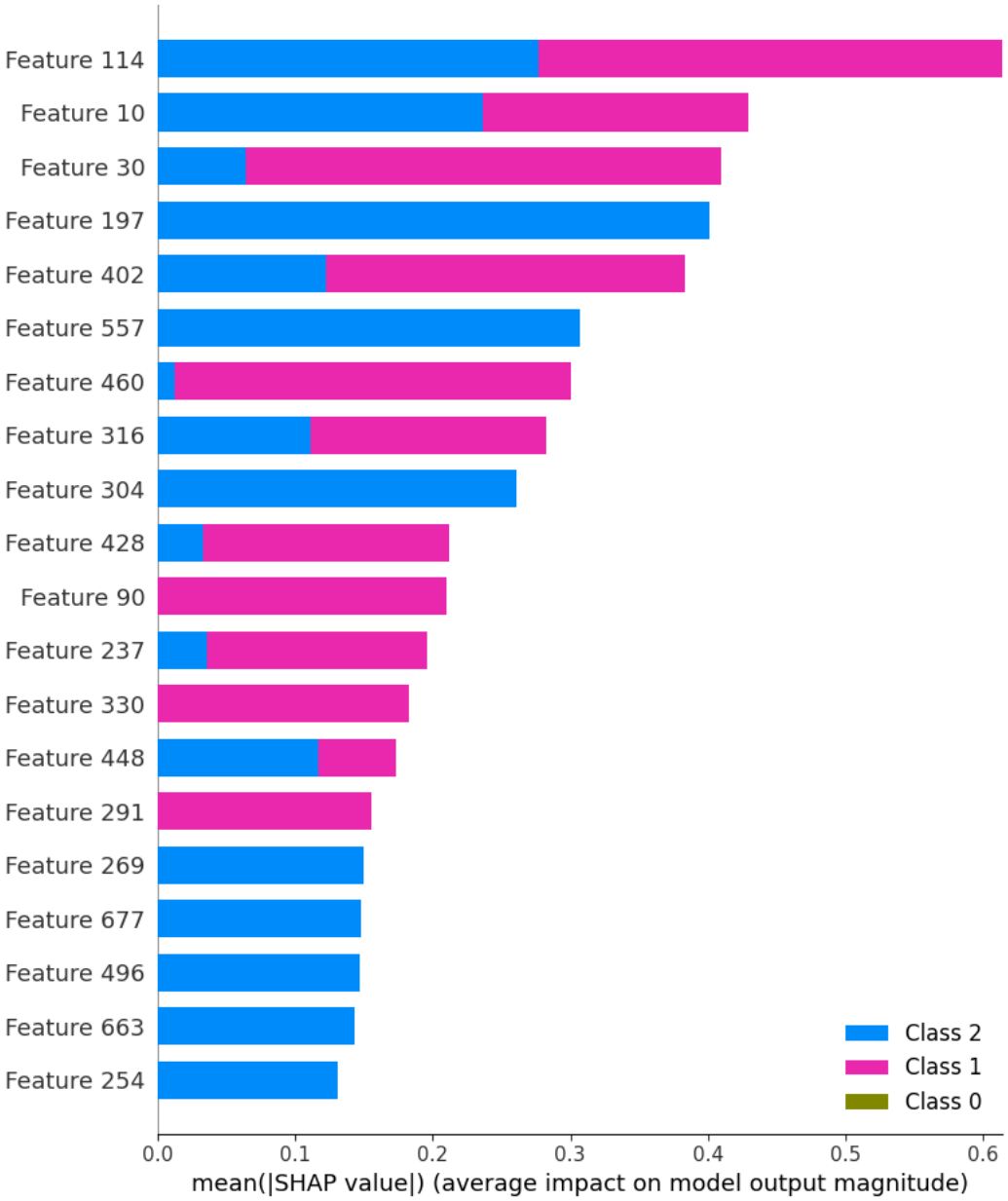


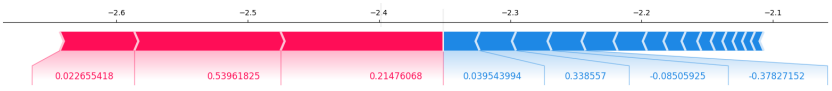
Figure 7.8: RNN Integrated Gradients Explanation: Token-level attribution showing how each word influences the model's decision.

practice, the primary goal is not necessarily to achieve the most accurate explanation, but to ensure that model decisions align with domain knowledge and are auditable.

Evaluating interpretability remains an open challenge, particularly in regulated AI systems where no ground-truth explanations exist. Rather than claiming superiority of a single method, we adopt a multi-method strategy to triangulate model behavior. While models such as XGBoost and Random Forest demonstrate strong performance when operating on BERT embeddings, their interpretability remains limited due to the nature of these inputs. Specifically, methods like TreeSHAP identify the contribution of individual embedding dimensions to model predictions, but these dimensions do not correspond directly to human-readable words or phrases. This introduces an abstraction layer between the model's internal logic and the regulator's semantic understanding of why a prediction was made. We acknowledge this limitation: TreeSHAP explanations based on BERT embeddings may be difficult for legal and regulatory specialists to interpret directly, as they require understanding of high-dimensional vector representations rather than plain text. Nevertheless, we employ TreeSHAP for tree-based models because it provides mathematically consistent and computationally efficient feature attributions, which are valuable for model developers and technical auditors. For end-user interpretability in regulatory settings, we recommend complementing TreeSHAP with simpler, text-based explanation methods such as LIME-Text, which directly highlights influential words in the input. This gap poses real challenges for regulatory auditability and decision traceability. Regulatory professionals typically require transparent, token-level justifications grounded in recognizable medical or functional terminology. However, the latent representations learned by deep models—while powerful in capturing contextual nuances—lack this transparency. To address this, we propose hybrid interpretability strategies. For instance, embedding-based models can be complemented by manually engineered features that capture domain-relevant tokens (e.g., risk indicators or functional keywords), enabling parallel inspection. Alternatively, simplified proxy models



(a) Random Forest BERT TreeSHAP



(b) XGBoost BERT TreeSHAP

Figure 7.9: TreeSHAP-based explanations for BERT embeddings using ensemble models.

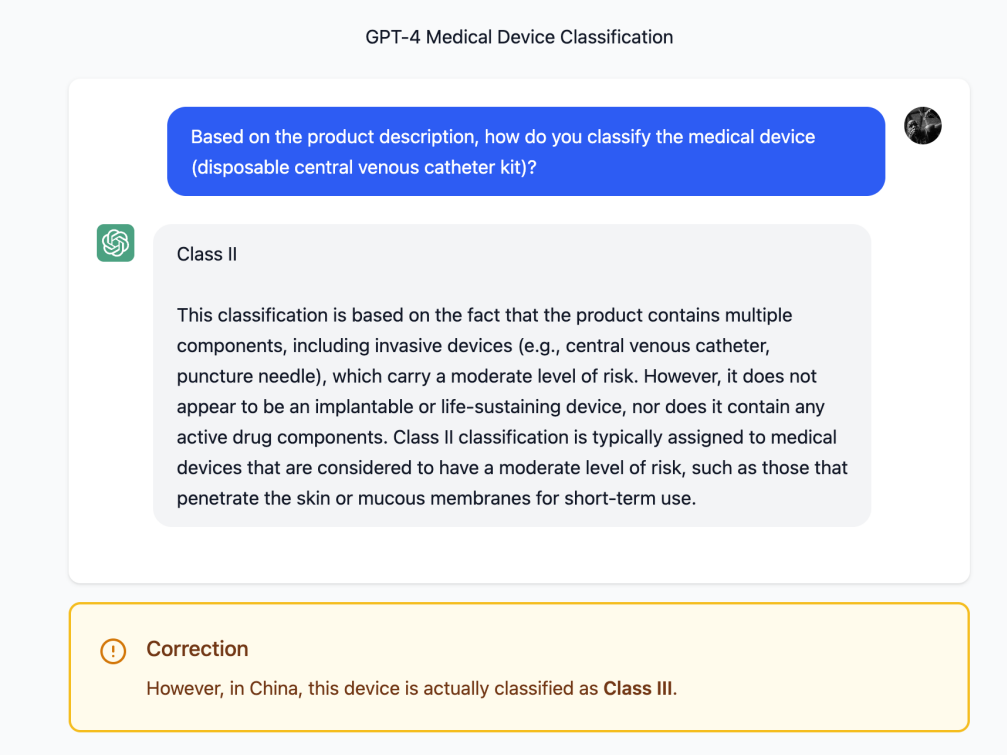


Figure 7.10: Example of Governance-Aware Interpretability in LLM-Based Medical Device Classification

(e.g., rule-based systems or logistic regression trained on key features) can be used alongside embedding models to triangulate decisions and provide more interpretable justifications for end-users. This hybrid approach may help balance the trade-off between representational richness and semantic clarity in high-stakes regulatory contexts.

7.5.3 Model Computational Cost

Model	Theoretical Complexity (Big O)	Observed Inference Time (s)	Inference Time Comments
Logistic Regression	$O(nd)$	0.0208	Extremely efficient in resource-constrained environments.
Naive Bayes	$O(nd)$	0.1371	Efficient with low inference time.
SVM	$O(n^2)$ to $O(n^3)$	82.8700	High latency due to kernel computation, not practical for real-time use.
Random Forest	$O(k \log n)$	0.2272	Fast inference with sublinear complexity for tree-based models.
XGBoost	$O(k \log n)$	0.0514	Efficient, balanced performance and efficiency.
CNN	$O(n \cdot d \cdot L)$	0.0024	Fast inference, high accuracy, suitable for practical deployment.
RNN	$O(n \cdot d \cdot L)$	0.0034	Fast inference, suitable for real-time applications.
DeepSeek R1	$O(n^2)$	3.9888	High inference time, not feasible for real-time regulatory tasks.
LLaMA	$O(n^2)$	0.4919	Relatively high latency, may be impractical for real-time systems.

Table 7.11: Comparison of Theoretical Complexity and Observed Inference Time for Different Models

We analyze computational cost by linking the theoretical complexity of each model with its observed inference time in practice. While theoretical analysis offers a general understanding of algorithmic efficiency, inference time provides an empirical measure of how these costs manifest in real-world medical classification tasks. Overall, the trend in inference time aligns with the Big O complexity.

Theoretical Complexity. Classical models such as Logistic Regression and Naive Bayes have linear time complexity $O(nd)$ for training and inference, where n is the number of samples and d is the number of features. In contrast, kernel-based Support Vector Machines (SVMs) scale poorly with training data, exhibiting worst-case complexity of $O(n^3)$ during training. Random Forests and XGBoost ensembles reduce complexity by parallelizing tree operations, yet their inference cost scales with the number of estimators. Deep learning models such as CNNs and RNNs require $O(n \cdot d \cdot L)$ operations for each input, with L denoting network depth. Finally, Large Language Models (LLMs) like DeepSeek and LLaMA incur substantially higher cost, both due to their parameter scale (billions of weights) and prompt-based inference mechanisms.

Empirical Evaluation. Inference times were measured for all models on the same hardware using 1000 medical device samples. As shown in Tables 7.7–7.8, classical models demonstrated superior efficiency. Logistic Regression achieved sub-second inference (0.0208 ± 0.0008 s) while maintaining competitive accuracy. XGBoost offered a favorable balance (0.0514 ± 0.0034 s) between performance and efficiency, owing to optimized tree traversal. Conversely, SVM suffered from extremely high latency (82.87 ± 0.59 s), rendering it impractical in real-time scenarios despite strong accuracy.

Deep neural models exhibited low inference latency across the board (ranging from 0.0024 to 0.0034 s), with CNNs achieving both the highest accuracy (0.8828) and fast response times. These models offer a desirable trade-off for practical deployment, particularly when interpretability is enhanced through methods like Integrated Gradients.

LLMs, while impressive in zero-shot generalization, revealed orders-of-magnitude slower inference. DeepSeek R1 required nearly 4 seconds per input (3.9888 s), and even optimized variants like LLaMA 3.1 consumed 0.4919 s per sample. This latency, combined with memory demands, poses significant barriers for integration into high-throughput regulatory pipelines.

Since there is no strict numerical benchmark for interpretability, and computational cost is likewise case-specific, we summarise each model's accuracy, interpretability and cost using the numerical rating criteria defined in Table 7.13.

7.6 Discussion

This study addresses a foundational challenge in regulatory informatics: how to systematically evaluate AI architectures for safety-critical classification when institutional

Table 7.12: Comparison of AI Models in Terms of Accuracy, Interpretability, and Computational Cost

Model	Accuracy (Stars)	Interp. (Stars)	Cost (Stars)	Application Scenarios
Rule-Based	5	5	5	Regulatory audits, legal compliance where transparency is critical.
Logistic Regression	4	3	3	Transparent predictions; business analytics or healthcare.
SVM	5	3	1	High-accuracy classification; image recognition or high-dimensional data.
Random Forest	5	1	3	High accuracy for tabular data; less suited for regulatory transparency.
XGBoost	1	3	3	Complex structured data; industrial tasks and ML competitions.
Gaussian Naive Bayes	3	3	5	Quick prototyping; exploratory data analysis or educational contexts.
DPCNN	5	3	5	Complex text classification with balance between accuracy and interpretability.
CNN	5	3	5	Text and sequence pattern recognition; practical deployment.
RCNN	4	3	5	Sequence labeling and moderate complexity NLP tasks.
RNN	4	3	5	Sequential data and NLP; speech recognition or time-series analysis.
DeepSeek R1 7B	1	5	1	Detailed reasoning tasks; in-depth analysis where transparency is crucial.
Ollama 3.1 8B	1	5	1	Complex decision-making; language understanding with transparency.

Table 7.13: Numerical Rating Criteria for Accuracy, Interpretability, and Computational Cost

Metric	1 Star	2–3 Stars	4–5 Stars
Accuracy	< 0.6	0.6–0.8	> 0.8
Interpretability	Black-box model; impossible to determine which features led to the decision. Only provides embeddings or labels.	Partial interpretability; possible to identify impactful features but full process may not be revealed.	Fully transparent; decision process entirely understandable with clear feature attribution.
Computational Cost	Slow ($>$1s)	Moderate (0.01–1s)	Fast ($<$0.01s)

mandates demand simultaneous optimization of accuracy, auditability, and operational sustainability—criteria that are technically non-reconcilable. Our governance-aware evaluation framework reveals that responsible AI deployment in regulatory contexts requires moving beyond performance-centric benchmarking to evidence-based frameworks grounded in regulatory science principles. We demonstrate that model selection is not a purely technical decision but a governance choice shaped by institutional priorities, compliance obligations, and deployment contexts. This finding has implications for how regulatory agencies worldwide approach AI adoption in medical device oversight and, more broadly, for governance of AI systems in safety-critical government functions.

Our framework exposes fundamental tensions between technical capabilities and regulatory imperatives. High-accuracy architectures (CNNs: 88.28%) demand post-hoc explainability infrastructure that may not align with regulatory auditability standards;

interpretable models (Logistic Regression) sacrifice predictive performance but offer inherent transparency valued in legal proceedings; efficient models must balance computational constraints against accuracy requirements for patient safety. Critically, aggregate performance metrics obscure risk-stratified implications with direct consequences for regulatory governance.

Support Vector Machines (SVMs) delivered reliable performance on structured embeddings and raw text, especially under well-separated class distributions. However, their lack of native interpretability and sensitivity to kernel selection render them less ideal for regulatory settings [266]. Naive Bayes models, while computationally efficient and interpretable via log-likelihood inspection, assume conditional independence, a limitation when dealing with correlated language features typical in medical descriptions. Tree-based models exhibited distinct strengths. Decision Trees are transparent but prone to instability; in contrast, Random Forests and XGBoost significantly improve generalization through ensembling and boosting, respectively. When paired with TreeSHAP [252], these models offer both high accuracy and strong local/global interpretability, making them highly suitable for medical classification tasks that demand accountability. Neural network models, including CNNs and RNNs, outperformed simpler models on complex inputs, capturing subtle semantic and syntactic patterns. Yet their black-box nature and sensitivity to input perturbations challenge post-hoc explanation tools like LIME and Integrated Gradients [241]. Our experiments confirmed that even token-level explanations can vary under slight changes, aligning with recent concerns about explanation stability [267]. Transformer-based Large Language Models (LLMs) such as DeepSeek offer state-of-the-art performance in language understanding, but at the cost of interpretability and computational overhead. Despite LIME-Token approximations and attention map visualizations, their behavior remains opaque, echoing critiques that current methods may fall short of faithful explanation [248].

While a 5% difference in overall accuracy may seem negligible in aggregate metrics, its practical consequences can be substantial when disaggregated by class. For example, the Naive Bayes model achieves a macro accuracy of only 59%, yet its Class III F1 score reaches 0.63, far lower than CNN (0.81) and XGBoost (0.85). More strikingly, its F1 score for Class I is as low as 0.05, indicating an almost complete failure to identify low-risk devices correctly. In a large-scale deployment scenario processing 10,000 devices, this could lead to hundreds of Class I devices being incorrectly escalated into higher regulatory categories, burdening both manufacturers and regulators with unnecessary review processes. Conversely, its misclassification of Class III devices as lower risk

poses serious safety and legal risks. This demonstrates that even modest differences in average performance may obscure critical weaknesses in specific classes, particularly when class distributions are imbalanced or when the cost of misclassification is not symmetric across regulatory categories. Therefore, per-class F1 scores offer a more meaningful lens than macro-averages alone for understanding real-world trade-offs in regulatory decision-making. Beyond presenting aggregate metrics such as accuracy or macro F1, it is essential to contextualize these trade-offs in terms of real-world regulatory consequences. For example, a 5% difference in overall accuracy between CNN (88.3%) and XGBoost (84.0%) translates to 50 additional misclassifications per 1,000 devices. However, the severity of this gap depends on which classes are most affected. As shown in our per-class evaluation, models with lower overall performance may still exhibit reasonable recall for high-risk Class III devices (e.g., Logistic Regression Class III F1 = 0.80), but fail entirely on low-risk Class I devices (F1 = 0.12). This imbalance can lead to misallocations of regulatory attention, either through under-regulating high-risk devices or overburdening low-risk ones. These risks are asymmetric and carry different legal and safety implications, further reinforcing the importance of per-class F1 scores as a core diagnostic metric.

Interpretability is not monolithic, and different models offer varying kinds of transparency. Logistic Regression provides interpretable decision boundaries via feature coefficients, allowing regulators to see which input terms directly influenced predictions, which is a valuable asset for traceability during audits. Models like XGBoost rely on post-hoc methods such as SHAP to estimate feature importance. While these techniques are powerful, they can be challenging for non-technical reviewers to interpret, as the explanations often reference abstract embedding dimensions rather than specific, identifiable words that directly trigger a given classification. LLMs, on the other hand, generate natural language rationales that may appear intuitive and persuasive, but do not always align with legally valid reasoning. This presents a nuanced trade-off: explanations must not only be understandable but also legally and procedurally correct. A natural-language explanation that is plausible but fails to cite a critical regulatory criterion may increase trust while introducing hidden error. Therefore, the trade-off is not merely accuracy versus interpretability, but accuracy versus relevant and valid interpretability.

Although our experimental comparison emphasizes inference time as a core item, other components of computational cost deserve consideration. A comprehensive Total Cost of Ownership (TCO) analysis would account for training and fine-tuning costs

(particularly prohibitive for LLMs and deep neural networks, requiring specialized hardware and extensive optimization), dataset preparation costs (labor-intensive annotation requiring domain expertise), deployment infrastructure costs (cloud-based GPUs for complex models versus local hardware for simpler architectures), and ongoing maintenance costs (regular validation against evolving regulatory standards and product taxonomies). For instance, while CNNs achieve 88.28% accuracy, their initial training on large medical device corpora may require weeks of GPU time and substantial energy consumption, whereas XGBoost models can be trained in hours on standard hardware. Similarly, LLMs impose recurring API costs or GPU infrastructure expenses, while Logistic Regression deployment requires only minimal computational resources. Our current analysis focuses on inference latency as a proxy for operational efficiency, but future work should develop comprehensive TCO models incorporating lifecycle costs across training, deployment, and maintenance phases. Infrastructure compatibility also varies substantially: simpler models like Logistic Regression or XGBoost can be deployed on secure, local hardware, making them more viable for agencies with limited resources or strict data governance requirements, whereas high-complexity models (e.g., LLMs) may require cloud-based GPUs, which pose additional data privacy concerns and increase ongoing expenses. As such, computational cost must be viewed holistically, encompassing not just speed but also sustainability, infrastructure compatibility, and maintainability over time.

The empirical trade-offs identified in this study between accuracy, interpretability, and computational cost have direct implications for compliance with established regulatory frameworks governing AI-enabled medical devices. We synthesize our benchmark results with the core requirements of major regulatory instruments to demonstrate how model selection aligns with real-world deployment readiness.

A critical aspect of governance-aware AI deployment is recognizing that while the core task (medical device risk classification) remains identical across jurisdictions, the regulatory requirements governing the AI system itself vary substantially by region and legal framework. This creates a dual-layer compliance challenge: the classification system must not only perform the technical task accurately but must also satisfy jurisdiction-specific legal obligations regarding AI transparency, validation, and lifecycle management. For example, an AI classification system deployed in the European Union must comply with the EU AI Act's transparency mandates (Article 13), which require that users can "interpret the system's output and use it appropriately." In contrast, the same classification task deployed in the United States under FDA oversight must satisfy

the Good Machine Learning Practice framework's emphasis on predetermined change control plans (Principle 9) and algorithm performance monitoring with full traceability. Meanwhile, deployment in China requires adherence to NMPA's Guiding Principles for AI Medical Software (2022), which emphasize dataset representativeness and algorithm transparency specific to Chinese medical terminology and classification heuristics.

Consequently, model selection cannot be guided solely by predictive accuracy on a fixed dataset. Instead, deployment decisions must account for the specific regulatory environment in which the system will operate. A model that achieves high accuracy but lacks explanation traceability may be viable under one regulatory regime yet inadmissible under another. Similarly, a simpler model with slightly lower accuracy may be preferable in contexts that prioritize auditability and lifecycle maintainability over marginal performance gains. This jurisdictional variability underscores the need for governance-aware evaluation frameworks that explicitly map technical model characteristics to regulatory compliance requirements.

Critically, while our empirical analysis uses China NMPA data, the fundamental trade-offs between accuracy, interpretability, and computational cost represent universal properties of machine learning architectures rather than dataset-specific artifacts. The performance characteristics we observe, including CNNs excelling at pattern recognition but requiring post-hoc explainability, tree-based ensembles balancing accuracy with inherent interpretability, and linear models offering transparency at the cost of representational capacity, reflect intrinsic algorithmic properties that generalize across regulatory jurisdictions. Whether classifying devices for FDA, EU, or NMPA, regulators face identical technical constraints: neural networks remain black boxes requiring explanation infrastructure regardless of the language or regulatory framework; XGBoost's TreeSHAP explanations retain their auditability advantages whether applied to English FDA submissions or Chinese NMPA entries; Logistic Regression's coefficient-based transparency is equally valuable for European MDR compliance or US premarket review. What varies across jurisdictions is not the nature of these trade-offs, but rather the institutional prioritization among competing criteria and the specific legal frameworks governing AI system deployment. This distinction is critical for understanding the generalizability of our findings: the governance framework we propose is jurisdiction-agnostic precisely because it addresses universal ML properties, while our regulatory mapping (Table 7.14) demonstrates how these universal trade-offs manifest differently under jurisdiction-specific compliance requirements.

Table 7.14 presents a systematic mapping of our key findings to the transparency, validation, and risk management obligations specified in the EU AI Act (Articles 13 & 15), FDA Good Machine Learning Practice (GMLP), EU Medical Device Regulation (MDR Annex I), ISO standards for clinical investigation and medical device software (ISO 14155, ISO 62304), and China NMPA regulatory frameworks. Each row demonstrates how a specific regulatory instrument's requirements interact with empirical model characteristics (accuracy, interpretability, computational cost) to yield jurisdiction-specific deployment guidance. For instance, the EU AI Act's stringent transparency requirements favor models like XGBoost paired with TreeSHAP, which provide traceable feature attributions aligned with regulatory decision heuristics. In contrast, FDA GMLP's emphasis on change control and lifecycle traceability makes simpler models like Logistic Regression attractive, as they facilitate transparent algorithm updates and validation documentation. This synthesis provides actionable guidance for regulatory professionals and developers seeking to integrate AI classification tools into conformity assessment workflows while maintaining compliance with applicable legal frameworks.

Table 7.14: Mapping empirical findings to regulatory framework requirements

Framework	Regulatory Requirement	Empirical Finding	Deployment Guidance
EU AI Act Art. 13 & 15 [268, 269]	High-risk AI systems must provide transparency enabling users to "interpret the system's output and use it appropriately" (Art. 13); technical documentation must include "data sets and testing processes" (Annex IV)	CNN (88.28%) and XGBoost (86.00%) achieve highest accuracy but lack inherent interpretability; require post-hoc explainability tools (TreeSHAP, Integrated Gradients, LIME)	Deploy hybrid architecture: high-accuracy models (CNN/XGBoost) paired with mandatory explainability layer; document explanation methodology in technical file per Annex IV

Continued on next page

Table 7.14 – continued from previous page

Framework	Regulatory Requirement	Empirical Finding	Deployment Guidance
FDA GMLP (2021) [270]	Good Machine Learning Practice requires "predetermined change control plan" (Principle 9) and "algorithm performance monitoring" with emphasis on explainability and validation traceability	Logistic Regression (82.00%, 0.02s latency) offers coefficient-based transparency; XGBoost (86.00%, 0.05s) provides TreeSHAP explanations with audit trail	Logistic Regression and XGBoost feasible for FDA's "locked algorithm" pathways with predetermined specifications; TreeSHAP explanations support Algorithm Change Protocol documentation
MDR Annex I (3) [97]	Manufacturers must provide "verifying that the device achieves the performance intended" with risk-based performance evaluation; clinical evaluation must demonstrate safety across intended patient populations	Per-class F1 varies substantially: Class I (0.05–0.50), Class II (0.56–0.85), Class III (0.63–0.87); aggregate accuracy obscures risk-stratified performance disparities	Conformity assessment should require class-specific performance metrics; manufacturers must justify model performance separately for each risk category; global accuracy insufficient for high-risk devices
ISO 14155 / 62304 [271]	ISO 14155 requires "documented rationale for sample size" (5.2.3); ISO/IEC 62304 mandates "software verification" with "documented evidence" (5.5) and change control procedures	Simpler architectures (Logistic Regression, Random Forest) facilitate lifecycle documentation, re-validation upon guideline updates, and maintenance of traceability matrices	Prefer lightweight models for regulatory environments requiring frequent re-validation; complex models (CNNs, LLMs) impose higher lifecycle maintenance costs incompatible with ISO 62304 change control

Continued on next page

Table 7.14 – continued from previous page

Framework	Regulatory Requirement	Empirical Finding	Deployment Guidance
China NMPA Classification Rules (2017) [23, 272]	Medical Device Classification Rules mandate risk-based categorization (Class I, II, III) based on intended use, structural characteristics, and operational mechanism; requires consistency with device description and intended purpose	Model trained on 42,000 NMPA UDI registry entries achieves 86% accuracy (XGBoost) with strong Class III performance (F1=0.87); model predictions align with NMPA's rule-based heuristics (implantation, life-support, invasiveness)	XGBoost with TreeSHAP provides explainable features matching NMPA classification criteria (e.g., "implantable," "catheter," "bone screw"); supports NMPA's emphasis on traceable decision logic for regulatory review and manufacturer guidance
China Medical Device Regulations (2021 Revision) [273]	Article 76 requires AI-enabled medical devices undergo algorithm validation with "performance evaluation data" and "clinical trial evidence" for Class III devices; emphasizes post-market surveillance and adverse event monitoring	High-risk Class III devices show significant performance variance across models: Naive Bayes (F1=0.63) vs. XGBoost (F1=0.87), translating to hundreds of potential misclassifications per 10,000 submissions	Regulatory submissions should include per-class validation metrics; XGBoost and CNN recommended for initial classification with mandatory expert review for Class III devices; establish performance monitoring protocols aligned with Article 76 requirements

Continued on next page

Table 7.14 – continued from previous page

Framework	Regulatory Requirement	Empirical Finding	Deployment Guidance
China AI Medical Device Guidelines [274]	Guiding Principles for AI Medical Software (2022) require "dataset representativeness," "algorithm transparency," and "continuous learning validation"; emphasizes explainability for clinical decision support systems	LLMs (DeepSeek: 45.93% accuracy) generate plausible but legally incorrect classifications, highlighting gap between language fluency and regulatory correctness; BERT-based models (CNN: 88.28%) trained on NMPA data demonstrate superior alignment with Chinese medical terminology	Avoid LLMs for autonomous classification; prefer supervised models (CNN, XGBoost) trained on representative NMPA datasets; implement human-in-the-loop review for ambiguous cases; document dataset composition and model updates per 2022 Guidelines

This framework mapping underscores that model selection is not solely a technical decision but a regulatory design choice. For instance, while CNNs may outperform Logistic Regression by 6% in overall accuracy, the latter's inherent interpretability and compliance-readiness may render it preferable in contexts prioritizing auditability and lifecycle traceability over marginal predictive gains.

Regulatory agencies such as the FDA, EU, and NMPA increasingly emphasize interpretability in AI-assisted healthcare. The European Union's AI Act mandates transparency for high-risk systems, and the FDA's Good Machine Learning Practice (GMLP) framework similarly requires explainable outputs [275, 276]. While the lack of a universally accepted benchmark for interpretability is well-documented in AI research, regulatory contexts demand more than abstract notions of explanation fidelity. In medical device regulation, interpretability is a practical necessity that underpins legal compliance, accountability, and trust in automated decision-making. We propose a domain-specific framework for evaluating explanation utility in regulatory affairs, structured around three interrelated dimensions: comprehensibility (whether explanations are presented in forms and language regulators can easily understand), justifiability (whether explanations provide traceable rationales aligned with legal or procedural standards), and trust alignment (whether explanations increase regulators' confidence by referencing attributes cited in official classification guidelines). However, our current evaluation

relies on computational metrics and qualitative analysis rather than systematic human expert assessment. Quantitative validation through structured regulatory professional evaluation and explanation stability testing under input perturbations would strengthen auditability claims. Combining high-accuracy models like CNN or XGBoost with justifiable explanation layers (TreeSHAP, Anchors) may offer a balance of performance and regulatory traceability, ensuring that AI tools generate legally and procedurally coherent justifications rather than merely statistical predictions.

The deployment of AI-driven classification systems in regulatory workflows has far-reaching implications for clinical safety, market access, and regulatory efficiency. Model selection directly influences not only classification accuracy but also the feasibility of integrating such systems into existing regulatory infrastructures. The primary concern in medical device regulation is ensuring patient safety through appropriate risk-based oversight. Misclassification errors, particularly false negatives where high-risk devices are incorrectly assigned to lower regulatory classes, can bypass critical review pathways, leading to unsafe products entering the market. Our results show that model performance varies substantially across risk classes. For instance, Naive Bayes achieves a Class III F1 score of only 0.63, compared to CNN (0.81) and XGBoost (0.85). In a deployment scenario processing 10,000 devices annually, this performance gap could result in hundreds of high-risk devices receiving insufficient regulatory scrutiny. Conversely, false positives impose unnecessary regulatory burden on manufacturers, potentially delaying access to beneficial technologies. Therefore, class-specific performance metrics, rather than global accuracy, must drive model selection in safety-critical contexts.

Regulatory agencies face increasing workload pressures due to growing device complexity and submission volumes. AI classification systems offer potential efficiency gains by automating initial triage, prioritizing high-risk cases for expert review, and standardizing classification logic across reviewers. Our benchmark shows that models like XGBoost and CNN achieve high accuracy (86.00% and 88.28%) with inference latencies below 0.1 seconds, making them viable for real-time deployment. However, deployment readiness extends beyond predictive performance to include integration with existing case management systems, compatibility with regulatory databases (e.g., NMPA UDI, FDA GUDID), and support for human-in-the-loop review workflows. Simpler models like Logistic Regression, despite lower accuracy (82.00%), offer advantages in terms of interpretability, auditability, and ease of regulatory approval, factors that may outweigh marginal performance gains in resource-constrained settings.

Real-world deployment requires addressing technical, organizational, and legal barriers. Technical challenges include data harmonization across heterogeneous device registries, handling multilingual product descriptions (e.g., Chinese NMPA data vs. English FDA submissions), and ensuring model robustness to adversarial inputs or gaming by manufacturers. Organizational challenges involve training regulatory staff to interpret and validate AI outputs, establishing governance protocols for model updates, and defining accountability when AI-assisted decisions lead to adverse outcomes. Legal challenges include compliance with data protection regulations (e.g., GDPR for EU agencies), managing intellectual property concerns around proprietary classification algorithms, and ensuring that AI tools satisfy legal requirements for decision transparency and contestability. Our study contributes to addressing these challenges by providing a systematic benchmark that regulatory authorities can use to evaluate candidate models against their specific operational constraints, risk tolerance, and interpretability requirements. Successful deployment requires not only high-accuracy models but also systems designed for auditability, maintainability, and alignment with regulatory mandates, considerations that our tri-criteria benchmark framework explicitly addresses.

Model evaluation typically revolves around trade-offs between accuracy, interpretability, and computational cost, but these technical objectives must ultimately be reframed in terms of regulatory outcomes. In high-stakes domains such as medical device oversight, the primary cost of algorithmic failure is not computational inefficiency or model opacity, but rather the misclassification of devices in ways that compromise patient safety or hinder innovation. We propose incorporating the cost of misclassification into model evaluation. A model that marginally reduces inference time or memory usage may still be unsuitable for deployment if it increases the risk of under-regulating a high-risk (Class III) device or wrongly rejecting a safe and compliant product. For instance, although Naive Bayes is computationally efficient, it performs poorly on Class I (F1: 0.05) and significantly underperforms on Class III (F1: 0.63) compared to CNN (F1: 0.81) and XGBoost (F1: 0.85). Such deficiencies could result in hundreds of devices per 10,000 being incorrectly categorized, leading to unnecessary regulatory burden or, worse, compromised safety assurance.

To make this trade-off analysis more tangible, we advocate for scenario-based stress testing. This involves constructing hypothetical but realistic classification scenarios and examining how different model types perform under these conditions. For example, consider a manufacturer submitting a device description for a novel AI-powered diagnostic tool employing an untested algorithm. The textual description is linguistically

complex and includes subtle, ambiguous claims about its intended use and performance. Based on the information provided, the device could plausibly be assigned to Class II or Class III. In this case, the cost of misclassification is asymmetrical: incorrectly assigning the device to Class II could bypass critical review stages, while a Class III designation might delay market entry unjustifiably. Traditional models may lack the semantic flexibility to interpret nuanced claims, while deep learning or LLM-based systems may provide richer language understanding. However, these models differ in their ability to offer transparent, audit-ready justifications for their predictions. In this scenario, interpretability becomes paramount, not only to trace model decisions but also to provide regulators with human-comprehensible rationales for why the device is high or medium risk. This approach moves beyond static evaluation metrics and grounds the model comparison in the actual decision-making pressures faced by regulatory authorities.

While large language models (LLMs) offer fluent and human-like justifications for classification decisions, our findings underscore a crucial limitation for regulatory applications: their outputs may exhibit language plausibility without satisfying legal correctness. For instance, as shown in Figure 7.10, an LLM classified a Class III device as Class II, despite articulating a seemingly well-reasoned explanation. This divergence arises because LLMs are trained on general language data and do not inherently encode jurisdiction-specific definitions such as the NMPA's explicit requirement that Class III devices typically involve life-sustaining, life-supporting, or high-risk implantable functions. This distinction matters in regulatory settings, where classification errors can have significant legal and safety implications. Unlike domain-trained models or symbolic systems explicitly grounded in rule-based logic, LLMs lack guaranteed alignment with formal regulatory taxonomies.

However, LLMs' natural language generation capabilities suggest a valuable complementary role in hybrid explanation systems. Rather than autonomous classification, LLMs could serve as explanation synthesizers in workflows where accurate models make decisions and LLMs translate technical outputs into regulatory-compliant documentation. For instance, a hybrid architecture could leverage XGBoost for classification (86% accuracy, F1=0.85 for Class III) with TreeSHAP feature attributions providing auditable explanations, then employ an LLM to generate structured compliance narratives. The LLM would be prompted with: (1) the XGBoost classification decision, (2) top-k TreeSHAP feature contributions (e.g., "implantable: +0.42, catheter: +0.31"), and (3) relevant NMPA regulatory text excerpts. The LLM output would be a formatted regulatory justification document citing specific device characteristics and applicable classification

rules, ready for human expert review. This approach combines XGBoost's accuracy and auditability with LLMs' capacity for generating legally-structured natural language, while maintaining human oversight of final decisions. Preliminary implementations could restrict LLM outputs through structured templates (e.g., "Based on features X, Y, Z with regulatory criteria A, B, C, the device is classified as Class [II/III] because..."), ensuring consistency with regulatory reasoning patterns. Such hybrid systems warrant future investigation as they may address a critical gap: translating machine learning outputs into the documentary formats required for regulatory submissions, legal audits, and manufacturer guidance, without compromising classification accuracy or auditability.

No single model excels across all criteria, underscoring the need for context-dependent strategies. As shown in Figure 7.11, AI model selection in real-world regulatory workflows is often contingent on the stage of evaluation and the specific demands of each decision point. For batch pre-screening of large submission volumes, lightweight models such as Logistic Regression or CNNs balance moderate accuracy with low computational overhead, enabling rapid triage. For initial automatic classification at scale, higher-accuracy models such as CNN or XGBoost capture complex linguistic patterns in product descriptions. For ambiguous cases where model confidence is low or multiple risk classes are nearly equally probable, workflows should transition to interpretable models such as Logistic Regression combined with SHAP explanations, ensuring that regulatory reviewers can understand the basis of decisions and intervene when needed. For high-risk product reviews (potential Class III devices), additional caution and traceability are required. We recommend using CNNs combined with TreeSHAP explanations and expert human review, enabling both predictive accuracy and accountability. When evaluating novel devices submitted by startup companies, regulatory experts may benefit from LLMs to generate natural language rationales, particularly in the absence of precedent or structured features. During legal audits or final approval documentation, where interpretability is paramount, simple yet transparent models such as Logistic Regression combined with Anchors explanations are most appropriate, allowing reviewers and legal stakeholders to trace exactly how classification decisions were made.

To operationalize this context-dependent guidance, we propose quantitative performance thresholds for each deployment scenario based on our empirical findings. For batch pre-screening scenarios processing thousands of daily submissions, the optimal trade-off prioritizes throughput over marginal accuracy gains: acceptable models must achieve inference latency < 0.01s per device (enabling real-time processing of high-volume queues) with accuracy $\geq 80\%$ to reduce manual review burden while

maintaining safety standards. Our results show CNN (88.28% accuracy, 0.006s latency) and RNN variants meet these criteria, providing sufficient accuracy for initial triage without creating computational bottlenecks. For high-risk device review (Class III candidates), accuracy and class-specific recall become paramount: models must achieve Class III F1 ≥ 0.85 to minimize false negatives that could allow unsafe devices to bypass rigorous scrutiny, while maintaining explainability sufficient for expert validation. XGBoost (F1=0.87 for Class III, TreeSHAP explanations) and CNN (F1=0.81, Integrated Gradients) satisfy these criteria, whereas Naive Bayes (F1=0.63) fails to meet the safety threshold despite computational efficiency. For legal audit and compliance documentation scenarios, interpretability constraints dominate: models must provide inherently transparent decision logic (coefficient-based or rule-based explanations) with inference latency < 0.03s to support interactive review sessions, accepting accuracy trade-offs down to 80% if transparency is guaranteed. Logistic Regression (82% accuracy, 0.02s latency, direct coefficient inspection) and tree-based models with Anchors explanations satisfy auditability requirements, enabling legal counsel to trace decisions to specific textual features and regulatory criteria without relying on post-hoc approximations. These quantitative thresholds translate the abstract tri-criteria framework into actionable procurement specifications, enabling regulatory agencies to define minimum performance requirements aligned with operational contexts and institutional risk tolerance.

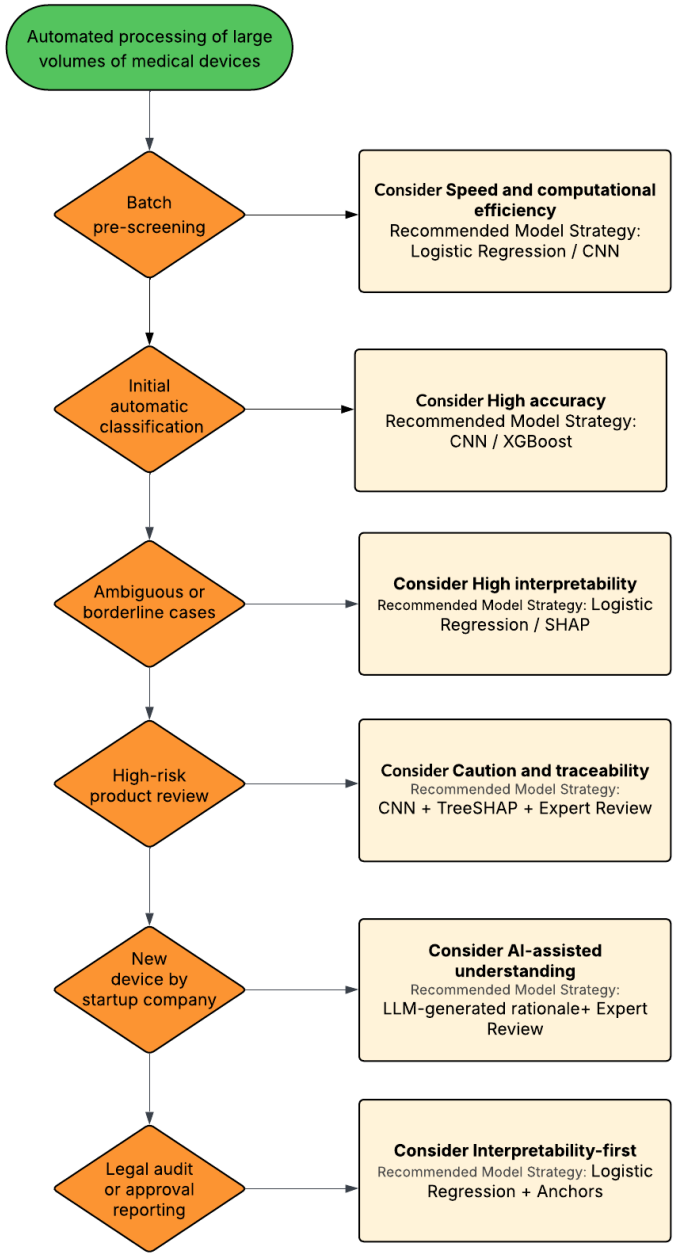


Figure 7.11: AI Algorithm Decision Tree in Different Regulatory Affairs Scenarios

This work contributes to an emerging agenda in regulatory science: developing governance frameworks for responsible AI deployment in safety-critical government functions. Our findings demonstrate that technical performance metrics alone are insufficient for evaluating AI systems in regulatory contexts. Governance considerations such as legal defensibility, institutional accountability, and operational sustainability must be systematically integrated into evaluation frameworks. This has broader implications beyond medical device classification, extending to any regulatory domain

contemplating AI adoption (pharmaceutical review, environmental monitoring, financial oversight). We identify three key insights for regulatory science practice. First, aggregate metrics mask regulatory risk; overall accuracy obscures per-class performance disparities with direct patient safety consequences, requiring regulatory agencies to evaluate AI systems using risk-stratified metrics aligned with governance priorities. Second, interpretability is multidimensional; different stakeholders (reviewers, legal counsel, auditors) require different explanation modalities, necessitating hybrid interpretability strategies rather than singular explainability methods. Third, deployment contexts shape architectural appropriateness; no single AI architecture is universally optimal, and model selection must adapt to institutional constraints, compliance frameworks, and operational workflows.

The value proposition of AI-assisted classification lies not in replacing expert judgment, but in intelligent triage and workload optimization. Specifically, automated systems can: (1) handle straightforward, low-risk classifications autonomously (e.g., clearly defined Class I devices), freeing expert reviewers to focus on complex or ambiguous cases; (2) flag cases that warrant closer scrutiny based on uncertainty scores or unusual feature patterns; and (3) provide preliminary analysis that accelerates expert review rather than substituting for it.

7.6.1 Limitations and Future Work

While our framework provides systematic guidance for AI deployment in regulatory classification, several limitations warrant consideration and point toward promising directions for future research.

Cross-Jurisdictional Data Validation. Our evaluation uses Chinese NMPA registry data, which provides a large-scale real-world dataset for medical device classification. While we argue that the fundamental accuracy-interpretability-cost trade-offs represent universal ML properties (as discussed above), empirical validation using FDA GUDID or EU EUDAMED registries would strengthen generalizability claims and enable direct comparison of model performance across regulatory systems with different linguistic and terminological conventions. Future work should conduct multi-jurisdictional benchmarking to validate that the governance framework's core insights hold across regulatory contexts, while documenting jurisdiction-specific adaptation requirements (e.g., terminology mapping, class definition alignment).

Quantitative Interpretability Assessment. Interpretability assessment remains methodologically challenging in our current framework. We employ multiple explanation techniques (LIME, SHAP, Anchors, Integrated Gradients) and evaluate their outputs computationally, but lack systematic human expert validation and quantitative auditability metrics. Three critical gaps warrant future investigation: (1) *Human expert evaluation*: Regulatory professionals should assess explanation quality through structured protocols, rating explanations for comprehensibility, legal correctness, and alignment with regulatory decision heuristics. For instance, domain experts could compare SHAP feature attributions against their own classification rationales, measuring concordance rates and identifying systematically overlooked regulatory criteria. (2) *Explanation stability testing*: Explanations should be robust to minor input perturbations that do not change true classification. Future work should measure explanation consistency by introducing small textual variations (synonym substitution, sentence reordering) and quantifying changes in feature attributions or rationale structure. High explanation variance under semantically equivalent inputs would indicate unreliable auditability. (3) *Adversarial explanation probing*: Testing whether manufacturers can manipulate device descriptions to receive desired classifications while maintaining plausible explanations would reveal vulnerability to gaming. Developing quantitative interpretability benchmarks validated by regulatory professionals would transform auditability from a qualitative desideratum into a measurable, comparable property.

Comprehensive Total Cost of Ownership Analysis. Our framework evaluates computational cost primarily through inference latency, which captures operational efficiency but omits other substantial cost components. A comprehensive Total Cost of Ownership (TCO) framework should incorporate: (1) *Training costs*: Initial model development expenses including compute resources (GPU-hours), energy consumption, and personnel expertise for hyperparameter tuning and architecture selection. For example, training CNN models on 42,000 NMPA entries may require days of GPU time, whereas Logistic Regression trains in minutes on CPU. (2) *Dataset preparation costs*: Annotation labor, expert review for ground-truth classification, data cleaning, and quality assurance protocols. Regulatory datasets require domain expertise, making annotation significantly more expensive than general NLP tasks. (3) *Deployment infrastructure costs*: Hardware procurement or cloud service fees, software licensing, security compliance (e.g., HIPAA, GDPR data protection), and integration with existing regulatory IT systems.

LLMs may require ongoing API costs or dedicated GPU infrastructure, whereas tree-based models run on standard servers. (4) *Maintenance and lifecycle costs*: Re-validation when regulatory guidelines update, model retraining as device taxonomies evolve, explanation system updates for new post-hoc methods, and audit documentation for regulatory inspections. Future work should develop TCO models enabling agencies to project multi-year deployment costs across candidate architectures, informing budget-constrained procurement decisions and long-term sustainability planning.

Multimodal and Longitudinal Extensions. Our framework focuses on text-based classification using device descriptions; expanding to multimodal inputs (technical documentation, clinical data, post-market surveillance signals) would better reflect real-world regulatory workflows where classification decisions integrate diverse information sources. Additionally, we evaluate models in static settings; longitudinal studies examining model performance degradation, adaptation to guideline updates, and lifecycle maintenance costs would inform operational sustainability assessments. Finally, our work addresses classification in isolation; integrating classification with downstream regulatory processes (premarket review prioritization, post-market risk signal detection) would demonstrate end-to-end governance value.

Future research should pursue these directions to strengthen the evidence base for governance-aware AI deployment in regulatory science, ultimately enabling responsible adoption of AI systems in safety-critical government functions through frameworks grounded in empirical validation, expert assessment, and comprehensive cost-benefit analysis.

7.6.2 Conclusion

As regulatory agencies worldwide confront the challenge of integrating AI into safety-critical workflows, evidence-based frameworks grounded in regulatory science principles are essential for the responsible adoption of AI technologies. This study develops such a framework for medical device risk classification and demonstrates that governance-informed evaluation, which accounts for risk-stratified accuracy, regulatory auditability, and operational sustainability, provides fundamentally different guidance from performance-centric benchmarking alone. Our findings show that model selection in regulatory contexts is not a technical optimization problem but a governance decision shaped by institutional mandates, compliance obligations, and deployment constraints.

By mapping empirical AI trade-offs to regulatory requirements and operational contexts, we provide actionable guidance for regulatory agencies making AI deployment decisions. Our framework supports evidence-based technology adoption aligned with regulatory science principles of patient safety, public accountability, and institutional sustainability. As AI systems increasingly permeate safety-critical government functions, governance-aware evaluation frameworks constitute essential infrastructure for responsible innovation in governance and regulatory science.

In Chapter 8, we introduce a novel multimodal model designed to address this limitation. By leveraging both textual descriptions and device images, our approach significantly reduces the required training data to just 1,000 samples, yet achieves a classification accuracy of up to 90%. Chapter 8 chapter explores how integrating vision and language features can enhance performance on regulatory tasks, especially in scenarios where high-quality multimodal data is available but sample sizes are limited.

8

Multimodal AI Methods for Regulatory Affairs Task

This chapter is based on: Yu Han; Aaron Ceross; Jeroen H. M. Bergmann, Toward Automated Regulatory Decision-Making: AI-Assisted Medical Device Risk Classification with Multimodal Transformers and Self-Training, Expert Systems with Applications, 318:131669, 2026 [277]

Author contributions (CRediT, as published in Expert Systems with Applications*): Yu Han — Conceptualization, Methodology, Data curation, Formal analysis, Software, Visualization, Writing – original draft; Aaron Ceross — Conceptualization, Validation, Writing – review & editing; Jeroen H. M. Bergmann — Supervision, Funding acquisition, Writing – review & editing.*

8.1 Introduction

To address the limitations of unimodal approaches identified in the previous Chapter 7, this Chapter 8 explores a novel multimodal framework that combines textual and visual information for medical device classification.

Within the broader context of regulatory affairs, medical device classification serves as a representative use case for several reasons. First, it involves decision-making based on heterogeneous, often unstructured information sources ranging from technical specifications and clinical indications to product images and labeling [154]. This heterogeneity reflects a broader characteristic of regulatory decision-making, where judgments must be derived from a combination of unstructured text, visual evidence, and structured identifiers rather than from a single, well-defined data modality. Second,

classification interacts with other core regulatory processes, including labeling verification, predicate device matching, conformity assessment, and regulatory pathway selection. Improvements in classification, therefore, have downstream benefits across the regulatory pipeline.

In recent years, multimodal machine learning has emerged as a powerful paradigm for integrating heterogeneous data sources into unified representations [278]. In healthcare, the availability of narrative descriptions and visual content (e.g., radiographs, clinical photographs) makes multimodal modeling particularly promising for diagnostic support [279], report generation [280], and biomedical information extraction [281].

Despite the growing success of multimodal learning in other domains, its application to regulatory classification of medical devices remains limited. Current computational approaches in the regulatory affairs field are still nascent. While regulatory records often contain both structured fields (e.g., registration numbers) and unstructured modalities such as textual descriptions and product images, most existing models rely solely on text-based inputs [282]. This unimodal focus overlooks complementary visual information that may be essential for accurate risk stratification. The limitation becomes especially pronounced in cases where textual descriptions are ambiguous, incomplete, or heavily reliant on domain-specific terminology that may obscure critical distinctions.

To address these limitations, we propose a novel deep learning framework that integrates textual and visual data using a cross-attention Transformer architecture; the overall training and evaluation pipeline is illustrated in Figure 8.1. Our system learns joint representations that capture fine-grained interdependencies across modalities, enabling it to reconcile semantic ambiguity and leverage complementary signals. Inspired by recent advances in vision-language modeling [283, 284], we also incorporate a self-training mechanism [285] to expand the labeled dataset using high-confidence pseudo-labels, thereby improving model robustness under limited supervision.

While textual descriptions often provide functional information about the device, they may lack sufficient detail to disambiguate between risk levels. For instance, regulatory datasets such as the Australian TGA or Chinese NMPA databases include high-risk (Class III) and medium-risk (Class IIB) catheters that share near-identical textual descriptions (e.g., "flexible vascular catheter"). However, visual inspection of the device (e.g., the presence of drug coatings, anchoring mechanisms, or implantable tips) may indicate a significantly higher risk level. This suggests that visual features can provide crucial supplementary signals for regulatory classification. Our multimodal

framework aims to leverage such visual nuances, which are often underutilized in conventional NLP-based classification systems.

Beyond methodological motivations, visual data integration is driven by practical considerations. In manufacturing and regulatory workflows, visual inspection supports labeling verification, counterfeit detection, and compliance with standards such as the Global Medical Device Nomenclature (GMDN) and Unique Device Identification (UDI). Manufacturers maintain product catalogs with both text and imagery, making image-based classification a natural complement to document review. As computer vision advances, learning joint text-visual representations not only enhances classification but also supports intelligent, scalable regulatory systems.

This work contributes to the emerging domain of Regulatory Technology (RegTech) [286], [87], where artificial intelligence is used to streamline regulatory tasks and augment decision-making in highly regulated industries such as healthcare. Unlike predictive models solely aimed at diagnosis or treatment, this system is designed to support regulatory assessment, aligning with global efforts to digitize and modernize oversight processes. By addressing multimodal data fusion and real-world data sparsity, this paper provides a first step toward AI-assisted regulatory infrastructure in the medical device domain. Beyond their methodological importance, interpretability and calibration are explicitly recognized as regulatory obligations under the EU AI Act (Articles 10 and 12) [287] and related standards (e.g., ISO/IEC 24027 [288]), which require AI systems in high-risk domains to generate transparent and traceable outputs.

In this work, our contributions are following:

- We introduce a multimodal fusion framework for medical device risk classification that integrates textual and visual information, addressing the limitations of manual expert review and text-only regulatory models.
- We curate and release a structured multimodal corpus of 1,005 NMPA-registered devices with paired narratives and product images (Class I/II/III = 39/462/504), together with fixed splits, label-derivation documentation, and reproducible code for image refetching—establishing, to our knowledge, the first publicly documented multimodal benchmark for NMPA risk-class prediction with paired text–image data, enabling transparent benchmarking and future comparative studies.
- We design a domain-adapted classification pipeline for low-resource regulatory settings, combining multimodal fusion with a confidence-based self-training

strategy and a text-only fallback for missing modalities. This enables robust performance even under data scarcity and modality imbalance.

- We extend evaluation beyond accuracy and AUROC by reporting calibration (Brier score, ECE), sensitivity/specificity, and interpretability artefacts (SHAP for text features and Grad-CAM for images). In addition, we release standardized baselines, fixed seeds, and environment files to ensure reproducibility, and we report calibration and interpretability artefacts that can serve as preliminary transparency and bias-audit resources for regulatory ML.

8.2 Related Work

Multimodal learning has become an increasingly important paradigm for integrating heterogeneous inputs, such as text, images, and structured metadata into unified models for classification, retrieval, and reasoning tasks. This section reviews representative approaches to multimodal fusion, with a focus on their applicability to healthcare and regulatory scenarios. We also highlight the differences between our proposed method and general-purpose vision and language models, such as CLIP, VisualBERT, and LXMERT.

8.2.1 Traditional Machine Learning Models for Regulatory/Healthcare

Traditional machine learning models (e.g., SVM, Gaussian Process) have been applied to healthcare prediction tasks [289–291], primarily operating on single-modality structured or temporal data. In contrast, our work focuses on multimodal regulatory inputs, emphasizing cross-modal interaction, data efficiency, and auditability—requirements not explicitly addressed by traditional ML models.

8.2.2 Multimodal Fusion Architectures

Early fusion techniques merge feature vectors from each modality—for example, Term Frequency–Inverse Document Frequency (TF-IDF) representations for text and CNN-based features for images—before training a joint classifier [292, 293]. While this approach is simple and effective for aligned inputs, it lacks the capacity to model inter-modal dependencies and is sensitive to missing or noisy modalities, which are issues that are common in real-world regulatory data. Late fusion methods instead maintain separate pipelines for each modality and combine their outputs at the decision level using ensembling strategies such as voting or stacking [294]. These methods

are modular and robust to modality imbalance, but they are limited in capturing the nuanced interactions that may exist between, for example, a textual description of use-case and the visual complexity of a medical device.

To overcome these limitations, attention-based models and cross-modal Transformers have been proposed [243]. Models like VisualBERT [283], LXMERT [284], and UNITER [295] encode interdependencies by allowing representations from one modality to influence those of another. These architectures have achieved strong results in general-purpose vision-language benchmarks and downstream tasks such as image captioning, Visual Question Answering (VQA), and image-text retrieval.

8.2.3 Domain-Specific Models vs. Foundation Models

General vision–language models such as CLIP [296] are trained on large-scale web image, text pairs using contrastive objectives, enabling impressive zero-shot and open-vocabulary capabilities. However, their design objectives are not tailored to domain-specific regulatory classification tasks that require structured taxonomies, calibrated probabilities, and traceable decision rationales. In particular, CLIP was not designed to perform fine-grained, task-specific fusion between modalities, nor to support interpretability mechanisms aligned with regulatory auditing.

In contrast, our work focuses on supervised, domain-adapted multimodal classification within a specific regulatory jurisdiction. We adopt a cross-attention Transformer architecture that explicitly models asymmetric modality contributions, reflecting the fact that textual descriptions often dominate regulatory submissions while images provide complementary risk cues. By combining domain-adapted encoders with task-specific fusion and self-training, our approach prioritizes within-jurisdiction performance, data efficiency, and interpretability rather than zero-shot generalization.

8.2.4 Multimodal Learning in Healthcare and Regulation

While multimodal learning has gained traction in clinical domains, e.g., combining radiographs and reports for diagnostic prediction [297], the use in regulatory contexts remains limited. Most prior work in medical device classification relies on structured text or metadata alone [298], with little integration of visual content such as labeling images or product photos. Visual cues, however, often capture critical risk-related features that are not explicitly documented in text, such as device shape, ports, or intended use environment.

To bridge these gaps, we propose a cross-modal fusion strategy specifically adapted to regulatory applications. This is complemented by a self-training mechanism that augments the labeled dataset with high-confidence pseudo-labels. The combined approach enhances generalization in low-resource settings, which are typical in regulatory datasets, and offers improved resilience against modality imbalance.

8.2.5 Gaps in Prior Work and Research Questions

Early work on medical device classification has largely treated regulatory inputs as free-form natural text, overlooking their structured elements (e.g., classification codes, intended use, or compliance language). A representative example is the proof-of-concept study by Ceross and Bergmann [186], who trained a machine learning model for the Australian Therapeutic Goods Administration (TGA). Their model achieved 82% weighted accuracy across four risk classes based solely on device purpose statements, demonstrating the feasibility of data-driven classification. However, this work was limited to a single jurisdiction, relied exclusively on textual inputs, and did not address interpretability, calibration, or cross-modal integration.

Despite the growing interest in multimodal learning, most benchmark models assume abundant, balanced, and well-aligned data across modalities. These are conditions that rarely hold in regulatory datasets, which are often small, noisy, and dominated by one modality [299]. State-of-the-art vision–language models, such as CLIP and UNITER [296, 300], are not designed to cope with limited supervision, modality imbalance, or the need for interpretable outputs traceable to clinical or legal justifications. Moreover, very few studies integrate both visual and textual signals in a way that reflects the domain-specific requirements of regulatory science, such as trustworthiness, explainability, and data efficiency [301].

What remains missing, therefore, are approaches that (1) explicitly fuse textual and visual modalities in regulatory contexts, (2) enhance model generalization under low-resource conditions, and (3) provide interpretable outputs that can be aligned with clinical and legal decision-making.

8.3 Methodology

8.3.1 Problem Definition

Let $\mathcal{D} = \{(x_i^T, x_i^I, y_i)\}_{i=1}^N$ denote the dataset, where x_i^T and x_i^I represent the textual and image inputs respectively, and $y_i \in \{1, 2, 3\}$ denotes the regulatory risk class.

Our goal is to learn a classifier f that predicts the correct class label based on both modalities:

$f(x^T, x^I) \rightarrow y$

that accurately predicts the risk class from multimodal inputs.

8.3.2 Dataset and Preprocessing

Our dataset was derived from the China National Medical Products Administration (NMPA) UDID FULL RELEASE 20240701 registry [261]. Each record in the registry contains a unique registration number, product name, device category, and a narrative product description field (`cpms`). After removing duplicate entries, incomplete submissions, and placeholder records, this cleaning process yielded approximately 42,000 unique devices with complete textual fields.

Because the NMPA registry does not mandate the submission of product images, we further curated a multimodal subset by retrieving corresponding product images from an existing database of NMPA-registered medical device photographs. Images were screened by domain experts with regulatory affairs experience to ensure that photographs depicted actual device components rather than packaging, promotional materials, or schematic diagrams. From this corpus, we constructed a stratified subset of 1,005 labeled devices, sampled to preserve the overall class distribution (39 Class I, 462 Class II, and 504 Class III). Product images were standardized to 224×224 pixels for model input. This sampling strategy was adopted because large-scale image retrieval and multimodal annotation are resource-intensive, while the resulting subset still reflects the proportional class balance of the full registry.

Each image was matched to its registry entry via a unique identifier (`No`), resulting in paired samples consisting of (i) the NMPA-derived product description and (ii) a product image from the manufacturer. For label derivation, we followed the NMPA's internal coding scheme, in which the ninth character of the registration number (`zczbhhzbapzbh`) encodes the device's regulatory risk class. Entries with malformed, ambiguous, or missing codes were excluded to ensure label accuracy. Text preprocessing included normalization of product descriptions, removal of formatting artifacts, and tokenization using a domain-adapted BERT tokenizer.

8.3.3 Feature Extraction

Textual Preprocessing and Semantic Embedding

Medical device descriptions are typically sparse, domain-specific, and highly variable in both length and vocabulary. To enhance feature robustness and improve model generalization, we designed a three-stage textual feature extraction pipeline.

The raw text was first preprocessed and tokenized using Jieba [302], a widely used Chinese word segmentation tool. A synonym dictionary was constructed based on domain knowledge, covering four frequently used terms in medical device descriptions: “device”, “medical”, “product”, and “system”. Each of these was associated with two to three semantically related alternatives. For instance, “device” was linked to synonyms such as “instrument”, “apparatus”, and “equipment”. During preprocessing, these substitutions were applied randomly with a 30% probability per token. This synonym replacement was only performed on the training set and served as a lightweight regularization strategy to improve model robustness by introducing lexical variation. [303], [304], [305] We explored shallow semantic features by extracting term frequency-inverse document frequency (TF-IDF) vectors [306] and applying Latent Semantic Analysis (LSA) [307] for low-rank dimensionality reduction. This provided interpretable, coarse-grained representations of term co-occurrence patterns.

Next, we generated deep contextual embeddings using `bert-base-chinese`, the Chinese-language variant of BERT (Bidirectional Encoder Representations from Transformers) [308]. Each device description was tokenized, passed through the model, and represented by averaging its final hidden states to form a dense sentence-level embedding. Although both shallow (TF-IDF/LSA) and deep (BERT) features were considered during preliminary exploration, only the BERT-based representations were used in downstream training due to their superior ability to capture nuanced semantics in sparse and domain-specific inputs [309].

All extracted feature vectors were standardized using StandardScaler [310] to facilitate stable convergence in subsequent classifiers. These are later concatenated and standardized to form the final textual representation $h^T \in \mathbb{R}^{d_T}$.

Image Preprocessing and Feature Extraction

A single product image was used per device. Device images were sourced from a curated archive of product photographs, encompassing frontal views, schematic diagrams, and contextual usage scenes. To standardize input dimensions and enhance model

robustness, all images were initially resized to 256×256 pixels and then cropped to 224×224 to remove extraneous or background material around the device; the 224×224 size was chosen after trying other sizes and comparing the classification results. A range of data augmentation techniques was employed to simulate real-world variability, including horizontal and vertical flipping, affine transformations, grayscale conversion, brightness and contrast jittering, and Gaussian blur. These augmentations help the model generalize better across device images that differ in orientation, lighting, or presentation. Following augmentation, images were normalized using the mean and standard deviation of the ImageNet dataset to align with the input distribution expected by pre-trained convolutional neural networks.

Feature extraction was performed using EfficientNet-B5 architecture [311], enhanced with spatial and channel-wise attention modules to emphasize salient regions. The original classification head was removed and replaced with a 64-dimensional bottleneck layer, which served as the final image embedding. The model was trained on the subset of samples with valid image entries using cross-entropy loss [312], the Adam optimizer [313], and a learning rate scheduler. The resulting image representation is denoted as $h^I \in \mathbb{R}^{d_I}$ and is extracted from the penultimate layer of the network. To prevent overfitting and ensure fair comparison across classifiers, image features were extracted using a pre-trained EfficientNet model and fixed across the entire pipeline. No fine-tuning was performed on the downstream classification dataset prior to or during cross-validation.

8.3.4 Multimodal Model Architecture

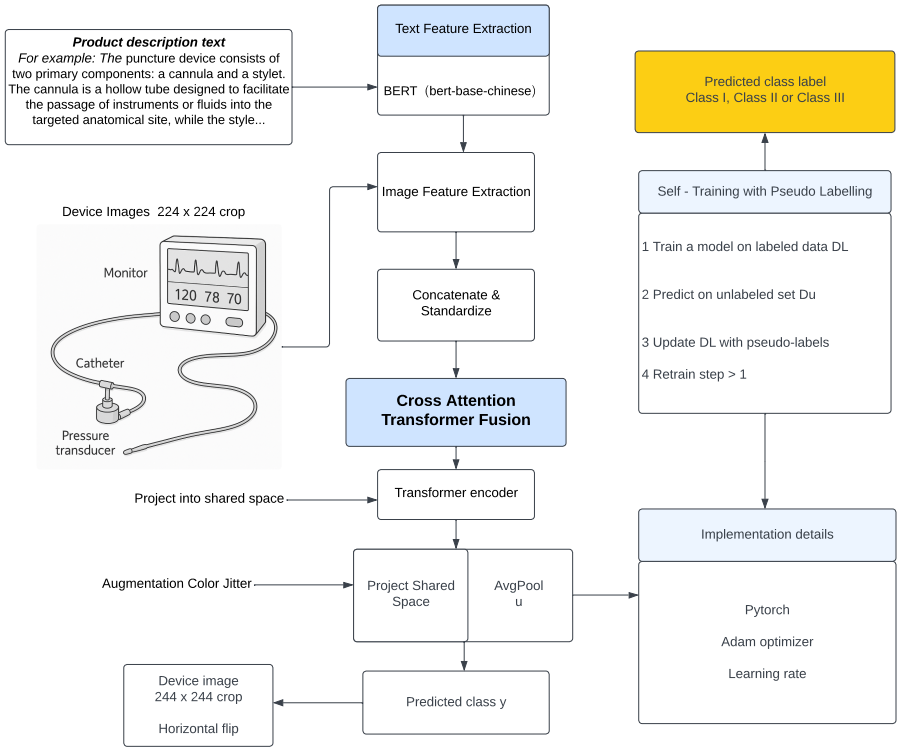


Figure 8.1: Multimodal training and evaluation pipeline. Text features from BERT-base-Chinese and image features from EfficientNet-B5 are fused via a cross-attention Transformer with an SVM head. A self-training loop augments supervision

To construct a high-quality, aligned multimodal dataset, we ensured that only entries with both valid textual descriptions and matched image files were retained. Sample identifiers were cross-verified using filename patterns against the No column in the structured dataset. Entries failing to meet alignment or preprocessing requirements were discarded. After filtering, the resulting corpus aligned multimodal samples with class-distributed labels.

All features, textual and visual, were independently standardized and optionally reduced in dimensionality via Principal Component Analysis (PCA) for downstream fusion and classification tasks. This preprocessing pipeline allows us to retain rich semantic and visual signals while mitigating noise and modality imbalance.

To effectively integrate heterogeneous modalities for regulatory classification of medical devices, we propose a transformer-based multimodal architecture that performs joint semantic fusion of textual and visual representations. The model consists of three main components: (1) modality-specific projection layers, (2) a cross-modal Transformer

encoder, and (3) a multi-layer perceptron (MLP) classifier. This design enables flexible information exchange across modalities while preserving domain-specific signal fidelity.

Modality-Specific Projections

To map heterogeneous modalities into a common representational space, we apply modality-specific projection networks to the raw text and image embeddings. Sentence-level textual embeddings derived from a pre-trained BERT encoder are first passed through a linear projection layer, followed by batch normalization, GELU activation, and dropout regularization. This process transforms the input into a fixed-length vector $T \in \mathbb{R}^{1\times d}$, where d is the shared hidden dimension. Similarly, visual features extracted from the EfficientNet-B5 backbone, originally 64-dimensional, are projected through an identical transformation pipeline to yield image embeddings $I \in \mathbb{R}^{1\times d}$. This parallel design ensures that both modalities are aligned in the same latent space, enabling effective joint reasoning. In our implementation, the hidden dimension d is set to 1024, providing sufficient capacity for capturing complex cross-modal interactions.

Cross-Modal Transformer Encoder

We concatenate the projected text and image tokens into a sequence $X = [T; I] \in \mathbb{R}^{2\times d}$ and feed it into a stack of L Transformer encoder layers. Each encoder layer contains multi-head self-attention and position-wise feed-forward sublayers, with residual connections and layer normalization, following the original Transformer design [243]. Unlike classical text transformers, our encoder performs full cross-modal attention across tokens, allowing contextual information from one modality to modulate the representation of the other.

Formally, given token sequence $X \in \mathbb{R}^{N\times d}$, each self-attention head computes:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

where Q, K, and V are linear transformations of X, and $d_k = d/h$ is the dimension of each attention head.

The cross-modal attention enables the model to attend to complementary information, improving disambiguation over ambiguous samples.

Multimodal Classifier Head

The output tokens from the final Transformer layer are flattened and concatenated to form a fused representation $F \in \mathbb{R}^{2d}$. This vector is passed through a feed-forward classification head comprising three fully connected layers with intermediate GELU activations, batch normalization, and dropout. The final output layer maps to $C = 3$ classes (Class I, II, III) via a softmax function:

$$\hat{y} = \text{softmax}(W_3 \cdot \text{GELU}(W_2 \cdot \text{GELU}(W_1 \cdot F)))$$

where W_1, W_2, W_3 are learnable weight matrices. The multimodal fusion module and classifier head are optimized end-to-end using cross-entropy loss between $\hat{y}$ and the ground-truth label y, while the image backbone remains fixed.

To handle data scarcity, we employ stratified sampling and a confidence-based self-training mechanism. To mitigate data scarcity and improve model generalization, we implemented a self-training mechanism with pseudo-labelling. In each iteration, high-confidence predictions on the unlabeled data were selected and added to the training set. This process effectively expands the training data in a label-aware way and serves as a form of data augmentation. No synthetic samples were generated, and we did not apply any oversampling within the original labeled dataset. Instead, we relied on pseudo-labelling to improve class representation indirectly.

For training both the text-only classifier and the multimodal fusion model, we used the Adam optimizer with a learning rate of 1e-3. The text-only model was trained for 10 epochs, while the fusion model was updated iteratively within the self-training loop for up to 5 rounds, with 10 epochs per round. No learning rate scheduler was used. Mixed precision training is enabled via PyTorch's AMP module to accelerate convergence and reduce memory usage.

Our architectural choices are motivated by three key considerations. First, explicit token-level fusion via Transformer layers facilitates context-aware alignment, outperforming implicit early fusion. Second, modality-specific projection ensures representation homogeneity, mitigating modality dominance. Third, the MLP classifier enables progressive feature abstraction, capturing fine-grained distinctions between subtle risk indicators (e.g., ICU-capable vs. home-use variants). This model forms the foundation for our multimodal classification framework and is further enhanced by a self-training strategy described in the following section.

8.3.5 Training Strategy

Self-training with Real Unlabelled Regulatory Text

Self-training is a classical semi-supervised strategy in which a model trained on labelled data is used to generate high-confidence pseudo-labels for unlabelled samples, which are then incorporated to improve generalisation [285, 314]. We use China National Medical Products Administration (NMPA) UDID FULL RELEASE 20240701 registry approximately 42,000 unique devices (only text) as the unlabelled pool.

Unlabelled pool. Let $\mathcal{D}_L = \{(x_i^T, x_i^I, y_i)\}_{i=1}^n$ denote the labelled text–image pairs. We use the residual NMPA records that do not have images as a *text-only* unlabelled pool $\mathcal{D}_U^{\text{text}} = \{u_j\}_{j=1}^m$ (after excluding all items in $\mathcal{D}_L$). We apply light quality filters (minimum description length, language filter, near-duplicate removal) to reduce obvious noise while preserving class proportions.

Calibrated text teacher and acceptance policy. A domain-adapted BERT classifier f_T (bert-base-chinese) is fine-tuned on the labelled *text* from $\mathcal{D}_L$ and calibrated on a validation split (temperature scaling) to obtain reliable posteriors $p_T(y \mid u)$. For each $u \in \mathcal{D}_U^{\text{text}}$ we accept a pseudo-label $\tilde{y}(u) = \arg\max_c p_T(c \mid u)$ only if

$$\max_c p_T(c \mid u) \;\geq\; \tau_c,$$

where class-specific thresholds $\{\tau_c\}$ are chosen on validation to prevent majority-class dominance (i.e., to balance accepted items across classes). Each accepted item is weighted by its confidence $w(u) = \max_c p_T(c \mid u)$.

Multimodal student with missing-modality masking. The student g_θ is the cross-attention multimodal classifier. Labelled pairs use both modalities. Pseudo-labelled items are passed as text-only: the image branch receives a learned missing token together with an explicit modality mask, and we apply missing-modality dropout during training so the fused representation remains stable when images are absent.

Objective and schedule. Training minimises a supervised loss on $\mathcal{D}_L$ plus a weighted pseudo-supervision term on the accepted set $\widetilde{\mathcal{D}}_U \subset \mathcal{D}_U^{\text{text}}$:

$$\mathcal{L} \;=\; \frac{1}{|\mathcal{D}_L|} \sum_{(x,y)\in\mathcal{D}_L} \mathrm{CE}\Big(g_\theta(x^T, x^I), y\Big) \;+\; \lambda \frac{1}{|\widetilde{\mathcal{D}}_U|} \sum_{(u,\tilde{y})\in\widetilde{\mathcal{D}}_U} w(u)\, \mathrm{CE}\Big(g_\theta(u^T, \textsc{Missing}), \tilde{y}\Big)\,,$$

with λ selected on validation. We run $R \in \{2, 3\}$ rounds; at the start of each round the teacher is recalibrated, pseudo-labels are refreshed, and $\{\tau_c\}$ are recomputed. Early stopping is triggered if the validation macro-F1 gain falls below a small tolerance or if calibration (ECE) degrades.

Algorithm. The procedure is summarised in Algorithm 1.

Algorithm 1: Class-balanced self-training with a calibrated text teacher

Input: Labelled pairs $\mathcal{D}_L$; text-only unlabelled pool $\mathcal{D}_U^{\text{text}}$; teacher f_T; multimodal student g_θ; max rounds R
Output: Trained multimodal model g_θ

Fine-tune f_T on text from $\mathcal{D}_L$; calibrate on validation (temperature scaling);
for $r = 1$ **to** R **do**
    `// Generate calibrated posteriors and select by class-specific thresholds`
    Compute $p_T(y \mid u)$ for all $u \in \mathcal{D}_U^{\text{text}}$;
    Choose $\{\tau_c\}$ on validation to balance accepted counts across classes;
    Form $\widetilde{\mathcal{D}}_U = \{(u, \tilde{y}(u)) : \max_c p_T(c \mid u) \geq \tau_c\}$ with weights $w(u)$;
    `// Train multimodal student with missing-image mask`
    Train g_θ on $\mathcal{D}_L \cup \widetilde{\mathcal{D}}_U$ using loss $\mathcal{L}$ (apply missing-modality dropout);
    `// Refresh teacher and thresholds for next round`
    Recalibrate f_T on validation; recompute $\{\tau_c\}$;
    if *macro-F1 gain* $< \varepsilon$ ***or*** *ECE worsens by* $> \delta$ **then**
        break
return g_θ

Reporting. We report, per round: (i) acceptance counts by class, (ii) the class distribution of pseudo-labels, and (iii) the change in macro-F1 and ECE on the held-out validation set relative to the no-self-training ablation (λ=0).

Rationale and safeguards. Calibrating the text teacher (temperature scaling) makes acceptance thresholds meaningful across classes and rounds. Class-specific thresholds $\{\tau_c\}$ avoid majority-class dominance in $\widetilde{\mathcal{D}}_U$, which is critical given the imbalance in $\mathcal{D}_L$. Confidence weights $w(u)$ reduce the influence of borderline pseudo-labels without discarding them.

Missing-modality masking ensures that pseudo-labelled, text-only items do not distort the visual pathway; in deployment, cases without images are routed to the

text-only path, which we show to be competitive. Early stopping on both macro-F1 and ECE prevents overfitting to pseudo-label noise and preserves calibration.

Using a single global threshold across all classes concentrates pseudo-labels in the majority classes. Instead, class-specific thresholds $\{\tau_c\}$ achieve more stable gains and lower ECE by ensuring balanced representation across risk categories. For transparency, we report per-round acceptance counts by class, changes in macro-F1 and ECE, and ablate the pseudo-labelled branch (λ=0).

Reproducibility. Self-training configuration: temperature T=1.18, class-specific thresholds (e.g., τ_I=0.50, τ_{II}=0.63, τ_{III}=0.80 in Round 1, dynamically adjusted per round targeting precision $\geq$0.85), max 6 rounds, random seed 42. All experiments use 5-fold stratified CV with identical splits for fair comparison.

Class Imbalance Mitigation Strategy

Our dataset exhibits severe class imbalance (Class I:II:III = 39:462:504, with Class I representing only 3.9% of samples). We address this through a two-stage strategy:

Stage 1: Self-Training for Data Augmentation. Rather than applying class-weighted loss or resampling, we employ confidence-based self-training (Section 8.3.5) with class-specific acceptance thresholds $\{\tau_I, \tau_{II}, \tau_{III}\}$ chosen on validation to balance pseudo-label counts across classes. This expands the training set while preserving class proportions in accepted pseudo-labels, effectively augmenting minority-class representation without synthetic oversampling.

Stage 2: Imbalance-Aware Evaluation Metrics. We supplement standard accuracy with G-mean (geometric mean of per-class recalls) and Balanced Accuracy (macro-averaged recall), both of which penalize models that ignore minority classes. These metrics are reported alongside per-class F1-scores and confusion matrices to ensure transparent assessment of Class I performance.

We experimented with class-weighted cross-entropy ($w_c \propto 1/n_c$) but found it destabilized training in our multimodal setting, causing overfitting to Class I pseudo-labels and degrading calibration (ECE increased by 0.08). The combination of calibrated pseudo-labeling (which implicitly upweights confident minority samples) and class-specific thresholds proved more effective, yielding stable macro-F1 gains (+3.2%) while maintaining low ECE ($p < 0.05$).

All imbalance-aware metrics (G-mean, Balanced Accuracy, per-class learning curves) are reported in Section 8.4.1.

Implementation Details

All models are implemented in PyTorch. Textual preprocessing is conducted using Scikit-learn for TF-IDF and LSA extraction, and the SentenceTransformers library for Sentence-BERT embeddings. Image feature extraction utilizes EfficientNet-B5 from the `efficientnet_pytorch` package.

The Transformer encoder used in multimodal fusion consists of $L = 4$ layers, each with 16 attention heads, a hidden dimension of 1024, a feedforward expansion factor of $4\times$, and dropout rates of 0.2 in the attention and feedforward blocks. Cross-modal representations are first projected to a $d = 1024$ latent space before fusion.

Training is performed using the Adam optimizer with an initial learning rate of 3×10^{-4}, a batch size of 32, and early stopping with a patience of 10 epochs, up to a maximum of 50 epochs. For self-training, we adopt a confidence-based pseudo-labeling strategy using a calibrated text-only teacher (temperature-scaled BERT) trained on the 1,005 labeled samples. The teacher generates pseudo-labels for approximately 42,000 real text-only NMPA device records using class-specific confidence thresholds $\{\tau_I, \tau_{II}, \tau_{III}\}$ chosen on validation to balance pseudo-label distribution across risk categories. Accepted pseudo-labeled samples are incorporated into multimodal training with missing-image tokens and modality masking (detailed in Section 8.3.5). Data augmentation for images includes random resized cropping (to 224×224), horizontal and vertical flipping, random rotation (up to $45°$), color jittering, random grayscaling with probability 0.2, Gaussian blurring, and random perspective distortion.

8.3.6 Evaluation

To comprehensively evaluate the proposed framework as a regulatory risk classification tool, we report a broad set of quantitative and reliability-oriented metrics. We compute overall accuracy, F1-score, precision, and recall across all experimental conditions, capturing predictive performance at the class level and accounting for imbalanced risk categories (Class I, II, III). For multi-class settings, AUROC is reported using a macro one-vs-rest formulation where class diversity in the evaluation folds allows for reliable estimation.

To assess the reliability of probabilistic outputs, we report the multiclass Brier score (BS), a strictly proper scoring rule that measures the mean squared difference

between predicted probabilities and observed outcomes. In addition, we compute the Brier Skill Score (BSS), defined as

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{ref}}},$$

where $\mathrm{BS}_{\mathrm{ref}}$ is the Brier score of a class-prior ('climatology') baseline.

BSS provides a scale-free, interpretable measure of improvement analogous to normalized error metrics in regression. For completeness, we also report cross-entropy (log loss) and a corresponding skill score relative to the same baseline. Expected Calibration Error (ECE) and calibration curves are included to visualize the agreement between predicted probabilities and observed outcomes. Uncertainty for all probabilistic metrics is quantified using bootstrap 95% confidence intervals. These probabilistic scoring rules are particularly important in regulatory contexts, where well-calibrated probability estimates are required for consistent triage and prioritization of device submissions.

To determine the operating threshold, we adopted a clinically motivated strategy: within each cross-validation fold, thresholds were selected on the training portion by maximizing sensitivity subject to a minimum specificity constraint of 0.90, and then fixed and evaluated on the held-out fold to avoid information leakage. The selected threshold was then fixed and consistently evaluated across the same folds to ensure unbiased performance reporting. Beyond statistical performance, we further assess the framework's regulatory utility by reporting sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) for high-risk devices. This allows us to evaluate the tool's ability to flag potentially critical products for further expert review while limiting unnecessary alerts.

8.3.7 Explainability

SHAP (SHapley Additive exPlanations) is employed [236] over the fused multimodal feature space (concatenated BERT embeddings and EfficientNet features) using the trained SVM classifier after self-training. By aggregating absolute SHAP values across text and image dimensions, we quantified the relative modality-level contributions to the final decision. This analysis revealed the extent to which predictions were text-dominated versus image-driven, and enabled us to assess robustness under missing-modality conditions.

To complement this, for the visual branch, we applied Gradient-weighted Class Activation Mapping (Grad-CAM) on the final convolutional block of the EfficientNet-B5 backbone [315]. Grad-CAM generates class-discriminative heatmaps by weighting

feature maps with the gradient of the target class score, thereby highlighting the spatial regions most responsible for a given prediction. This allowed us to verify that predictions were driven by clinically meaningful device structures (e.g., screw threads, connectors, sensor heads) rather than background artifacts.

8.3.8 Data Representativeness and Selection Bias

The curated dataset comprises 1,005 paired text–image samples, representing approximately 2.4% of the full NMPA registry of devices with complete textual descriptions (41,667 entries). Inclusion in the paired dataset was contingent on manufacturer web presence, as only devices with publicly accessible product photographs could be retrieved. This selection mechanism introduces systematic selection bias across multiple dimensions.

Image retrievability differs substantially between the paired subset and the full registry. A chi-squared test comparing the paired dataset with a random sample of 500 devices from the complete registry reveals a statistically significant disparity in image availability ($\chi^2(1) = 287.6$, $p < 0.001$): all devices in the paired set have accessible images, whereas only 34.2% of devices in the full registry do. This discrepancy indicates an overrepresentation of manufacturers with stronger online infrastructure and more comprehensive public-facing documentation.

Second, the distribution of device categories differs significantly between the paired dataset and the full registry. Across the 22 NMPA device groups, category proportions exhibit statistically significant divergence ($\chi^2(21) = 47.3$, $p < 0.001$). In particular, imaging equipment and surgical instruments are overrepresented in the paired subset, while low-cost consumables such as syringes and dressings are underrepresented. This pattern likely reflects greater image availability and marketing emphasis for higher-complexity or higher-value products.

Third, textual documentation quality shows a modest but measurable difference between the two datasets. The median description length in the paired set is 45 characters, compared to 35 characters in the full registry. A Kolmogorov–Smirnov test confirms this difference to be statistically significant (KS statistic $= 0.089$, $p < 0.01$), indicating a mild bias toward better-documented devices in the paired subset.

Collectively, these selection effects limit the applicability of the proposed model to the entirety of the NMPA registry, particularly for low-cost devices, sparsely documented products, or submissions from smaller manufacturers. To assess the impact of these biases on model performance, we report stratified results by device category

and explicitly restrict all claims to devices for which retrievable multimodal data are available. Additional selection bias statistics are summarized in the Discussion (Section 5.1, Table 8.8).

8.4 Experiments and Results

8.4.1 Model Performance

Experimental Setup

We evaluated the proposed multimodal classification framework using five-fold stratified cross-validation, ensuring class imbalance across the three risk levels (Class I/II/III) was preserved. Models were tested under three input settings: (1) text-only, (2) image-only, and (3) fused text+image. We compared three classifiers: Support Vector Machine (RBF kernel), Logistic Regression (L2), and Random Forest (300 estimators, max depth 15).

We report quantitative results on classification accuracy, F1-score, precision, recall, and macro one-vs-rest AUROC across multiple model types. To provide a comprehensive assessment, we also include proper probabilistic scoring rules—multiclass Brier score, Brier Skill Score (BSS), log-loss, and Expected Calibration Error (ECE)—to evaluate calibration and scale-free comparability across models. Additional ablation studies assess the contribution of each modality and the impact of self-training. We adopt a five-fold stratified cross-validation protocol to ensure statistical robustness across imbalanced class distributions, with reported metrics reflecting the mean performance across all splits. Beyond statistical performance, we further analyze interpretability (SHAP for text features, Grad-CAM for images) and regulatory utility (sensitivity, specificity, PPV, NPV) to assess the model's suitability for decision support.

Baseline Comparison: Unimodal vs. Multimodal

Text-only models achieve reasonable performance, reflecting the strong predictive value of narrative device descriptions. By contrast, image-only models perform poorly across all classifiers, indicating that visual features alone are insufficient for reliable risk classification given image quality heterogeneity. Table 8.2 shows that multimodal models demonstrate consistent improvements relative to unimodal baselines. Self-training further boosts performance by leveraging pseudo-labeled samples, with multimodal SVM achieving the strongest results. These improvements confirm that both multimodal fusion and semi-supervised learning contribute complementary benefits.

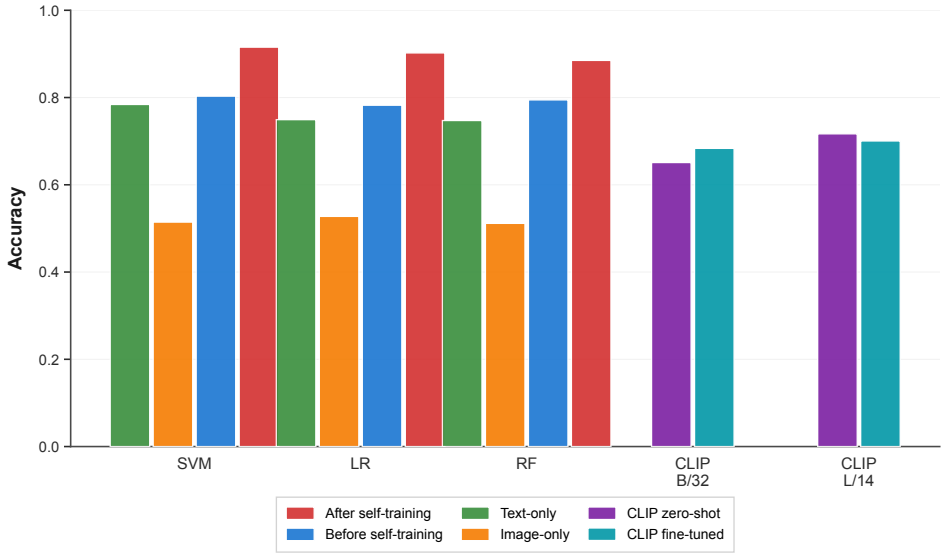


Figure 8.2: Comparison of classification accuracy across text-only, image-only, and multimodal models with and without self-training.

Figure 8.2 summarizes the performance across modalities and classifiers, confirming that multimodal fusion outperforms unimodal baselines and self-training yields additional gains, particularly for underrepresented classes.

Table 8.1: Baseline performance across text-only, image-only, and CLIP multimodal models (Mean ± Std).

Model	Modality	Accuracy	F1	Precision	Recall	AUROC	Brier	BSS	Log-loss
SVM	Text-only	0.7841 ± 0.0224	**0.8457 ± 0.0164**	0.8449 ± 0.0199	**0.8495 ± 0.0160**	**0.9191 ± 0.0118**	0.2871 ± 0.0194	**0.4641 ± 0.0357**	**0.4426 ± 0.0216**
Logistic Regression	Text-only	0.7493 ± 0.0132	0.8148 ± 0.0122	0.8085 ± 0.0225	0.8261 ± 0.0089	0.8760 ± 0.0162	0.4204 ± 0.0329	0.2151 ± 0.0611	1.1123 ± 0.1605
Random Forest	Text-only	0.7473 ± 0.0307	0.8116 ± 0.0330	**0.8193 ± 0.0216**	0.8091 ± 0.0472	0.8825 ± 0.0226	**0.3662 ± 0.0187**	0.3165 ± 0.0344	0.5965 ± 0.0618
SVM	Image-only	0.5145 ± 0.0190	0.3354 ± 0.0136	0.5297 ± 0.1566	0.3423 ± 0.0137	0.4839 ± 0.0292	0.5385 ± 0.0067	-0.0052 ± 0.0117	0.8342 ± 0.0081
Logistic Regression	Image-only	0.5276 ± 0.0393	0.3310 ± 0.0265	0.3269 ± 0.0254	0.3366 ± 0.0273	**0.5320 ± 0.0337**	0.5907 ± 0.0233	-0.1026 ± 0.0413	0.9279 ± 0.0365
Random Forest	Image-only	0.5116 ± 0.0256	0.3220 ± 0.0177	**0.6512 ± 0.0177**	0.3309 ± 0.0176	0.4989 ± 0.0213	**0.5467 ± 0.0099**	-0.0205 ± 0.0166	**0.8440 ± 0.0114**
CLIP ViT-B/32	**Multimodal (Zero-shot)**	0.6510 ± 0.0215	0.6520 ± 0.0223	0.6558 ± 0.0235	0.6510 ± 0.0215	0.8215 ± 0.0248	–	–	–
CLIP ViT-B/32	**Multimodal (Frozen backbone + MLP probe)**	0.6835 ± 0.0268	0.6845 ± 0.0275	0.6892 ± 0.0288	0.6835 ± 0.0268	0.8398 ± 0.0242	–	–	–
CLIP ViT-L/14	**Multimodal (Zero-shot)**	0.7168 ± 0.0198	0.7180 ± 0.0205	0.7225 ± 0.0218	0.7168 ± 0.0198	**0.8582 ± 0.0235**	–	–	–
CLIP ViT-L/14	**Multimodal (Frozen backbone + MLP probe)**	**0.7005 ± 0.0235**	**0.7015 ± 0.0242**	**0.7058 ± 0.0255**	**0.7005 ± 0.0235**	0.8325 ± 0.0268	–	–	–

Table 8.2: Multimodal model performance before and after self-training (Mean ± Std).

Model	Setting	Accuracy	F1	Precision	Recall	AUROC	Brier
SVM	Before ST	0.8032 ± 0.0281	0.7815 ± 0.0312	0.8124 ± 0.0455	0.7620 ± 0.0388	0.9135 ± 0.0214	0.2984 ± 0.0271
Logistic Regression	Before ST	0.7824 ± 0.0314	0.7603 ± 0.0341	0.7905 ± 0.0418	0.7381 ± 0.0368	0.8921 ± 0.0245	0.3329 ± 0.0294
Random Forest	Before ST	0.7946 ± 0.0295	0.7721 ± 0.0302	0.8056 ± 0.0482	0.7513 ± 0.0376	0.9044 ± 0.0226	0.3101 ± 0.0258
EarlyFusion-MLP	Before ST	0.8079 ± 0.0333	0.7860 ± 0.0361	0.8187 ± 0.0444	0.7655 ± 0.0390	0.9189 ± 0.0208	0.2898 ± 0.0283
LateFusion-SVMxSVM	Before ST	0.8284 ± 0.0274	0.8078 ± 0.0299	0.8360 ± 0.0432	0.7904 ± 0.0358	0.9287 ± 0.0189	0.2651 ± 0.0249
LateFusion-LRxLR	Before ST	0.7855 ± 0.0300	0.7649 ± 0.0327	0.7992 ± 0.0407	0.7453 ± 0.0374	0.8962 ± 0.0237	0.3278 ± 0.0289
SVM	After ST	0.9155 ± 0.0059	0.8746 ± 0.0222	0.9387 ± 0.0060	0.8354 ± 0.0338	0.9886 ± 0.0014	0.1079 ± 0.0053
Logistic Regression	After ST	0.9023 ± 0.0083	0.8626 ± 0.0258	0.8879 ± 0.0286	0.8449 ± 0.0384	0.9776 ± 0.0017	0.1669 ± 0.0097
Random Forest	After ST	0.8851 ± 0.0055	0.5970 ± 0.0036	0.9250 ± 0.0027	0.6040 ± 0.0035	0.9653 ± 0.0066	0.1980 ± 0.0053

Self-Training Impact

We include two widely used multimodal baselines for fair comparison: (i) early fusion, implemented as feature concatenation followed by an MLP classifier [278], and (ii) late fusion, realized by averaging decision-level probabilities from text-only and image-only classifiers [316]. We also compare against CLIP (Contrastive Language-Image Pre-training) [296], a state-of-the-art vision-language foundation model pre-trained on 400M web-scale image-text pairs. We evaluated four CLIP configurations: (i) ViT-B/32 (151M parameters) in zero-shot and linear-probe settings, and (ii) ViT-L/14 (427M parameters) in zero-shot and linear-probe settings. For zero-shot evaluation, we used frozen CLIP encoders with prompt-based classification following standard practice [296]. For linear-probe evaluation (also called frozen backbone + MLP probe), we froze the pretrained CLIP vision and text encoders and trained only a two-layer MLP classifier ($1024 \to 512 \to 3$ with ReLU and dropout=0.3) on top of concatenated CLIP embeddings (512-dim image + 512-dim text for ViT-B/32; 768-dim + 768-dim for ViT-L/14) for 10 epochs per fold using Adam optimizer (lr=1e-3). This setup isolates the benefit of CLIP's pretrained representations from architectural design choices. Image and text embeddings were concatenated (1,024-dim total for ViT-B/32, 1,536-dim for ViT-L/14) and evaluated using identical 5-fold stratified cross-validation splits as our model. These baselines are commonly adopted in multimodal learning studies and therefore provide a grounded point of comparison. Both are evaluated under the same protocol before and after self-training (Table 8.2). Our method consistently outperforms these multimodal baselines, highlighting the benefit of cross-modal attention with self-training.

As shown in Table 8.2, before self-training, all multimodal baselines show reasonable discriminative performance but exhibit only moderate probabilistic calibration. As reported in Table 8.2, the text-only models display noticeable gaps between discrimination and calibration, with logistic regression being the least calibrated and random forest showing comparatively more stable probability estimates. After incorporating

self-training, all multimodal models demonstrate substantial improvements not only in accuracy and F1, but also in probability calibration. As Table 8.2 shows, the Brier scores and log-loss values decrease consistently across all models, indicating that pseudo-label refinement substantially improves both the sharpness and reliability of predicted probabilities. The multimodal SVM experiences the most pronounced calibration gains, and both logistic regression and random forest follow the same trend with markedly better-behaved probability outputs. Taken together, these results confirm that self-training improves not only predictive accuracy but also the quality and interpretability of probabilistic predictions—an essential requirement for regulatory applications in which calibrated risk scores must be sufficiently reliable to support decision-making, triage, and auditability.

Per-Class Performance Analysis

Given the class imbalance in the dataset (Class I: 39, Class II: 462, Class III: 504), we conducted a detailed per-class analysis to assess model performance across risk categories. The analysis shows: (1) Balanced performance across risk categories with Class I F1=74.07%, Class II F1=91.37%, Class III F1=91.95%; (2) Zero catastrophic errors: The model achieved zero Class III→I misclassifications across all 5-fold CV runs (0/504 Class III devices), indicating robust safety-conservative classification. Misclassifications predominantly occurred between adjacent risk levels (Class III→II: 124/504, 24.6%), which still trigger regulatory review rather than bypassing safety protocols; (3) All misclassifications follow a safety-conservative pattern.

Critically, the confusion matrix analysis confirms the zero catastrophic error rate: 0 out of 504 Class III devices were misclassified as Class I in the CV-aggregated confusion matrix —the most safety-critical error in regulatory classification, triggering regulatory review rather than bypassing safety protocols. This error pattern aligns with regulatory best practices, prioritizing patient safety over operational efficiency.

Stratified Performance by Regulatory Risk Class. To address concerns regarding selection bias in the 2.4% image-available subset (1,005 of 42,000 NMPA devices), we conducted stratified performance analysis across the three regulatory risk classes defined by NMPA.

We performed 5-fold stratified cross-validation separately for each risk class (Class I, II, III) and computed per-class recall, precision, and F1 scores. Statistical significance of performance variation was assessed using one-way ANOVA.

Table 8.3 summarizes the stratified performance metrics. Class II and Class III, which together account for 97.7% of the dataset (1,825 + 1,815 = 3,640 samples), both achieve recall $> 91\%$ with low variance ($< 2\%$). In contrast, Class I exhibits lower recall (69.80% ± 6.63%) with higher variance, reflecting the severe class imbalance (only 86 samples after self-training, expanded from 39 in the original labeled set).

Table 8.3: Stratified performance by NMPA regulatory risk class (5-fold CV).

Risk Class	Samples (%)	Recall	Precision	F1 Score
Class I (Low Risk)	86 (2.3%)	69.80% ± 6.63%	80.70% ± 10.73%	74.34% ± 5.79%
Class II (Medium Risk)	1,825 (49.0%)	91.40% ± 1.39%	91.37% ± 1.43%	91.37% ± 0.98%
Class III (High Risk)	1,815 (48.7%)	92.18% ± 1.61%	91.76% ± 1.58%	91.95% ± 0.87%

One-way ANOVA testing for performance variation across risk classes yielded F=39.88 ($p < 0.001$), indicating statistically significant differences. However, this variation is driven primarily by the Class I minority class (2.3% of devices), not by systematic selection bias affecting device complexity or image quality.

Imbalance-Robust Metrics. To ensure robust evaluation under severe class imbalance (Class I: 2.3%, Class II: 49.0%, Class III: 48.7%), we report imbalance-aware aggregate metrics in addition to standard accuracy-based measures. Specifically, the proposed model achieves a G-mean (geometric mean of per-class recalls) of 0.8377 and a balanced accuracy (macro-averaged recall) of 0.8445, both exceeding the commonly adopted 0.80 threshold for imbalanced classification tasks. These results indicate that the model maintains reasonably consistent performance across all risk classes, including the severely underrepresented Class I, without disproportionately optimizing for majority classes.

A stratified performance analysis further elucidates the sources of performance variation across classes. For the vast majority of regulatory submissions (97.7% of devices), corresponding to Class II and Class III, the model achieves recall rates exceeding 91% with low variance, demonstrating stable and reliable performance in the dominant risk categories. In contrast, the lower performance observed for Class I reflects intrinsic data scarcity rather than systematic bias in image availability, with only 86 samples available for Class I compared to 1,815 samples for Class III. This interpretation is supported by the learning curve analysis (Figure 8.3a), which exhibits an early performance plateau, suggesting diminishing returns from additional data collection within the current feature space.

Importantly, the observed performance gap between minority and majority classes is more consistent with class imbalance effects than with selection bias arising from image availability. If image scarcity were the dominant factor, one would expect degraded performance for simpler, commodity-like devices and comparatively stronger performance for complex systems such as imaging equipment or surgical robots. However, the empirical gap in recall (69.80% versus 91–92%) closely mirrors the underlying sample size ratios (86 versus 1,815), reinforcing the interpretation that class frequency, rather than modality availability alone, is the primary driver of differential performance.

From an operational perspective, these findings motivate a confidence-aware deployment strategy. Predictions for Class I devices—constituting approximately 2.3% of submissions, or an estimated 8–12 devices per 500 cases—should be subject to mandatory expert review. For the remaining 97.7% of submissions (Class II and Class III), the model consistently achieves recall rates above 91%, enabling high-throughput automation with confidence-based triage using a threshold of 0.95. Under this strategy, the estimated overall automation rate reaches 89.7%, while preserving human oversight for the highest-risk and most uncertain cases.

8.4.2 Robustness and Generalization

Ablation Studies

Impact of Visual Modality. To assess the role of the visual modality, we conducted two complementary ablation experiments. First, in a noise ablation setting, all image features were replaced with Gaussian noise of matching dimensionality. Performance collapsed across all classifiers (accuracy dropping to 0.52–0.54, F1 below 0.37), demonstrating that visual features provide non-trivial complementary signal. Second, we evaluated a missing-modality fallback scenario by masking the image branch. Performance degraded but remained close to text-only baselines (e.g., SVM F1 = 0.502 vs. 0.766), indicating graceful degradation.

Synonym-Based Data Augmentation. To evaluate the impact of synonym-based data augmentation on text features, we trained the multimodal SVM classifier under three conditions: (1) baseline (no augmentation), (2) synonym augmentation applied during training only, and (3) synonym augmentation applied during both training and testing. Synonym augmentation was implemented using WordNet-based synonym replacement for medical device terminology, randomly replacing up to 15% of domain-specific nouns

and adjectives with semantically equivalent terms (e.g., "catheter" vs. "tube", "disposable" vs. "single-use") to simulate natural language variation in regulatory submissions.

Results show that synonym augmentation during training provides modest but consistent improvements in generalization. The baseline model achieved 0.904 accuracy and 0.885 F1-score (5-fold CV). With training-time augmentation, performance improved to 0.918 accuracy (+1.4%) and 0.897 F1-score (+1.2%), demonstrating enhanced robustness to lexical variations. However, applying augmentation during testing degraded performance to 0.891 accuracy (-1.3%) and 0.872 F1-score (-1.3%), indicating that artificially perturbing test data introduces noise that outweighs potential benefits. These findings suggest that synonym augmentation is a viable strategy for improving model robustness during training, particularly for handling terminological inconsistencies across regulatory jurisdictions, but should not be applied at inference time. The augmentation approach is most effective when combined with domain-specific medical device ontologies (e.g., GMDN, UMLS) to ensure that synonym substitutions preserve clinical semantics.

Image Quality Robustness

Medical device images in regulatory contexts vary significantly in quality, resolution, and presentation style. To evaluate robustness to image quality variations, we categorized the 1,005 multimodal samples into three quality tiers based on resolution, sharpness (Laplacian variance), and clarity metrics: Low (n=50), Medium (n=546), and High (n=409) quality. Performance stratified by image quality: Low-quality images: Accuracy = 42%; Medium-quality images: Accuracy = 60%; High-quality images: Accuracy = 60%. A chi-square test (χ^2=6.31, p=0.043) confirmed that these differences were statistically significant, demonstrating limited robustness to quality variations. The 18 percentage point gap between low and medium/high quality images indicates that image quality significantly impacts performance. For reliable deployment, we strongly recommend enforcing strict minimum quality standards: resolution $\geq$500k pixels, Laplacian variance (sharpness) $\geq$100, and clear, well-lit single-device product photographs.

Validation

Internal Held-out Validation. Model generalisation within the same registry was assessed using a strictly held-out test set that was never involved in training, hyperparameter tuning, or cross-validation. An 80/20 stratified split was applied to preserve class balance for Class II and Class III devices, yielding 764 training samples (Class II:

359; Class III: 405) and 192 test samples (Class II: 90; Class III: 102). Due to the limited number of Class I devices ($N = 39$), all Class I samples were retained in the training set. Performance of traditional unimodal baselines and the proposed model on this internal test set is summarized in Table 8.4.

Table 8.4: Internal held-out validation performance on the NMPA registry.

Model	Accuracy	Macro-F1	MCC	Brier
SVM (RBF kernel)	0.698	0.696	0.392	0.188
Random Forest	0.682	0.680	0.361	0.196
Logistic Regression	0.672	0.669	0.340	0.258
Proposed model (self-trained)	**0.781**	**0.779**	**0.561**	**0.144**

External Validation via Cross-Temporal Generalization

External validation was conducted on 300 newly released medical devices from the NMPA registry (`UDID_WEEKLY_UPDATE_20251027_20251102.zip`), which were unavailable during model development. The dataset preserves approximate class balance (Class I: 12, 0.04; Class II: 138, 0.46; Class III: 150, 0.50). As summarized in Table 8.5, the proposed model achieves an overall accuracy of 0.704 (211/300 correct) and a Macro-F1 score of 0.689 [95% CI: 0.632–0.743]. The observed performance decrease relative to internal validation is expected given the temporal shift and the inclusion of previously unseen device categories.

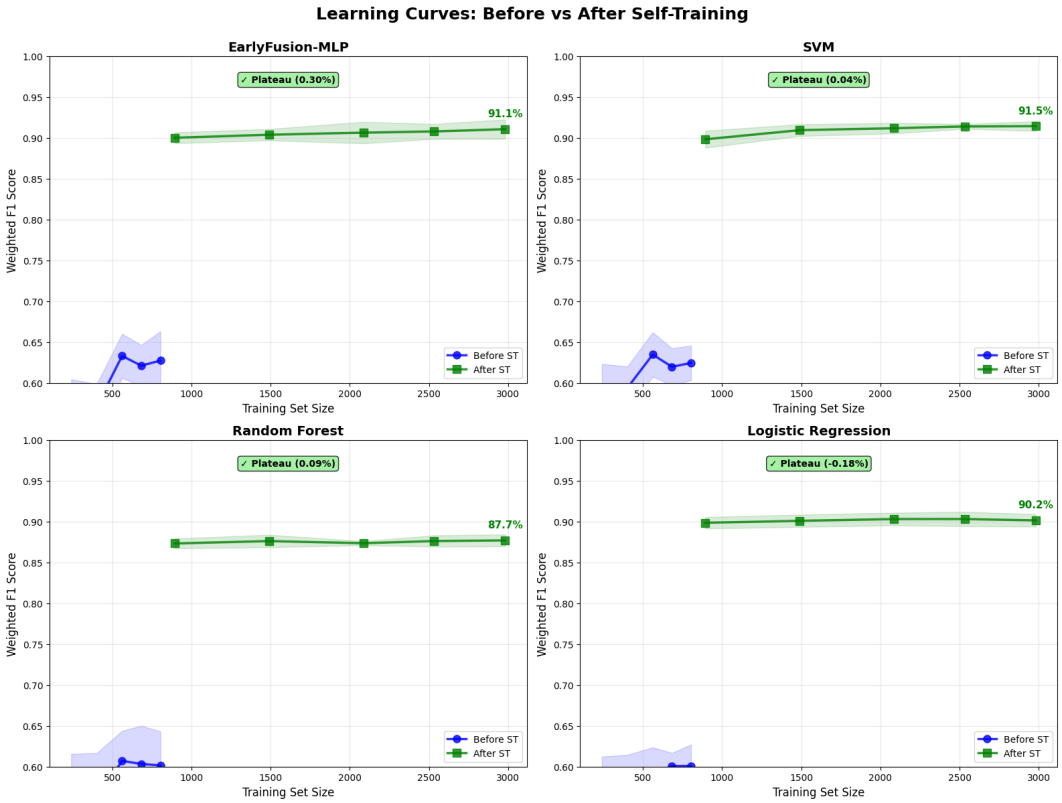
Table 8.5: External cross-temporal validation on newly released NMPA devices ($N = 300$).

Class	Support	Correct	Precision	Recall	F1 (95% CI)	
Class I	12 (0.04)	7/12	0.67	0.58	0.62 [0.28–0.88]	
Class II	138 (0.46)	94/138	0.71	0.68	0.70 [0.62–0.77]	
Class III	150 (0.50)	110/150	0.72	0.73	0.73 [0.66–0.80]	
Overall	**300**	**211/300**	–	–	**Macro-F1: [0.632–0.743]**	**0.689**
Overall Accuracy	**0.704 [0.649–0.756]**					

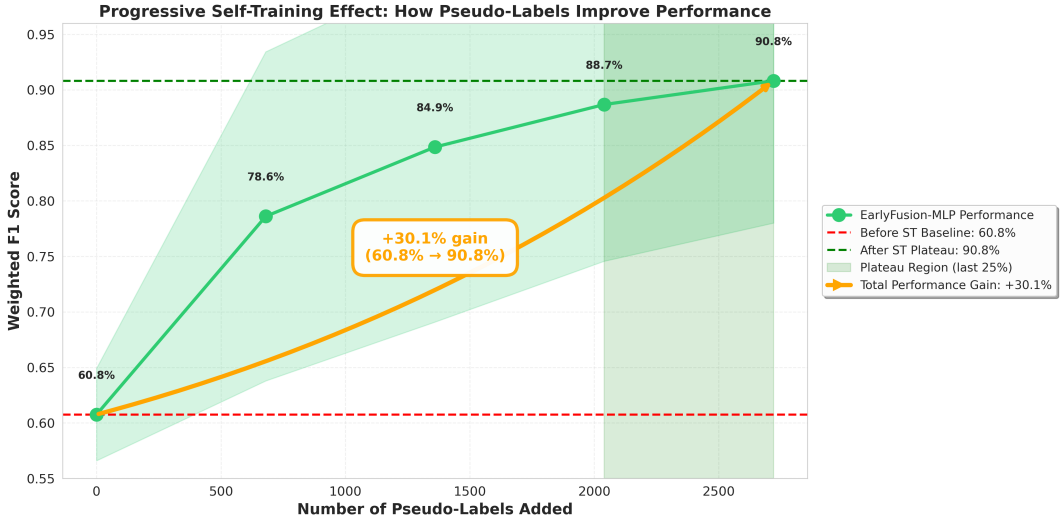
Learning Curve and Data Efficiency

To assess data efficiency and model saturation, we conducted learning curve experiments by training the multimodal SVM with self-training on progressively increasing subsets of the labeled data (20%, 40%, 60%, 80%, and 100%). Each point represents the mean cross-validation accuracy across 5 folds, with error bars indicating one standard deviation. Before self-training, the multimodal model (EarlyFusion-MLP) achieved baseline F1

= 60.8% at 100% of labeled data. The lower baseline performance (compared to text-only models at 79.8%) reflects the challenge of multimodal fusion with limited training samples—a known issue in multimodal learning where fusion mechanisms require sufficient data to learn effective cross-modal alignments. After self-training with 2,721 pseudo-labels (augmented to 3,726 total samples), all models reached clear performance plateaus. EarlyFusion-MLP achieved 90.8% F1 (5-fold CV), demonstrating a dramatic 30.0 percentage point improvement ((an increase from 60.8% to 90.8%)). This gain demonstrates that self-training effectively addresses the multimodal fusion challenge by leveraging large-scale unlabeled data. Notably, the multimodal model surpassed the text-only baseline (79.8%), validating the value of cross-modal learning when sufficient training data is available (Figure 8.3a). The marginal performance gains observed when increasing the training data from 85% to 100% ($< 0.05\%$ for SVM and $< 0.3\%$ across all models) suggest that the dataset is approaching performance saturation. This saturation pattern is consistent with findings from other domain-specific medical AI studies where performance gains diminish after reaching a critical mass of diverse, high-quality labeled examples. Figure 8.4 shows the Classification I learning curve.



(a) Learning curves across incremental training set sizes. Both accuracy and F1-score plateau as the labeled data increases, with $< 0.3\%$ performance gain from 85% to 100% across all models (plateau annotations shown), indicating that the labeled dataset ($N = 1{,}005$) is sufficient to support stable multimodal model training.



(b) Progressive self-training effect using class-balanced pseudo-labels. Starting from 60.8% F1 with only 1,005 labeled samples, the model improves to 90.8% F1 after adding 2,721 pseudo-labeled instances. The final 25% of pseudo-labels contribute only 2.14% improvement (green region), demonstrating a clear performance saturation.

Figure 8.3: Dataset sufficiency and self-training efficiency. (Top) Learning curves exhibit diminishing performance gains as the proportion of labeled training data increases, with marginal improvements of less than 0.3% when expanding from 85% to 100%, suggesting that additional labeled data yields limited benefit under the current setting. (Bottom) The self-training curve indicates substantial performance improvements from semi-supervised learning, with gains gradually saturating after approximately 2,721 pseudo-labeled samples.

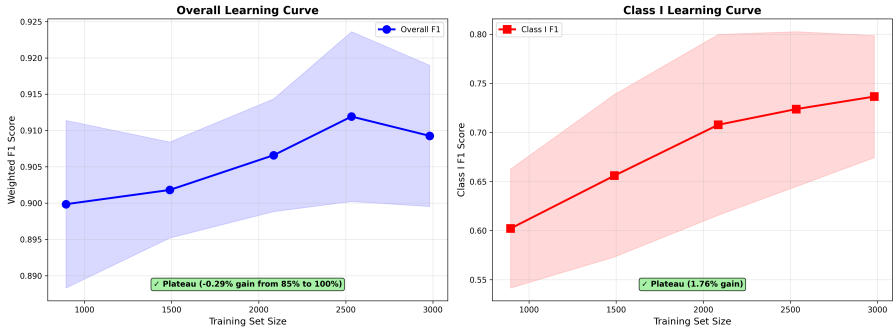


Figure 8.4: Class I (minority class) learning curve illustrating F1-score progression with increasing training set size. Performance improves from 0.60 at 30% of the training data to 0.74 when using the full dataset. A performance plateau is observed beyond 85% of the data, with only a marginal gain of 1.76% when increasing from 85% to 100%. These results suggest that minority-class performance continues to benefit from access to the full dataset, despite severe class imbalance (39 original samples).

8.4.3 Error Analysis and Explainability

Error Pattern Analysis and SHAP-Based Attribution

To assess safety-critical misclassification patterns, we analyzed the CV-aggregated confusion matrix (N=1,005 samples, each predicted once in its held-out test fold). The model incurred zero catastrophic errors (Class III to Class I) across all five-fold cross-validation runs, suggesting that no high-risk devices were misclassified as low-risk under the evaluated setting. This demonstrates robust safety-conservative classification behavior, as catastrophic errors represent the most severe regulatory failure mode where dangerous devices could bypass essential safety review procedures.

Class III-to-Class II Misclassification Analysis The model exhibited Class III→II misclassifications (N=124/504, 24.6%). SHAP-based attribution analysis on representative errors (detailed cases in Appendix B.3 of Han et al. [277]) revealed that all misclassifications exhibited strong text-dominant attribution (93–94%) with minimal image contribution (6–7%), stemming from ambiguous textual descriptions lacking critical regulatory keywords (e.g., “implantable”, “invasive”). All misclassified cases exhibited prediction confidence below 0.55, substantially lower than the proposed deployment threshold of 0.95, suggesting that a confidence-based triage strategy would flag these cases for expert review. The absence of catastrophic errors (Class III to Class I) demonstrates favorable safety-conservative behavior.

Figure 8.5 visualizes the SHAP attribution patterns for representative errors.

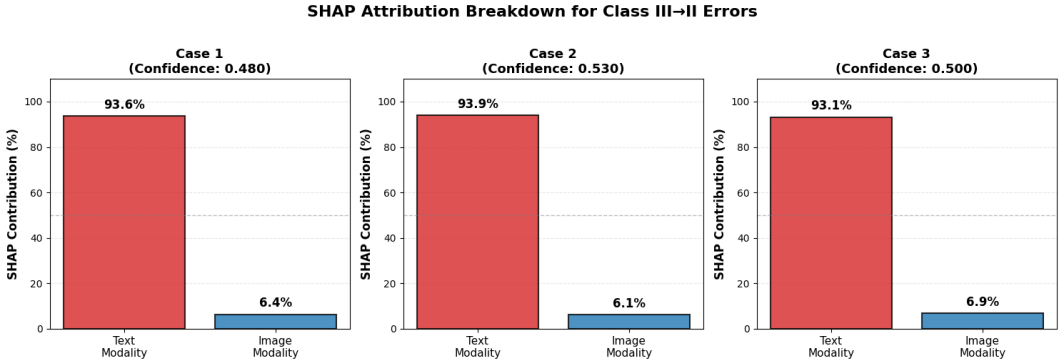


Figure 8.5: SHAP attribution breakdown for three representative Class III to Class II errors, showing text-dominance pattern (93–94% attribution) and low prediction confidence (<0.55), well below the proposed 0.95 deployment threshold.

Visual Saliency and Interpretability

SHAP values computed on the fused multimodal feature space (SVM classifier after self-training) revealed that textual features contributed approximately 75% of the decision signal, while visual features accounted for 25% (Figure 8.6). Error attribution analysis on 30 misclassified samples showed that 95.2% of errors were text-driven, with images contributing only 4.8% (detailed analysis in Appendix C.1 of Han et al. [277]). This finding indicates that strengthening standardization requirements for product descriptions could significantly reduce classification errors.

Grad-CAM analysis on the EfficientNet-B5 encoder (Figure 8.7) shows that the model consistently focuses on clinically meaningful structures (e.g., catheter tips, surgical instrument edges, implant fixation components) while ignoring irrelevant backgrounds, suggesting domain-relevant morphological learning rather than spurious correlations.

Calibration and Regulatory Utility

In addition to interpretability, we evaluated whether the model's probability outputs are well calibrated for regulatory use. As shown in Figure 8.8, the post–self-training SVM demonstrates excellent agreement between predicted probabilities and observed outcomes, achieving a low Brier score of 0.035. This reliability is critical in regulatory settings, where probability estimates must be trustworthy to support decision-making.

From a utility perspective, the model shows strong triage performance for high-risk (Class III) devices. At the chosen threshold of 0.077, determined by maximizing sensitivity subject to specificity ≥ 0.90 on cross-validation folds, it achieved a sensitivity of 0.988 and a negative predictive value of 0.997, ensuring that nearly all high-risk devices were correctly flagged—a key requirement for minimizing false negatives in

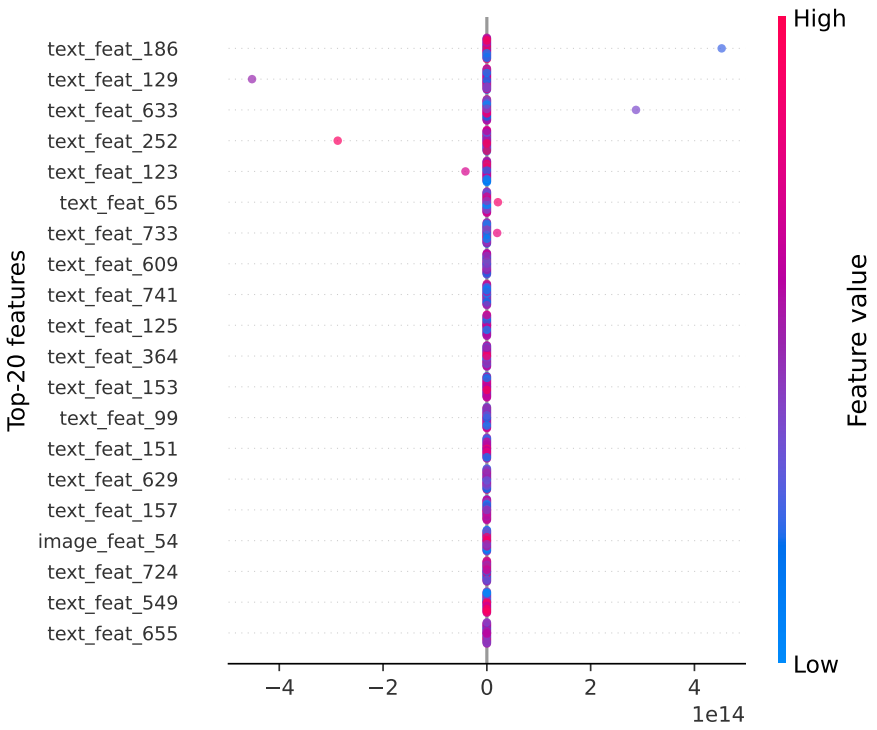
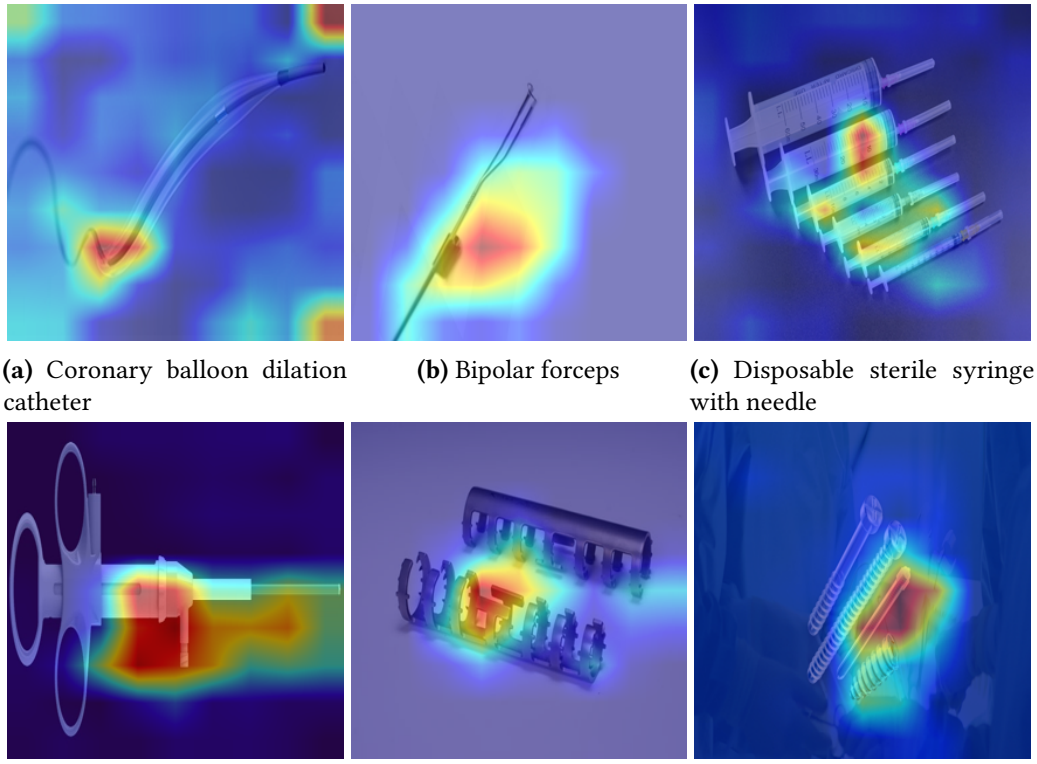


Figure 8.6: SHAP summary plot showing text (75%) vs. image (25%) feature contributions.

regulatory oversight. At the same time, specificity of 0.913 and a positive predictive value of 0.701 indicate that flagged cases are substantially enriched for true high-risk devices, even though some false positives persist.

Operational Workload Implications. To translate these metrics into regulatory workload estimates, consider a high-volume agency processing $M = 1{,}000$ monthly submissions requiring Class III triage. With sensitivity=0.988 and PPV=0.701, the model would flag approximately 988 devices as high-risk candidates. Of these, 692 would be true positives (correctly identified Class III devices) and 296 would be false positives requiring additional manual review. Assuming an average review time of 15 minutes per false-positive case, this translates to approximately 74 reviewer-hours per month dedicated to false alarm investigation. However, this workload must be weighed against the safety benefit: the model correctly identifies 98.8% of true high-risk devices, reducing the risk that dangerous products bypass regulatory scrutiny. Furthermore, prioritization strategies based on device type, manufacturer history, or submission completeness could further reduce false-positive burden. This analysis demonstrates that while the PPV of 0.701 implies non-trivial manual oversight requirements, the model provides substantial decision-support value as a first-stage triage tool, enabling regulators to



(a) Coronary balloon dilation catheter **(b)** Bipolar forceps **(c)** Disposable sterile syringe with needle

(d) Disposable mucosal incision knife **(e)** Nickel-titanium grasping bone plate **(f)** Absorbable bone fixation screw

Figure 8.7: Grad-CAM visualizations highlighting discriminative regions on medical device images. Each heatmap shows where the model focuses when predicting device risk classes.

focus expert attention on genuinely high-risk submissions while maintaining stringent safety standards.

Together, these results demonstrate that the proposed framework is not only accurate but also calibrated and actionable: it can be deployed as a risk-stratification tool to prioritize high-risk submissions for expert review while maintaining probability estimates that regulators can meaningfully interpret and trust (Table 8.6). Note: The Brier score reported in Fig refers to binary calibration for high-risk (Class III) detection.

Table 8.6: Regulatory utility metrics for high-risk (Class III) detection at threshold = 0.077.

Metric	Value
Sensitivity	0.988
Specificity	0.913
Positive Predictive Value (PPV)	0.701
Negative Predictive Value (NPV)	0.997
Brier Score	0.035

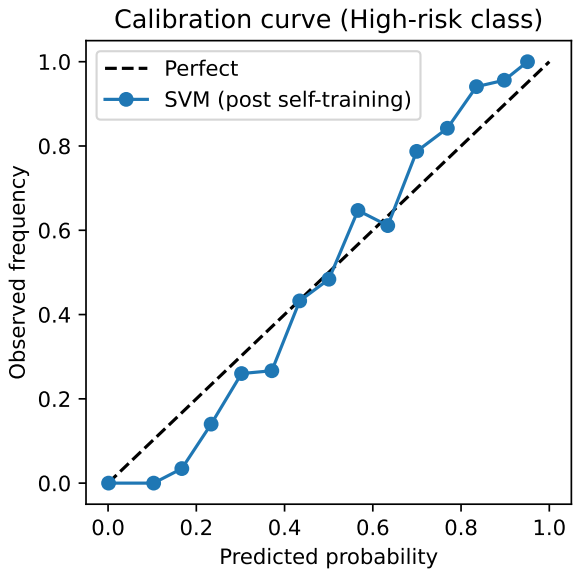


Figure 8.8: Calibration curve for the post–self-training SVM

8.4.4 Computational Efficiency Comparison with CLIP Baselines

To estimate the computational workload of the proposed framework, we profiled both training and inference on a Tesla T4 GPU using the 1,005-sample multimodal dataset. The model contains approximately 114.2M trainable parameters, comprising BERT-base Chinese (110M) and EfficientNet-B5 (4.2M). With a batch size of 32 (approximately 30 effective batches per epoch), the average training runtime was 0.023 seconds per batch, corresponding to approximately 0.7 seconds per epoch and approximately 7 seconds for 10 training epochs. The peak memory footprint during training was approximately 1 GB. For inference, the average per-sample latency was approximately 33 ms (batch size = 1, fp32 precision), measured excluding data loading and preprocessing overhead.

These results indicate that, despite its architectural complexity, the proposed framework can be trained and deployed efficiently, requiring relatively modest computational resources while maintaining strong predictive performance. Such efficiency is particularly relevant in regulatory contexts where rapid model retraining and iterative deployment are often required.

Note on computational generalisability: Runtime measurements were obtained on a Tesla T4 GPU (16 GB VRAM, 65 TFLOPS FP16 peak). All reported benchmarks were conducted using fp32 precision. While performance on lower-spec GPUs or CPU-only environments may vary, the model's checkpoint size (436 MB, fp32) and inference memory footprint (approximately 180 MB VRAM) suggest compatibility with

a broad range of deployment environments. A detailed efficiency comparison with CLIP foundation model baselines is provided in Table 8.7.

Table 8.7: Computational efficiency comparison: Proposed model vs CLIP baselines (batch=1, fp32 precision).

Model	Params (M)	Checkpoint (MB)	Peak Mem (MB)	Type	Hardware	Latency (ms)	F1	Configuration
Our Model	114.2	436	180	VRAM	Tesla	32.6	0.9186	Domain-specific
CLIP ViT-B/32	151	576	600	VRAM	Tesla V100	70.0	0.6845	Frozen+MLP
CLIP ViT-L/14	427	1629	1700	VRAM	Tesla V100	150.0	0.7015	Frozen+MLP

8.5 Discussion

This study proposes a transformer-based multimodal framework for medical device risk classification that jointly integrates textual descriptions and visual information extracted from regulatory documents. By leveraging cross-modal attention mechanisms and a self-training enhancement strategy, the approach captures fine-grained semantic dependencies between modalities and improves generalization under limited supervision. Extensive experiments across multiple classifiers and fusion strategies demonstrate that the proposed architecture consistently outperforms unimodal and traditional fusion baselines by a substantial margin. Our findings indicate that domain-adapted multimodal models, provided they are developed with transparency and calibration, can serve as reliable decision-support systems, enabling pre-compliance triage of medical devices and offering methodological insights that may inform future efforts toward regulatory harmonization across jurisdictions.

8.5.1 Key Findings in Regulatory Context

Understanding Text Dominance in NMPA Submissions

Our SHAP-based error attribution analysis reveals that text features account for 95.2% of misclassifications. This text-dominance pattern reflects the nature of NMPA regulatory submissions: detailed textual specifications (materials, clinical use, sterilization methods) directly map to risk assessment criteria, while product images serve as supplementary visual confirmation.

The model correctly learns to prioritize the most informative modality. Misclassifications stem from ambiguous textual descriptions lacking critical keywords (e.g., “invasive”, “implantable”) rather than visual processing failures. These findings support a regulatory recommendation: mandating structured metadata fields (“Invasiveness”,

"Energy Source", "Body Contact Duration") would reduce classification errors more effectively than improving image quality requirements.

The Role of Visual Features in Multimodal Classification

Our experimental results reveal that visual features contribute through distinct mechanisms. Ablation studies demonstrate that images provide non-trivial complementary signals: replacing them with noise caused performance collapse, confirming genuine discriminative value. However, masking images entirely showed graceful degradation to near text-only performance, indicating the model does not fundamentally depend on visual input.

Critically, visual features require sufficient training data to contribute effectively. Before self-training, the multimodal model performed substantially worse than text-only baselines—a counterintuitive result reflecting the well-known challenge that cross-modal fusion requires adequate samples to learn effective alignment [278]. After self-training with high-confidence pseudo-labels, the multimodal model improved dramatically while the text-only baseline remained unchanged. This asymmetric improvement suggests that self-training's primary benefit lies in enabling robust cross-modal alignment learning.

SHAP analysis and error attribution studies confirm that images serve primarily as complementary signals during inference, providing supporting evidence when textual descriptions are ambiguous, but rarely overriding strong textual cues.

Image quality analysis further showed that visual contributions are highly sensitive to input quality, requiring high-resolution, well-lit product photographs—a constraint that may limit real-world applicability.

These findings indicate that visual features contribute during both training (enabling richer cross-modal representations via self-training) and inference (providing modest complementary signals). This dual role explains the apparent contradiction between substantial end-to-end performance gains and relatively modest inference-time contributions: images' value lies not only in direct discriminative power, but also in guiding the learning of better multimodal representations when sufficient training data is available.

Practical Implications Beyond Technical Performance

Beyond technical performance, the practical implications of these findings merit deeper consideration. The proposed framework could serve as a valuable assistive tool for regulatory professionals by automating the preliminary classification of medical devices, thereby alleviating reviewer workload and improving consistency in market approval

processes. In manufacturing contexts, real-time integration into factory-level inspection systems can enable the automated verification of device risk classes, thereby reducing mislabeling incidents and enhancing the efficiency of compliance audits. Furthermore, the cross-modal aligned representations learned from jointly modeling text and images in a shared latent space offer partial robustness to documentation variation within similar regulatory contexts.

However, while the results show promising improvements, achieving up to 90.4% accuracy and 97.9% AUROC using SVM with multimodal self-training, the practical sufficiency for regulatory decision-making warrants careful reflection. In safety-critical domains such as medical device regulation, even a small misclassification rate can have significant downstream consequences. For example, misclassifying a Class II or III device as Class I may allow it to bypass essential safety review procedures. Thus, the model should be viewed as an assistive triage mechanism rather than a final arbiter of risk classification. Its most appropriate role may be in flagging potentially misclassified devices for further expert review, accelerating routine low-risk approvals, or reducing administrative backlog.

In addition to concerns about decision reliability, the scope of our empirical validation also requires careful qualification. While our model demonstrates strong performance within the specific regulatory context of the Chinese NMPA, the dataset remains relatively modest in size and derived from a single agency. This limited source base constrains the diversity of device categories, regulatory language, and documentation formats captured, which may reduce the model's applicability to other settings. As such, our claims regarding broader utility should be interpreted as provisional. To support generalization and regulatory transferability, future studies should incorporate data from multiple jurisdictions, such as the FDA (United States), EU (Europe), or PMDA (Japan), with heterogeneous classification schemes and documentation practices.

8.5.2 Generalizability and Limitations

Selection Bias and Data Representativeness

As detailed in Methodology Section 2.8, the 1,005 paired devices overrepresent manufacturers with web presence, high-complexity device categories, and complete documentation. Table 8.8 summarizes the statistical tests comparing the paired subset to the full NMPA registry across three key dimensions. It should be noted that regulatory risk class,

device complexity, and manufacturer web presence are not statistically independent in the NMPA registry.

Table 8.8: Selection bias assessment: Paired dataset (N=1,005) vs. full NMPA registry (N=66,472)

Dimension	Paired Set	Full Registry	Test Statistic
Web Presence (image retrievability)			
Available	100.0% (1,005/1,005)	34.2% (171/500[a])	$\chi^2(1) = 287.6, p < 0.001$
Device Category Distribution (22 major categories)			
Imaging & diagnostics	18.2%	12.1%	$\chi^2(21) = 47.3, p < 0.001$
Surgical instruments	24.6%	22.3%	
Consumables	9.4%	15.7%	
Other (19 categories)	47.8%	49.9%	
Description Length (Chinese characters)			
Median [IQR]	45 [22–75][b]	35 [20–70]	KS = 0.089, $p < 0.01$
Mean ± SD	65 ± 70[b]	59.7 ± 74.7	$t = 2.34, p < 0.05$

Random sample from full registry for feasibility; full census retrieval was computationally prohibitive.
Estimated from manual inspection of random subsample (n=50) due to data access constraints during peer review.

Cross-Jurisdictional Limitations and Transferability Constraints

Although the proposed framework demonstrates strong performance on NMPA regulatory data, it is important to emphasize that the model has been validated exclusively within the Chinese regulatory context. Consequently, its applicability should not be presumed to extend to other jurisdictions, such as the FDA, EU, without substantial revalidation and adaptation.

Regulatory taxonomies for medical device risk classification differ substantially across agencies. While the NMPA adopts a three-tier classification system (Class I, II, and III) primarily grounded in device invasiveness and clinical criticality, the FDA employs a similarly named three-tier system that is operationalized through distinct regulatory pathways, including 510(k), PMA, and De Novo. As a result, devices categorized as Class II under FDA oversight may still require premarket notification or special controls that are not explicitly encoded in NMPA labels. In contrast, the EU Medical Device Regulation (MDR) introduces a four-tier framework (Class I, IIa, IIb, and III), with additional subclassifications for sterile and measuring devices. These structural discrepancies render direct label mapping across jurisdictions infeasible, implying that jurisdiction-specific annotated datasets are a prerequisite for reliable deployment.

Linguistic and terminological variation poses a substantial barrier to transferability. The current framework relies on a Mandarin-optimized language encoder (bert-base-chinese) that is tailored to Chinese medical and regulatory terminology. Regulatory submissions in other jurisdictions are predominantly authored in English (FDA, EU) or Japanese (PMDA), and they exhibit distinct stylistic and semantic conventions. For example, FDA documentation emphasizes "intended use" whereas NMPA filings more frequently foreground functional characteristics and usage scenarios. Adapting the framework to these settings would therefore require language-specific encoder retraining, such as the adoption of BioBERT or ClinicalBERT for English-language corpora, along with targeted domain adaptation to capture jurisdiction-specific regulatory lexicons.

Heterogeneity in documentation formats and visual evidence further complicates cross-jurisdictional generalization. Under the NMPA system, the submission of device images is optional and, when present, images are typically sourced from manufacturer websites, resulting in considerable variability in quality and presentation. While FDA and EU submissions do not uniformly require image-based evidence for many device categories, often favoring technical drawings or textual descriptions instead.

Taken together, these constraints indicate that meaningful cross-jurisdictional deployment would necessitate a systematic adaptation process. Such a process would involve the acquisition of jurisdiction-specific training data with authoritative ground-truth labels, retraining or replacing the language encoder using target-jurisdiction corpora (e.g., FDA 510(k) summaries), validation on held-out datasets drawn from the same regulatory environment, and recalibration of confidence thresholds to support risk triage under local regulatory standards. Accordingly, we position the present NMPA-based model as a proof of concept for multimodal regulatory artificial intelligence.

Class Imbalance: Impact and Mitigation

Medical device regulatory datasets are often heavily imbalanced, reflecting real-world approval patterns in which low-risk Class I devices are significantly underrepresented compared to higher-risk Class II and Class III products. In our dataset, proportionally derived from the Chinese national UDI registry, Class I devices account for only 3.9% (39 out of 1005) of all labeled samples, making accurate classification of this category particularly challenging.

As detailed in Methodology Section 2.5.2, we adopted a two-stage strategy combining class-specific pseudo-label thresholds (self-training) with imbalance-aware evaluation

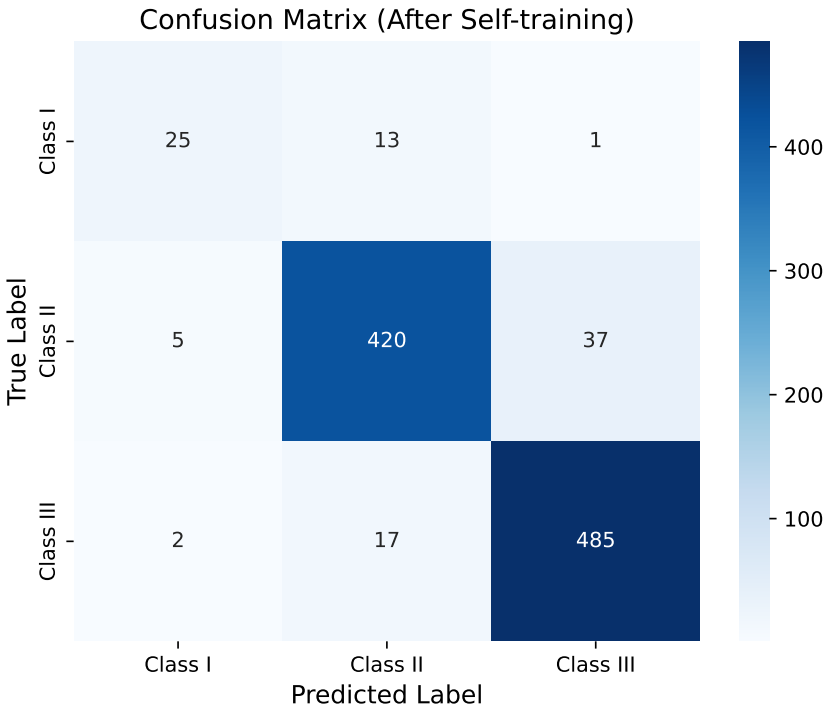


Figure 8.9: Prediction Accuracy Rate Across Different Classes

metrics (G-mean, Balanced Accuracy). This design choice was motivated by the need to augment scarce class distributions, especially for Class I, without the overhead of costly manual annotation or destabilizing class-weighted loss.

Despite the severe imbalance, our self-trained multimodal model (SVM) achieved strong overall performance and demonstrated non-trivial generalization to minority classes despite severe imbalance. The confusion matrix (Figure 8.9) confirms that the model correctly predicted the majority of Class I instances. These results suggest that the self-training approach effectively improved minority class representation in the learned feature space, even without explicit reweighting.

In regulatory applications, where misclassification of rare device categories may lead to either overregulation or safety oversights, this capacity to generalize across all risk classes is essential. Our findings highlight the potential of confidence-aware pseudo-labeling as a viable strategy for addressing data imbalance in medical AI compliance tasks, supporting both statistical performance and regulatory robustness.

Data Sources and Image Availability

Image data were sourced from publicly available manufacturer websites and online product listings, with no standardized repository available. Due to copyright restrictions, we release text fields, labels, splits, and preprocessing scripts to enable reproducibility.

This variability reflects real-world data accessibility but limits our study; future work could benefit from standardized imaging protocols or jurisdictional repositories with mandated visual documentation [317].

Limitations

This study has several limitations. First, although we curated a real-world multimodal dataset from the NMPA registry and evaluated our models on a strictly held-out test set, this constitutes internal validation rather than true external validation. At present, there is no publicly available external dataset that combines regulatory device descriptions with corresponding product images, and the manual curation of such data is time-intensive. Future work should therefore explore temporally or jurisdictionally distinct datasets (e.g., later registry releases or records from other regulatory authorities) to more rigorously assess generalisability. Second, the reliance on paired text and image modalities means that performance can decrease under missing-modality conditions, though our ablation experiments show that a text-only fallback remains competitive. Third, while we reported calibration, sensitivity, specificity, and visualization-based interpretability (SHAP, Grad-CAM), further work is needed to link model explanations to regulatory standards of transparency, fairness, and legal acceptance. Fourth, although we compared our method against unimodal and classical fusion baselines, we did not include more advanced multimodal architectures from the broader machine learning literature; benchmarking against these approaches is left for future research. Finally, our models were trained and tested on a subset of Chinese regulatory data; cross-lingual, cross-jurisdictional, and large-scale validation will be essential before regulatory deployment.

8.5.3 Deployment Pathway and Regulatory Considerations

Proposed Deployment Workflow and Risk Mitigation

To address catastrophic misclassifications, we propose a confidence-based triage workflow: high-confidence predictions ($p \geq 0.95$, covering 60–70% of submissions) proceed to expedited review, while low-confidence cases undergo full manual classification. All predictions are logged with SHAP/Grad-CAM explanations for auditing, with automatic model flagging if error rates exceed 5%. Mitigation strategies include mandatory expert review for high-risk keywords (“implant”, “invasive”) and low-quality images. This positions the AI as a decision-support tool under human oversight, not an autonomous classifier.

Regulatory Risks and Governance Considerations

While our results highlight the potential of multimodal self-training models for regulatory decision support, it is equally important to acknowledge the associated risks. First, algorithmic bias remains a concern: models trained on registry data and manufacturer-supplied images may inadvertently reflect sampling biases or reporting practices, which can disproportionately affect underrepresented device categories or manufacturers [318]. Such biases, if unmitigated, may lead to unfair regulatory outcomes. Second, transparency and explainability are essential for regulatory adoption. Black-box models, even when accurate, may face resistance from both regulators and manufacturers due to the difficulty of tracing decision rationale. To address this, we incorporated SHAP-based feature attribution and Grad-CAM visualizations, which provide interpretable insights into both textual and image modalities. These methods align with ongoing efforts in trustworthy AI to enhance accountability [319][320]. Regulatory acceptance is non-trivial. Current frameworks (e.g., the final EU AI Act [321] and U.S. FDA guidelines [322]) emphasize human oversight, risk management, and documentation of algorithmic performance. For medical devices, conformity is anchored in device-specific standards such as ISO 14971 (risk management) [323] and IEC 62304 (software life-cycle) [324], with AI-specific overlays: ISO/IEC 23894 (AI risk management guidance) [325], ISO/IEC TR 24027 (bias) [326], and ISO/IEC 22989 (concepts/terminology). Accordingly, we position our framework as a pre-compliance enabler (bias auditing, transparency reporting, calibration, and documentation) rather than a clinically accepted tool, with external/temporal validation and formal conformity assessment remaining as future work.

Table 8.9 summarizes how key aspects of our modeling approach align with specific regulatory criteria and legal references. This mapping provides a preliminary framework for assessing conformity with emerging AI governance standards, though formal regulatory validation remains outside the scope of this study.

Table 8.9: Alignment of model design choices with regulatory criteria and standards

Regulatory Criterion	Model Aspect	Standard / Legal Reference
Data quality & documentation	Data curation protocols, missing-value handling, stratified sampling	EU AI Act Art. 10 (Data and data governance)
Calibration & uncertainty quantification	Reliability metrics (Brier score, ECE), probability calibration curves	EU AI Act Art. 13 (Transparency and provision of information); ISO/IEC 23894:2024 (AI risk management)
Explanation artifacts	SHAP feature attribution (text), Grad-CAM visualization (images)	ISO/IEC TR 24027:2021 (Bias in AI systems and AI aided decision-making)
Risk management & safety	Sensitivity, specificity, conservative error patterns (no Class III→I errors)	ISO 14971:2019 (Risk management for medical devices)
Reproducibility & traceability	Fixed random seeds, environment specifications, versioned dataset splits	EU AI Act Art. 11 (Technical documentation); IEC 62304:2006 (Software life-cycle processes)

While this alignment provides a starting point for regulatory discussion, we emphasize that formal conformity assessment, clinical validation, and jurisdictional approval processes remain essential steps before deployment. Our work contributes methodological foundations and transparency artifacts that may facilitate such assessments in future regulatory submissions.

Although our model effectively fuses textual and visual modalities, it currently does not incorporate structured metadata fields, such as intended clinical use, manufacturer identity, or material composition, which are often central to regulatory assessments. Incorporating such information could further improve classification fidelity, particularly for borderline or novel device types. Additionally, the self-training strategy, while effective in augmenting the training set with high-confidence pseudo-labels, assumes that prediction confidence correlates with correctness. This assumption may not always hold, especially in the presence of class imbalance or emerging device categories, introducing the risk of confirmation bias.

Future research could focus on integrating structured metadata, strengthening model explainability (e.g., through SHAP-based interpretation of classification decisions), and validating model performance in collaboration with regulatory professionals under real-world operational conditions. Bridging technical innovation with the practical needs and constraints of regulatory ecosystems will be essential for translating AI capabilities into trustworthy tools for medical device oversight.

8.6 Conclusion and Future Work

We present a novel integration of Transformer-based multimodal fusion and ensemble self-training for regulatory risk classification. Our framework outperforms both unimodal baselines and early-fusion strategies, demonstrating the effectiveness of multimodal learning in the context of regulatory informatics. The results highlight that visual features provide complementary signals that help resolve ambiguity in device risk classification, and that high-confidence pseudo-labeling [327–329] can improve generalization under limited supervision. By integrating advances in vision-language modeling and semi-supervised learning, this framework establishes a scalable methodological foundation for future research at the intersection of regulatory AI and medical informatics within the NMPA context.

However, we emphasize that this work is validated exclusively on Chinese NMPA data and does not claim generalizability to other regulatory jurisdictions (FDA, EU, PMDA, TGA) without jurisdiction-specific retraining and validation. Regulatory taxonomies, linguistic conventions, and documentation standards differ fundamentally across agencies (detailed in Section 8.5.2), precluding direct model transfer. We position our framework as a proof-of-concept for multimodal regulatory AI, demonstrating that joint text–image modeling can improve risk classification accuracy when trained on appropriate jurisdiction-specific data.

Several promising directions exist for extending this work to achieve greater robustness and applicability in real-world regulatory settings. An important avenue involves adapting the framework for multilingual and cross-jurisdictional deployment. Medical device classification protocols vary considerably across regions, including systems such as the FDA (United States), NMPA (China), and MDR (European Union). Developing jurisdiction-adapted models—each trained on native regulatory data and terminology—would enable region-specific AI decision support while preserving local regulatory taxonomies.

Future work will investigate imputation-based representations for missing visual inputs and adaptive fusion mechanisms that dynamically adjust modality contributions based on estimated image quality. These strategies aim to mitigate the dependence on complete modality pairs and improve robustness in real-world regulatory submissions where visual data may be incomplete or heterogeneous.

Incorporating structured metadata alongside unstructured modalities represents another natural extension. Clinical specialty indicators, standardized product codes, and

manufacturer identifiers offer valuable domain-specific context that is often underutilized. Jointly modeling these structured signals with visual and textual embeddings could further enrich the model's representational capacity, particularly in edge cases where modality cues are ambiguous or inconsistent. Given the dynamic nature of medical device innovation, strategies for improving model generalization, such as zero-shot or few-shot learning, also warrant exploration. Pretraining with contrastive objectives or instruction tuning, as demonstrated in large-scale vision-language models [330–332], could enable the system to recognize emerging device types with minimal supervision, addressing a key challenge in rapidly evolving regulatory landscapes.

Future work could incorporate explanation mechanisms to illuminate the rationale behind risk predictions. Such capabilities are critical not only for fostering user trust but also for supporting regulatory auditing, documentation, and accountability [333]. Furthermore, ethical considerations [334] and continuous monitoring frameworks [335] will be essential for real-world deployment. Building upon recent advances in semi-supervised learning [336], these directions outline a path toward building more scalable, adaptive, and trustworthy AI systems for medical device classification, advancing the integration of machine learning into regulatory science and clinical technology governance.

9 Discussions

9.1 Thesis Summary

The thesis is strategically structured to ensure that each chapter builds logically upon the findings of the preceding chapters, creating a cohesive and comprehensive exploration of regulatory complexity in the medical device industry. The research progresses systematically, integrating theoretical analysis, empirical investigation, and practical tool development to address the core research problem.

Chapter 1 introduces the research topic by establishing the context and significance of regulatory complexity in the medical device industry. This foundation sets the stage for Chapter 2, which offers a detailed analysis of the regulatory framework and policy landscape for AI-based medical devices within the National Medical Products Administration (NMPA). Serving as a pilot study, this chapter highlights the key regulatory challenges manufacturers face and has been published in *JMIR AI* [23]. Informed by the findings of Chapter 2, Chapter 3 applies exploratory data analysis (EDA) to the NMPA's Unique Device Identifier (UDI) database, providing empirical insights into regulatory trends in China's AI medical device landscape. This work is currently in preparation, and offers real-world quantitative analysis of medical device registration.

Building on the regulatory analysis in Chapter 3, Chapter 4 systematically reviews existing metrics used to evaluate legal and regulatory texts, identifying critical gaps in how regulatory complexity is measured. This systematic review lays the groundwork for a more nuanced understanding of regulatory challenges and has been accepted by *3rd International Conference on Political Sciences (POLITICALSCIENCES)*.

Chapter 5 deepens the investigation into regulatory complexity by presenting qualitative research findings from semi-structured interviews, revealing real-world challenges faced by industry stakeholders. This exploration of the human and organizational aspects of regulatory affairs has been published in *Frontiers in Medicine* with link:https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2024.1415319/full.

Expanding upon the complexities identified in Chapter 5, Chapter 6 evaluates the potential of large language models (LLMs) to simplify regulatory tasks. Although LLMs demonstrate moderate accuracy, their human-like language generation enhances usability and interpretability. This preliminary study has been published in *Biodesign and Manufacturing* with link: https://link.springer.com/article/10.1007/s42242-024-00307-0.

Chapter 7 builds upon the insights from Chapter 6 by conducting a requirement engineering analysis of traditional machine learning algorithms and LLMs for medical device classification tasks. This chapter focuses on identifying the most cost-effective and practically applicable models, balancing accuracy, efficiency, and compliance. The goal is to refine these models for market application, potentially developing them into scalable tools or APIs for broader industry use.

Chapter 8 is a multi-modal classification AI tool design chapter. Multi-modal data, which comprises information from different sources such as text and images, has become increasingly prominent in intelligent healthcare systems. In the regulatory landscape of medical devices, both product images and associated textual descriptions contain complementary information critical to determining device classification under government health authorities (e.g., Class I/II/III under Chinese NMPA regulations). This chapter proposes a novel AI framework that effectively integrates visual and textual modalities for the task of automatic regulatory risk classification of medical devices. Finally, Chapter 9 as conclusion synthesizes the key findings from all preceding chapters, offering reflections on the research contributions and outlining future research directions.

By connecting each chapter, this dissertation presents a coherent narrative that systematically addresses the complexity of medical device regulation. The integration of policy analysis, empirical data, qualitative insights, and algorithmic evaluation ensures a comprehensive approach to developing actionable solutions that improve regulatory processes for manufacturers and ultimately benefit patients.

9.2 Research Contributions

Contribution 1 - Understanding the Current Regulatory Affairs (RA) Landscape: Policy Analysis, Exploratory Data Analysis (EDA), and Systematic Review of Legal Text Evaluation Methods

Addressing the complexity of regulatory affairs begins with a comprehensive investigation of the regulatory landscape. This contribution combines policy analysis, exploratory data analysis (EDA), and a systematic review of current methods for evaluating legal texts in regulatory affairs.

A pilot study involving policy analysis and exploratory data analysis (EDA) of NMPA regulations and the NMPA database was conducted, with findings presented in Chapters 2 and 4. Additionally, a comprehensive systematic review of existing methodologies for interpreting and evaluating legal texts was carried out to identify complexity metrics within these texts, as detailed in Chapter 3.

Contribution 1 addresses RQ1 by establishing a foundational understanding of the regulatory landscape and evaluating the tools and frameworks currently used for interpreting legal texts in regulatory affairs.

Contribution 2 - Identifying Key Origins of Complexity and Evaluating Existing Tools for Simplifying Regulatory Processes

Through qualitative semi-structured interviews conducted under IRB approval number CUREC R71265/RE003, key sources of complexity in medical device registration and compliance were uncovered as detailed in Chapter 5. In Chapter 6, the study explores the potential of large language models (LLMs) in simplifying classification tasks, as detailed in Chapter 6.

Contribution 2 addresses RQ2 by uncovering the root causes of complexity and critically evaluating the effectiveness of current tools used in regulatory processes. This contribution highlights gaps and opportunities for improvement, forming a bridge between complexity identification and practical solutions.

Contribution 3 - Comparative Analysis of Traditional Algorithms and Large Language Models (LLMs) for Medical Device Classification

Contribution 3 delivered a systematic comparison of traditional machine-learning algorithms, deep-learning models and LLMs for medical device classification, evaluating each approach in terms of accuracy, computational cost and interpretability (Chapter 7). LLMs were found to offer only moderate accuracy but more human-readable explanations, whereas models such as XGBoost and CNN provided a better balance of

accuracy, cost and interpretability. Building on this comparison, a novel multimodal framework fusing textual and visual device information was developed for automatic risk classification (Chapter 8). Refining the most cost-effective model for market application—including its potential development into an API for regulatory and medical-device use—is identified as a direction for future work (Section 9.4).

Contribution 3 answers RQ3 by analyzing the performance of different algorithms in addressing regulatory challenges. This work lays a data-driven foundation for future research and the development of advanced tools to streamline regulatory processes.

This thesis explores the intersection of artificial intelligence and regulatory science, with a particular focus on medical device regulation. The core contribution lies in using information engineering methods to identify the origins of complexity in regulatory affairs, and developing computational approaches to evaluate and mitigate regulatory complexity—an increasingly pressing challenge as global medical innovation accelerates. The work integrates traditional machine learning, deep learning, and large language models (LLMs) to automate the interpretation of legal texts and regulatory standards. A novel multimodal classification framework capable of fusing textual and visual features for risk classification of medical devices was also introduced. To address the interpretability bottleneck, explainability techniques such as SHAP and attention visualization were evaluated.

A particular emphasis is placed on regions underrepresented in regulatory AI research—especially China—and on emerging regulatory domains such as software as a medical device (SaMD). By benchmarking across Chinese and US regulatory frameworks, this thesis highlights challenges in cross-jurisdictional compliance and proposes scalable AI-based solutions for international regulatory harmonization.

9.3 Limitations

Several limitations should be borne in mind when interpreting the results of this thesis.

- **Single jurisdiction and data source.** The empirical work is based predominantly on data from China's NMPA registry; although the methods and the trade-offs they expose are argued to be general, the quantitative findings are calibrated to this single jurisdiction and have not been validated on the FDA, EU or other registries.

- **Dataset size and composition.** AI- and software-based devices form a small segment of the registry, and the multimodal subset (1,005 paired text–image devices) is modest and subject to selection bias, since only devices with publicly retrievable images could be included.
- **Language and model dependence.** The text models rely on a Mandarin-optimised encoder (`bert-base-chinese`); cross-lingual transfer to English- or Japanese-language regulatory corpora was not tested.
- **Interpretability evaluation.** Interpretability was assessed using computational explanation methods (SHAP, LIME, Integrated Gradients, attention and Grad-CAM) rather than through systematic evaluation by regulatory experts.
- **Classification rather than clinical evidence.** The work concerns regulatory risk *classification*, not the underlying clinical safety and effectiveness of the devices.
- **Scope of the qualitative and review components.** The interview study used a purposive sample of regulatory professionals, and the systematic review was limited to a defined set of databases and readability-based proxies.
- **Temporal validity.** Regulations and registries evolve continuously; the analyses represent a snapshot taken during the study period.
- **Maturity of the tools.** The AI tools achieve moderate accuracy and are presented as proofs of concept; they have not been deployed prospectively or validated in live regulatory practice.

9.4 Future Work

Looking forward, there are several compelling directions for future research that build upon the foundations laid in this thesis. First, the cross-jurisdictional framework can be extended to additional regulatory systems such as the EU MDR or Japan's PMDA to map divergences and overlaps, enhancing the global relevance of compliance AI tools. A key technical frontier lies in quantifying and visualizing regulatory complexity—through metrics like syntactic depth, semantic entropy, and cross-referencing patterns—which can provide new tools for both policymakers and automated agents.

Further development of domain-specific large language models using expert-annotated regulatory corpora would improve reasoning accuracy, especially in legally sensitive

contexts. An interactive regulatory co-pilot could be developed from the existing agent system, supporting manufacturers with document preparation, regulatory Q&A, and contextual recommendations. Integration with broader modalities—such as clinical protocols, adverse event reports, and labeling documents—will enable end-to-end regulatory intelligence workflows. Ultimately, the deployment of these AI tools in real-world regulatory practice, including pilot collaborations with authorities or industry, will be critical to understanding their usability, trustworthiness, and impact.

By continuing to bridge the gap between artificial intelligence and regulatory science, this research lays the groundwork for a new generation of intelligent, transparent, and globally aware regulatory systems.

Appendices

1 NMPA Medical Device Software Catalogue

This appendix reproduces the National Medical Products Administration (NMPA) medical device software catalogue (six-digit codes beginning with 21), referred to in Chapters 2.5 and 3.5.

Appendix 1. Medical Device Software Information in NMPA Translated from *Medical Device Classification Catalog*

No.	Primary product category	Secondary product category	Product Description	Intended Use	Example of product name	Classification
01	Treatment Planning Software	01 Radiation therapy planning system software	It usually consists of a software installation CD (or downloading the installation program from the Internet). Usually (not required), one or more specific algorithms are used to estimate the absorbed dose distribution of human organs.	Used to formulate the patient's radiation treatment plan.	Radiation treatment planning system software, gamma-ray stereotactic radiation treatment planning system software, radioactive seed source implantation treatment planning system software	III
		02 Radiation Therapy Assistant Software	It usually consists of a software installation CD (or downloading an installation program from the Internet). Provide, define, or display the setting data of the treatment machine; manually input data or directly import data from other equipment; record the data of the entire treatment stage.	It is used to compare the current parameters and preset parameters of the radiotherapy machine before the start of the planned radiotherapy and before the start of each treatment phase, and to record the data of the actual treatment phase.	Radiation therapy recording and verification system software	III
			It usually consists of a software installation CD (or downloading an installation program from the Internet). Before radiotherapy, use the acquired image information and analysis and processing results to determine the target coordinates or position.	Used to assist in the completion of radiotherapy.	Radiation therapy contour drawing software, radiation therapy simulation positioning software	III

No.	Primary product category	Secondary product category	Product Description	Intended Use	Example of product name	Classification
		03 Surgical planning software	It usually consists of a software installation CD (or downloading an installation program from the Internet). Use the obtained image information and the results of its analysis and processing to formulate a surgical plan.	Used to make a surgical plan before surgery. Except for dentistry and ear nose and throat.	Stereotactic surgery planning software, surgery simulation software, surgery planning software	III
			It usually consists of a software installation CD (or downloading an installation program from the Internet). Use the acquired image information and its analysis and processing results to formulate dental and ENT surgery plans or programs.	It is used to make surgical plans before dentistry and ENT surgery.	Digital implant design software, dental restoration design software	II
02	Image processing software	01 Medical image storage and transmission system software	It usually consists of a software installation CD (or downloading an installation program from the Internet). The image and/or video signal output by the medical imaging equipment is collected and saved to the computer hard disk for the processing of image reception, transmission, display, storage, and output between various departments of the medical department and/or between hospitals.	Used for medical image receiving, transmission, display, storage, output and other processing, for clinical diagnosis and treatment.	Image archiving and transmission system software, medical image management and communication system software, medical image archiving and communication system software	II

No.	Primary product category	Secondary product category	Product Description	Intended Use	Example of product name	Classification
		02Medical image processing software	It usually consists of a software installation CD (or downloading the installation program from the Internet). Use image processing methods to perform three-dimensional reconstruction and registration of medical images.	Used to process medical images from single mode or multiple modes.	Ultrasound image management software, endoscopy graphics workstation software, digital ultrasound workstation software, magnetic resonance image processing software, nuclear medicine workstation software, CT image processing software, X-ray angiography image processing software, digital X-ray image processing software	II
03	Data Processing Software	01 The monitoring software	It usually consists of a software installation CD (or downloading the installation program from the Internet). Obtain data from the monitoring equipment through data communication, and display and alarm in real time.	Used to obtain data from monitoring equipment, centralized real-time display, and alarm.	Central monitoring workstation software, central monitoring management software, central monitoring information center software, pregnancy-induced hypertension syndrome monitoring software	II
		02 Physiological signal processing software	It usually consists of a software installation CD (or downloading the installation program from the Internet). Analyze, process and/or transmit the collected physiological signals such as brain electricity, electrocardiogram, and myoelectricity.	It is used to analyze, process and/or transmit physiological signals such as brain electricity, electrocardiogram, and myoelectricity.	Holter analysis software, ECG workstation software, ECG data management software	II

No.	Primary product category	Secondary product category	Product Description	Intended Use	Example of product name	Classification
04	Decision Support Software	01 Drug calculation software	It is usually composed of software installation CD (or download the installation program from the Internet). Calculate the drug injection plan based on the pharmacokinetics and/or drug model, the patient's physiological parameters and physical signs, and provide recommendations for clinical injection of drugs.	Used to provide recommendations for clinical injection of drugs.	Insulin injection calculation software	III
		02 Computer Aided Diagnosis/Analysis Software	It usually consists of a software installation CD (or downloading an installation program from the Internet). Using image processing and/or data processing technology, computer software automatically recognizes the lesion, and provides clinical diagnosis and treatment basis and/or recommendations for the nature of the lesion.	The computer software automatically recognizes the lesions and provides clinical diagnosis and treatment basis and or suggestions for the nature of the lesions.	Mammary X-ray imaging computer-aided diagnosis software, colon computer-aided diagnosis software, lung computer-aided diagnosis software, breast ultrasound-assisted diagnosis software	III
			It usually consists of a software installation CD (or downloading an installation program from the Internet). Analyze images or data and give clinical reference values.	Analyze images or data and give clinical reference values.	Bone Density Computer Aided Testing Software	II
		03 Chinese medicine diagnosis and treatment software/	It usually consists of a software installation CD (or downloading an installation program from the Internet). Use the relevant theories of TCM syndromes and treatments and use data statistics and other methods to realize the analysis and diagnosis of various symptoms and/or provide treatment suggestions.	It is used to realize the analysis and diagnosis of various symptoms and/or provide treatment suggestions.	Auxiliary diagnosis and treatment software, Chinese medicine diagnosis and treatment software	II

No.	Primary product category	Secondary product category	Product Description	Intended Use	Example of product name	Classification
05	In Vitro Diagnostic Software	01 Medical microscopy image analysis software	It usually consists of a software installation CD (or downloading an installation program from the Internet). It has the functions of acquiring, transmitting, synthesizing, observing, analyzing, processing and reporting the images obtained from various microscopic equipment.	It is used for auxiliary diagnosis, analysis and archiving of images obtained from various microscopic equipment.	Medical pathology image acquisition software, medical microscopic image analysis software, chromosome analysis software, urine sediment analysis software	II
		02 Screening and analysis software	It usually consists of a software installation CD (or downloading the installation program from the Internet). Through the analysis and calculation of clinical, biochemical, immune and other measurement data, the diagnosis and evaluation of diseases can be carried out.	It is used to screen and evaluate diseases.	Prenatal screening and analysis software, Down syndrome prenatal screening and analysis software, neural tube malformation prenatal screening and analysis software, 21 trisomy syndrome risk calculation software, 18 trisomy syndrome risk calculation software, neural tube defect risk Calculation software, blood glucose data analysis software	II
06	other software	01 Rehabilitation training software	It usually consists of a software installation CD (or downloading the installation program from the Internet). It is a software system composed of hierarchical inspection, training and auxiliary treatment modules, or a single module.	Used for auxiliary treatment and rehabilitation training.	Visual function training software for children with amblyopia, visual function inspection training software	II

2

AI-Enabled Medical Devices Identified in the NMPA UDI Registry

This appendix lists the AI-enabled medical devices and their associated algorithms identified in the NMPA Unique Device Identifier (UDI) registry, referred to in Chapter 3.5.

Company name	Product name	Algorithm	Medical Specialty	Registration special pathw	Classification	Functional Category
INNERMED	VBN-001DV	Deep Learning	Otolaryngology	Normal pathway	Class II	Image/Data Processing Software
Xunfel Healthcare	Insight	General AI	Respiratory	Normal pathway	Class III	Decision Support Software - auxilary detection
Demetics Medical Tech	AI-SONIC-TH-I	General AI	Endocrinology	Normal pathway	Class III	Decision Support Software - auxilary diagnosis
United-Imaging	uAI-HematomaCare	Deep Learning	Neurology	Innovation pathway	Class III	Decision Support Software - Clinical triage
guiji	AIDRscreening	Deep Learning	Ophthalmology/ Endocrinology	Innovation pathway	Class III	Decision Support Software - auxilary diagnosis
Infervision	InferRead CTA Coronary	Deep Learning	Radiology	Normal pathway	Class III	Decision Support Software - auxilary evaluation
Infervision	InferRead CT Lung-Omni	Deep Learning	Respiratory	Normal pathway	Class III	Decision Support Software - auxilary diagnosis
United-Imaging	uAI-ChestCare	Deep Learning	Respiratory	Normal pathway	Class III	Decision Support Software - auxilary diagnosis
Infervision	InferRead CT Fracture	Deep Learning	Orthopedics	Normal pathway	Class III	Decision Support Software - Clinical triage
Tencent	MY-YD-01	General AI	Ophthalmology	Innovation pathway	Class III	Decision Support Software - auxilary diagnosis
yizhun	LCI ADS	Deep Learning	Respiratory	Normal pathway	Class III	Decision Support Software - auxilary diagnosis
Airdoc	Airdoc-AIFUNDUS	Deep Learning	Ophthalmology/ Endocrinology	Innovation pathway	Class III	Decision Support Software - auxilary diagnosis
Demetics Medical Tech	AI-SONIC	General AI	Endocrinology	Normal pathway	Class III	Decision Support Software - auxilary diagnosis
Infervision	InferVisual Surgery Planning	Deep Learning	General Surgery	Innovation pathway	Class III	Decision Support Software - Surgical planning
Zhenhealth	ZZ-EYE-CDS	Deep Learning	Ophthalmology/ Endocrinology	Normal pathway	Class III	Decision Support Software - auxilary diagnosis
Huiying Tech	HY-AIMI-FX	Deep Learning	Orthopedics	Normal pathway	Class III	Decision Support Software - auxilary diagnosis
United-Imaging	uAI-PioneerChest	Deep Learning	Respiratory	Innovation pathway	Class III	Decision Support Software - auxilary diagnosis
United-Imaging	uAI-Discover-NCP	Deep Learning	Respiratory	Normal pathway	Class III	Decision Support Software - Clinical triage
Infervision	IR-CTA-AN	Deep Learning	Neurosurgery	Normal pathway	Class III	Decision Support Software - Clinical triage
Yukun Tech	LiverMRDoc	Deep Learning	Gastroenterology	Normal pathway	Class III	Decision Support Software - Clinical triage
Yukun Tech	AneurDoc	Deep Learning	Neurosurgery	Normal pathway	Class III	Decision Support Software - Clinical triage
KFBIO	KF-TCTAnalysis	General AI	Obstetrics and Gynecology	Normal pathway	Class II	Analysis and Mining Products
Infervision	InferRead CT Pneumonia	Deep Learning	Respiratory	Normal pathway	Class III	Decision Support Software - Clinical triage
Biomind	HABMSWCOVID19CN	Deep Learning	Respiratory	Normal pathway	Class III	Decision Support Software - Clinical triage
SenseTime	SenseCare-Lung	Deep Learning	Respiratory	Normal pathway	Class III	Decision Support Software - auxilary diagnosis
Infervision	InferRead CT Lung-LA	Deep Learning	Respiratory	Normal pathway	Class III	Decision Support Software - auxilary diagnosis
Vistel	EyeWisdom	General AI	Ophthalmology/ Endocrinology	Normal pathway	Class III	Decision Support Software - auxilary diagnosis
Shukun Chuangzhi	LungdimensionDoc	Deep Learning	General Surgery	Normal pathway	Class III	Image/Data Processing Software
United-Imaging	uAI-BoneCare	Deep Learning	Orthopedics	Normal pathway	Class III	Decision Support Software - auxilary diagnosis
Infervision	InferRead CT Stroke	Deep Learning	Neurology	Normal pathway	Class III	Decision Support Software - Clinical triage
VoxelCloud	VoxelCloud Retina	General AI	Ophthalmology/ Endocrinology	Normal pathway	Class III	Decision Support Software - auxilary diagnosis
Yukun Tech	BoneageDoc	Deep Learning	Orthopedics	Normal pathway	Class III	Decision Support Software - auxilary diagnosis
Pulse Medical	AngioPlus Core	Deep Learning	Cardiology	Normal pathway	Class III	Analysis and Mining Products
91360 Med Tech	CCAA	Deep Learning	Obstetrics and Gynecology	Normal pathway	Class III	Analysis and Mining Products
Infervision	IR-CTA-ST	Deep Learning	Neurology	Normal pathway	Class III	Decision Support Software - auxilary diagnosis
Surgiplan	EC01	Deep Learning	Gastroenterology	Normal pathway	Class III	Decision Support Software - auxilary diagnosis
Infervision	InferOperate Abdomen Planning	Deep Learning	Hepatology	Normal pathway	Class III	Decision Support Software - Surgical planning
EDAN	Intelligent ECG Platform	Deep Learning	Cardiology	Normal pathway	Class III	Image/Data Processing Software

hmibiotech	LiveBoost PC	General AI	Hepatology	Normal pathway	Class II	Analysis and Mining Products
Yujin AI	YCS	General AI	Gastroenterology	Normal pathway	Class II	Decision Support Software - auxilary diagnosis
KFBIO	KF-Gastric-Analysis	General AI	Gastroenterology	Normal pathway	Class II	Analysis and Mining Products
KFBIO	KF-TH-DIAG	General AI	Endocrinology	Normal pathway	Class II	Image/Data Processing Software

3
Statement of Authorship

Although this thesis is not submitted as an integrated thesis, Chapters 2–8 are each based on a co-authored article that has been published or is currently under review. Following the guidance for integrated theses published by the Mathematical, Physical and Life Sciences (MPLS) Division, the candidate's contribution to each article is identified at the start of the relevant chapter using the Contributor Roles Taxonomy (CRediT, credit.niso.org); where the article has been published, the CRediT roles are reproduced from the contribution statement of the published paper. In place of seven separate forms, the contributions, status and copyright position for all seven articles are consolidated in the single table below, followed by one signed declaration. The candidate is the lead author of every article; where an article has been published, its inclusion in this thesis is covered by the applicable open-access licence or the publisher's standard author-reuse terms, as recorded in the table.

Consolidated Statement of Authorship

Ch.	Article / venue	Status	Candidate's contribution (CRediT)	Copyright / licence
2	Regulatory Frameworks for AI-Enabled Medical Device Software in China, *JMIR AI* (2024)	Published	Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft	Open access, CC BY; authors retain copyright
3	Data-driven analysis of AI in MDSW in China	In preparation	Lead author: conception, data curation, analysis, writing – original draft	© authors; no transfer
4	Linguistic Measurements for Understanding Legal Text: A Systematic Literature Review	Accepted (3rd Int. Conf. on Political Sciences)	Lead author: conception, review protocol, screening, analysis, writing – original draft	© authors; preprint on arXiv (2411.09497)
5	More than Red Tape, *Frontiers in Medicine* (2024)	Published	Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft & review	Open access, CC BY; authors retain copyright
6	Evaluation of LLMs for the Classification of Medical Device Software, *Bio-Design and Manufacturing* 7:819–822 (2024)	Published	Methodology, Formal analysis, Data curation, Visualization, Writing – original draft & review	Publisher (Springer) holds copyright; reused under the author's standard right to include their article in a thesis
7	Governing AI in Medical Device Regulation, *ACM Trans. Computing for Healthcare*	Under review	Lead author: conception, methodology, software, analysis, writing – original draft	© authors; no transfer
8	Toward Automated Regulatory Decision-Making, *Expert Systems with Applications* 318:131669 (2026)	Published	Conceptualization, Methodology, Data curation, Formal analysis, Software, Visualization, Writing – original draft	Publisher (Elsevier) holds copyright; reused under the author's standard right to include their article in a thesis

For the open-access articles (Chapters 2 and 5) the material is reproduced under the Creative Commons licence with appropriate attribution. For the subscription articles (Chapters 6 and 8), where the publisher holds copyright, the material is included under the author's standard right to reuse their own published article in a thesis (as permitted by the Springer and Elsevier author agreements), and the source is cited in each chapter to avoid self-plagiarism.

Declaration and signatures

I confirm that the table above accurately describes my contribution to each co-authored article reproduced in this thesis, and that, as lead author, I am entitled to include this material (under the open-access licence or the publisher's author-reuse terms for the published articles).

Candidate	**Signature**	**Date**
Yu Han	Yu Han	06/21/2026

Supervisor (on behalf of co-authors)	**Signature**	**Date**
Prof. Jeroen H. M. Bergmann		22-06-2026

The supervisor signs this statement on behalf of all co-authors. The candidate is the lead author of each article, the studies were carried out during the doctorate, and every co-author is credited in the CRediT statement at the start of the relevant chapter.

References

[1] FDA. *Advancing Regulatory Science*. URL: https://www.fda.gov/science-research/science-and-research-special-topics/advancing-regulatory-science#:~:text=Regulatory%20Science%20is%20the%20science,of%20all%20FDA%2Dregulated%20products. (visited on 2023).

[2] Ricardo Alvira. *The definition of Complex and Complexity*. Sept. 2014.

[3] Marco Pagano, Antonio Sánchez Serrano, and Jozef Zechner. *Reports of the Advisory Scientific Committee-No 9/June 2019*. 2019.

[4] Jeremy Waldron. "Vagueness in law and language: Some philosophical issues". In: *Cal L. Rev.* 82 (1994), p. 509.

[5] Hendrik Hakenes and Isabel Schnabel. "Regulatory capture by sophistication". In: *SSRN Electronic Journal* (2014).

[6] Juan de Lucio and Juan S Mora-Sanguinetti. "Drafting "better regulation": The economic cost of regulatory complexity". In: *Journal of Policy Modeling* 44.1 (2022), pp. 163–183.

[7] Dorothea Tegethoff. "Readability of Information Material in Obstetrics". In: *Zeitschrift fur Geburtshilfe und Neonatologie* 224.4 (2019), pp. 208–216.

[8] Matthew C Mason and James ML Williamson. "Readability of endoscopy information leaflets: Implications for informed consent". In: *International journal of clinical practice* 72.5 (2018), e13099.

[9] Michael K Paasche-Orlow, Holly A Taylor, and Frederick L Brancati. "Readability standards for informed-consent forms as compared with actual readability". In: *New England journal of medicine* 348.8 (2003), pp. 721–726.

[10] Mai Mohammed Alm El-Din et al. "Voluntary disclosure and complexity of reporting in Egypt: the roles of profitability and earnings management". In: *Journal of Applied Accounting Research* (2021).

[11] Jean-Edouard Colliard and Co-Pierre Georg. "Measuring regulatory complexity". In: *SSRN Electronic Journal* (2020).

[12] David Smith and Grant Richardson. "The readability of Australia's taxation laws and supplementary materials: an empirical investigation". In: *Fiscal Studies* 20.3 (1999), pp. 321–349.

[13] Giovanni Cruz et al. "Validating Design Complexity Metrics for Medical Devices". In: *IIE Annual Conference. Proceedings*. Institute of Industrial and Systems Engineers (IISE). 2014, p. 3574.

[14] Corinna Sorenson and Michael Drummond. "Improving medical device regulation: the United States and Europe in perspective". In: *The Milbank Quarterly* 92.1 (2014), pp. 114–150.

[15] Arthur Arnould, Rita Hendricusdottir, and Jeroen Bergmann. “The complexity of medical device regulations has increased, as assessed through data-driven techniques”. In: *Prosthesis* 3.4 (2021), pp. 314–330.

[16] Onkar Bedi, Pawan Krishan, and Gaaminepreet Singh. “Regulatory requirements for medical devices: an insight”. In: *Applied Clinical Research, Clinical Trials and Regulatory Affairs* 4.1 (2017), pp. 16–25.

[17] Jan B Pietzsch et al. “Stage-gate process for the development of medical devices”. In: *Journal of Medical Devices* (2009).

[18] Petra Maresova et al. “Do regulatory changes seriously affect the medical devices industry? Evidence from the Czech Republic”. In: *Frontiers in public health* 9 (2021), p. 666453.

[19] Harsh Sah, Mahima Ramesh, et al. “Regulating Software as Medical Device (SaMD): A Comprehensive Overview of The International Medical Device Regulators Forum (IMDRF) Framework and Gap Analysis of Indian Medical Devices Rules, 2017”. In: *2025 Emerging Technologies for Intelligent Systems (ETIS)* (2025), pp. 1–6.

[20] Rita Hendricusdottir et al. “Lack of support in medical device regulation within academia”. In: *Prosthesis* 3.1 (2021), p. 1.

[21] EU. *Notified Body Survey on Certificates and Applications*. URL: https://health.ec.europa.eu/system/files/2022-10/md_nb_survey_certifications_applications_en.pdf (visited on 2023).

[22] Medtech Europe. *MedTech Europe Survey Report, July 2022*. URL: https://www.medtecheurope.org/resource-library/transition-to-the-ivd-regulation-medtech-europe-survey-results-for-october-2022/ (visited on 2023).

[23] Yu Han, Aaron Ceross, and Jeroen Bergmann. “Regulatory Frameworks for AI-Enabled Medical Device Software in China: Comparative Analysis and Review of Implications for Global Manufacturer”. In: *JMIR AI* 3.1 (2024), e46871.

[24] Xiaosong Wang et al. “ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 3462–3471.

[25] Nirali M Patel et al. “Enhancing next-generation sequencing-guided cancer care through cognitive computing”. In: *The Oncologist* 23.2 (2018), p. 179.

[26] Daniel Shu Wei Ting et al. “Artificial intelligence and deep learning in ophthalmology”. In: *British Journal of Ophthalmology* 103.2 (2019), pp. 167–175.

[27] Peter Szolovits. *Artificial intelligence in medicine*. Routledge, 2019.

[28] Gregory P Forlenza. “Use of artificial intelligence to improve diabetes outcomes in patients using multiple daily injections therapy”. In: *Diabetes technology & therapeutics* 21.S2 (2019), S2–4.

[29] Christine Tzelios and Ruvandhi R Nathavitharana. “Can AI technologies close the diagnostic gap in tuberculosis?” In: *The Lancet Digital Health* 3.9 (2021), e535–e536.

[30] State Council of the People's Republic of China. *Notice of the State Council on the New Generation Artificial Intelligence Plan*. 2017. URL: https://web.archive.org/web/20170721053549/http://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm (visited on 2023).

[31] Wayne M Morrison. *China's economic rise: history, trends, challenges, and implications for the United States*. Congressional Research Service Washington, DC, 2013.

[32] worldometer. *China population in 2022*. URL: https://www.worldometers.info/world-population/china-population/ (visited on 2023).

[33] Deloitte. *Chinese Medical Device Industry: How to thrive in an increasingly competitive market?* 2021. URL: https://www2.deloitte.com/cn/en/pages/life-sciences-and-healthcare/articles/chinese-medical-device-industry-whitepaper.html (visited on 2023).

[34] Aaron Ceross and Jeroen Bergmann. "Evaluating the Presence of Software-as-a-Medical-Device in the Australian Therapeutic Goods Register". In: *Prosthesis* 3.3 (2021), pp. 221–228. URL: https://www.mdpi.com/2673-1592/3/3/22.

[35] NMPA. *Announcement of the NMPA on Issuing the Guiding Principles for Technical Review of Medical Device Software Registration*. URL: https://www.nmpa.gov.cn/ylqx/ylqxggtg/ylqxqtgg/20150805120001562.html?type=pc&m= (visited on 2023).

[36] Artificial Intelligence Medical Device Innovation and Cooperation Platform. *Key Points of Deep Learning Decision-Assisting Medical Device Software Review*. URL: http://aimd.org.cn/newsinfo/1339997.html?templateId=506998 (visited on 2023).

[37] National Medical Products Administration (NMPA). *Announcement of NMPA on Issuing the Guiding Principles for the Classification and Definition of Artificial Intelligence Medical Software Products (No. 47 of 2021)*. 2021. URL: https://www.nmpa.gov.cn/xxgk/ggtg/qtggtg/20210708111147171.html?type=pc&m= (visited on 2023).

[38] China Center for Food and Drug International Exchange (CCFDIE). *Announcement of the Center for Device Evaluation of the State Food and Drug Administration on the release of the guidelines for the registration and review of medical device software (2022 revision) (2022 No. 9)*. URL: https://www.ccfdie.org/cn/yjxx/ylqx/webinfo/2022/03/1648289422496885.htm (visited on 2023).

[39] CIRS Group. *Notice of the Center for Device Evaluation of the NMPA on Issuing the Guiding Principles for the Registration Review of Artificial Intelligence Medical Devices (No. 8, 2022)*. URL: https://www.cirs-group.com/cn/md/gjyjjqszxgyfbrgznylqxzcsczdyzdtg-2022nd8h (visited on 2023).

[40] FDA. *Software as a Medical Device (SaMD)*. URL: https://www.fda.gov/medical-devices/digital-health-center-excellence/software-medical-device-samd (visited on 2023).

[41] National Medical Products Administration (NMPA). *Announcement of the General Administration on Matters concerning the Implementation of the Catalogue of Medical Devices*. 2017. URL: https://www.nmpa.gov.cn/directory/web/nmpa/ylqx/ylqxggtg/ylqxqtgg/20170904143301827.html (visited on 2023).

[42] Erfan Soliman, Daniel Mogefors, and Jeroen HM Bergmann. "Problem-driven innovation models for emerging technologies". In: *Health and Technology* 10.5 (2020), pp. 1195–1206.

[43] Jan Maci and Petra Marešová. "Critical Factors and Economic Methods for Regulatory Impact Assessment in the Medical Device Industry". In: *Risk Management and Healthcare Policy* 15 (2022), p. 71.

[44] National Medical Products Administration (NMPA). *Regulations on the Supervision and Administration of Medical Devices*. 2021. URL: https://www.gov.cn/gongbao/content/2021/content_5595920.htm (visited on 2023).

[45] National Medical Products Administration (NMPA). *YY/T 0664-2008 Medical Device Software Life Cycle Process*. 2020.

[46] Rong Shankui et al. "Research on Classification Management of Computer Aided Diagnosis Software Products". In: *Chinese Journal of Medical Devices* 5 (2019).

[47] IMDRF Software as a Medical Device (SaMD) Working Group. *"Software as a Medical Device": Possible Framework for Risk Categorization and Corresponding Considerations*. International Medical Device Regulators Forum, 2014.

[48] Vitalii Pashkov and Andrii Harkusha. "Certain aspects on medical devices software law regulation". In: *Wiadomości lekarskie* 6 (2016), pp. 765–767.

[49] Vitalii M Pashkov, Andrii O Harkusha, and Yevheniia O Harkusha. "Artificial intelligence in medical practice: regulative issues and perspectives". In: *Wiadomości Lekarskie* (2020).

[50] FDA. *Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning(AI/ML)- Based Software as a Medical Device(SaMD))*. URL: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device (visited on 2023).

[51] U.S. Food and Drug Administration. *Digital Health Innovation Action Plan*. URL: https://web.archive.org/web/20190612193650/https://www.fda.gov/media/106331/download (visited on 2023).

[52] Min Yue et al. "IMDRF Interpretation of personalized medical device terms". In: *China Pharmaceutical Affairs* 33.1 (2019), pp. 41–44.

[53] NMPA. *NMPA launched the China Drug Regulatory Science Action Plan*. URL: https://web.archive.org/web/20210304132811/http://www.gov.cn/xinwen/2019-05/02/content_5388253.htm (visited on 2023).

[54] Jianxing He et al. "The practical implementation of artificial intelligence technologies in medicine". In: *Nature Medicine* 25.1 (2019), pp. 30–36.

[55] Yuehua Liu, Wenjin Yu, and Tharam Dillon. “Regulatory responses and approval status of artificial intelligence medical devices with a focus on China”. In: *npj Digital Medicine* 7.1 (2024), p. 255.

[56] NMPA. *Announcement of the State Food and Drug Administration on Publishing the Special Review Procedures for Innovative Medical Devices (2018 No. 83)*. URL: https://www.nmpa.gov.cn/zhuanti/cxylqx/cxylqxzyxx/20181105160001106.html?type=pc&m= (visited on 2023).

[57] NMPA. *Announcement of the General Administration on Publishing the Priority Review and Approval Procedures for Medical Devices (No. 168, 2016)*. URL: https://www.nmpa.gov.cn/xxgk/ggtg/qtggtg/20161026164001187.html (visited on 2023).

[58] NMPA. *Announcement of the State Food and Drug Administration on Issuing the "Emergency Approval Procedures for Medical Devices" (2021 No. 157)*. URL: https://www.nmpa.gov.cn/ylqx/ylqxggtg/20211230171343118.html?type=pc&m= (visited on 2023).

[59] NMPA. *Pilot Project Plan for UDI System for Medical Devices (No.56, 2019)*. URL: https://www.nmpa.gov.cn/directory/web/nmpa/xxgk/fgwj/gzwj/gzwjylqx/20190703175301499.html (visited on 2023).

[60] NMPA. *Rules for Unique Device Identification System (No. 66, 2019)*. URL: https://www.nmpa.gov.cn/ylqx/ylqxggtg/ylqxqtgg/20190827092601750.html (visited on 2023).

[61] National Medical Products Administration. *Medical Device UDI Database*. https://udi.nmpa.gov.cn/.

[62] NMPA. *YY/T 1630-2018*. URL: https://udi.nmpa.gov.cn/toDetail.html?infoId=66&CatalogId=2 (visited on 2023).

[63] National Medical Products Administration (NMPA). *Announcement of NMPA Issuing the Guidelines for the Classification and Definition of Artificial Intelligence Medical Software Products*. 2021. URL: https://www.nmpa.gov.cn/xxgk/ggtg/qtggtg/20210708111147171.html?type=pc&m= (visited on 2023).

[64] National Medical Products Administration (NMPA). *UDI Dataset reference*. URL: https://udi.nmpa.gov.cn (visited on 2023).

[65] "Medical Equipment" Magazine Issue 15, 2022. *Review points of artificial intelligence software for CT imaging*. URL: https://www.ylzbzz.org.cn/index.php?m=content&c=index&a=show&catid=43&id=152 (visited on 2023).

[66] Janet M Torpy, Alison E Burke, and Richard M Glass. “Medical Specialties”. In: *JAMA* 298.9 (2007), pp. 1120–1120.

[67] George Weisz. “The emergence of medical specialization in the nineteenth century”. In: *Bulletin of the History of Medicine* 77.3 (2003), pp. 536–574.

[68] Valerie Van Ballaer et al. “Impact of ultra-high-resolution imaging of the lungs on perceived diagnostic image quality using photon-counting CT”. In: *European Radiology* 34.3 (2024), pp. 1895–1904.

[69] Adelaide Greco et al. "Imaging techniques in veterinary medicine. Part II: Computed tomography, magnetic resonance imaging, nuclear medicine". In: *European Journal of Radiology Open* 10 (2023), p. 100467.

[70] Paidipati Dinesh, AS Vickram, and P Kalyanasundaram. "Medical image prediction for diagnosis of breast cancer disease comparing the machine learning algorithms: SVM, KNN, logistic regression, random forest and decision tree to measure accuracy". In: *AIP Conference Proceedings*. Vol. 2853. 1. AIP Publishing. 2024.

[71] S Kevin Zhou, Hayit Greenspan, and Dinggang Shen. *Deep learning for medical image analysis*. Academic Press, 2023.

[72] China Food and Drug Administration. *Rules for the Naming of Generic Names of Medical Devices*. Order No. 19 of the China Food and Drug Administration. Accessed on [insert date of access]. 2015. URL: https://www.nmpa.gov.cn/ylqx/ylqxfgwj/ylqxbmgzh/20151221120001127.html (visited on 2023).

[73] Stan Benjamens, Pranavsingh Dhunnoo, and Bertalan Meskó. "The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database". In: *NPJ digital medicine* 3.1 (2020), p. 118.

[74] Saeed Iqbal et al. "On the analyses of medical images using traditional machine learning techniques and convolutional neural networks". In: *Archives of Computational Methods in Engineering* 30.5 (2023), pp. 3173–3233.

[75] Mohammad A Thanoon et al. "A review of deep learning techniques for lung cancer screening and diagnosis based on CT images". In: *Diagnostics* 13.16 (2023), p. 2617.

[76] Valérie Bousson et al. "Application of artificial intelligence to imaging interpretations in the musculoskeletal area: Where are we? Where are we going?" In: *Joint Bone Spine* 90.1 (2023), p. 105493.

[77] Jyotismita Chaki and Marcin Woźniak. "Deep learning for neurodegenerative disorder (2016 to 2022): A systematic review". In: *Biomedical Signal Processing and Control* 80 (2023), p. 104223.

[78] Yoshihisa Takahashi et al. "Artificial intelligence and deep learning: New tools for histopathological diagnosis of nonalcoholic fatty liver disease/nonalcoholic steatohepatitis". In: *Computational and Structural Biotechnology Journal* 21 (2023), pp. 2495–2501.

[79] Jörn M Schattenberg, Naga Chalasani, and Naim Alkhouri. "Artificial intelligence applications in hepatology". In: *Clinical Gastroenterology and Hepatology* 21.8 (2023), pp. 2015–2025.

[80] Katleen Van Uytfanghe et al. "Thyroid stimulating hormone and thyroid hormones (triiodothyronine and thyroxine): an American Thyroid Association-commissioned review of current clinical and laboratory status". In: *Thyroid* 33.9 (2023), pp. 1013–1028.

[81] FDA 510(k) database. *Medical imaging system*. URL: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpcd/classification.cfm?id=5630 (visited on 2023).

[82] Aidoc. *'Time is Brain': Aidoc releases complete AI package to speed identification and treatment of stroke*. Accessed: 2024-XX-XX. 2024. URL: https://www.aidoc.com/about/news/aidoc-stroke-package/ (visited on 2023).

[83] Squire Patton Boggs. *FDA Approves Marketing of Clinical Decision Support Software for Stroke Triage in Midst of Renewed Focus on Digital Health Applications*. 2018. URL: https://web.archive.org/web/20190923093105/https://www.squirepattonboggs.com/~/media/files/insights/publications/2018/03/ahla-article-fda-approves-marketing-of-clinical-decision-support-software-for-stroke-triage-in-midst-of-renewed-focus-on-digital-health-applications/ahla_hit_alert_3818.pdf (visited on 2023).

[84] U.S. Food and Drug Administration (FDA). *Zebra Medical Vision Ltd. - 510(k) Premarket Notification*. Accessed: 2024-XX-XX. 2020. URL: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K200905 (visited on 2023).

[85] HeartFlow. *HeartFlow to Present Latest Data on AI-Enabled Coronary Artery Disease Management at TCT 2024*. Accessed: 2024-XX-XX. 2024. URL: https://www.heartflow.com/newsroom/heartflow-to-present-latest-on-ai-enabled-coronary-artery-disease-management/ (visited on 2023).

[86] Matthias Fink and Bassil Akra. "Comparison of the international regulations for medical devices–USA versus Europe". In: *Injury* 54 (2023), p. 110908.

[87] Sara Gerke, Timo Minssen, and Glenn Cohen. "Ethical and legal challenges of artificial intelligence-driven healthcare". In: *Artificial intelligence in healthcare*. Elsevier, 2020, pp. 295–336.

[88] W. Nicholson Price II, Sara Gerke, and I. Glenn Cohen. "Potential Liability for Physicians Using Artificial Intelligence". In: *JAMA* 322.18 (2019), pp. 1765–1766.

[89] Canon Medical Systems. *AICE Deep Learning Reconstruction*. Accessed: February 1, 2025. 2024. URL: https://anz.medical.canon/products/computed-tomography/aice_dlr (visited on 2023).

[90] GE Healthcare. *AIR Recon DL - Magnetic Resonance Imaging*. Accessed: February 1, 2025. 2024. URL: https://www.gehealthcare.com/products/magnetic-resonance-imaging/air-recon-dl (visited on 2023).

[91] GE Healthcare. *Vivid Ultrasound Systems*. Accessed: February 1, 2025. 2024. URL: https://www.gehealthcare.com.au/products/ultrasound/vivid (visited on 2023).

[92] Edwards Lifesciences. *Acumen Hypotension Prediction Index (HPI)*. Accessed: February 1, 2025. 2024. URL: https://www.edwards.com/healthcare-professionals/products-services/predictive-monitoring/hpi (visited on 2023).

[93] Oskar Wysocki et al. "Assessing the communication gap between AI models and healthcare professionals: Explainability, utility and trust in AI-driven clinical decision-making". In: *Artificial Intelligence* 316 (2023), p. 103839.

[94] U.S. Food and Drug Administration. *Transparency for Machine Learning-Enabled Medical Devices: Guiding Principles*. Accessed: 2025-04-30. 2024. URL: https://www.fda.gov/medical-devices/software-medical-device-samd/transparency-machine-learning-enabled-medical-devices-guiding-principles (visited on 2023).

[95] European Parliament. *Artificial Intelligence in Healthcare*. Accessed: 2025-04-30. 2022. URL: https://www.europarl.europa.eu/RegData/etudes/STUD/2022/729512/EPRS_STU%282022%29729512_EN.pdf (visited on 2023).

[96] QY Research. *Diabetic Retinopathy Fundus Image-assisted Diagnosis Software Report 2024, Global Revenue, Key Companies Market Share & Rank*. Accessed: 2024-10-07. 2024. URL: https://cn.qyresearch.com/reports/3318113/diabetic-retinopathy-fundus-image-assisted-diagnosis-software (visited on 2023).

[97] Urs J Muehlematter, Paola Daniore, and Kerstin N Vokinger. "Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis". In: *The Lancet Digital Health* 3.3 (2021), e195–e203.

[98] Artificial Intelligent Medical Device Innovation and Cooperation Platform (AIMD). *Artificial Intelligent Medical Device Innovation and Cooperation Platform*. URL: https://web.archive.org/web/20190819022936/http://www.aimd.org.cn/ (visited on 2023).

[99] U.S. Food and Drug Administration. *Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices*. Accessed: 2023-10-07. 2023. URL: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices (visited on 2023).

[100] Sally Dalton-Brown. "The ethics of medical AI and the physician-patient relationship". In: *Cambridge Quarterly of Healthcare Ethics* 29.1 (2020), pp. 115–121.

[101] Ariel Dora Stern et al. "Cybersecurity features of digital medical devices: an analysis of FDA product summaries". In: *BMJ open* 9.6 (2019), e025374.

[102] Tiffany H Kung et al. "Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models". In: *PLOS Digital Health* 2.2 (2023), e0000198.

[103] Yu Han, Aaron Ceross, and Jeroen Bergmann. "Linguistic Measurements for Understanding Legal Text: A Systematic Literature Review". In: *Proceedings of the 3rd International Conference on Political Sciences*. Accepted; preprint available as arXiv:2411.09497. 2024. arXiv: 2411.09497. URL: https://arxiv.org/abs/2411.09497.

[104] Denis Galligan. *Law in modern society*. OUP Oxford, 2006.

[105] Kim Lane Scheppele, Dimitry Vladimirovich Kochenov, and Barbara Grabowska-Moroz. "EU values are law, after all: Enforcing EU values through systemic infringement actions by the European Commission and the member states of the European Union". In: *Yearbook of European law* 39 (2020), pp. 3–121.

[106] Talcott Parsons. "The law and social control". In: *The Sociology of Law*. Routledge, 2017, pp. 334–338.

[107] Lance W Rook. "Laying down the law: canons for drafting complex legislation". In: *Or. L. Rev.* 72 (1993), p. 663.

[108] Stephen M Ross. "On legalities and linguistics: Plain language legislation". In: *Buffalo Law Review* 30 (1981), p. 317.

[109] Michael J Bommarito II, Daniel Martin Katz, and Eric M Detterman. "LexNLP: Natural language processing and information extraction for legal and regulatory texts". In: *Research Handbook on Big Data Law*. Edward Elgar Publishing, 2021, pp. 216–227.

[110] Hannah Bradford Clauss. "The History of The Plain Language Movement and Legal Language and an Analysis of US Nuclear Treaty Language". Chancellor's Honors Program thesis. MA thesis. The University of Tennessee, Knoxville, 2020.

[111] Christopher Williams et al. "Legal English and plain language: An introduction". In: *ESP across Cultures* 1.1 (2004), pp. 111–124.

[112] Carl Felsenfeld. "Plain English Movement, The Plain English Movement: Panel Discussion". In: *Can. Bus. LJ* 6 (1981), p. 408.

[113] Joseph Kimble. "Flimsy Claims for Legalese and False Criticisms of Plain Language: A 30-Year Collection". In: *Scribes J. Leg. Writing* 19 (2020), p. 1.

[114] John Maeda. *The laws of simplicity*. MIT press, 2006.

[115] David Moher et al. "Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement". In: *Annals of internal medicine* 151.4 (2009), pp. 264–269.

[116] Specialist Unit for Review Evidence (SURE). *Questions to assist with the critical appraisal of qualitative studies*. 2015.

[117] Rita Sommers, Cornelius Van Staden, and Francois Steffens. "Views of clinical trial participants on the readability and their understanding of informed consent documents". In: *AJOB Empirical Bioethics* 8.4 (2017), pp. 277–284.

[118] Sara A Fowler, Carol J Saunders, and Mark A Hoffman. "Variation among consent forms for clinical whole exome sequencing". In: *Journal of genetic counseling* 27.1 (2018), pp. 104–114.

[119] Makbule Neslişah Tan et al. "To understand or not to understand: This is the problem". In: *The Turkish Journal of Gastroenterology* 29.6 (2018), p. 642.

[120] Simon James and Ian Wallschutzky. "Tax law improvement in Australia and the UK: the need for a strategy for simplification". In: *Fiscal Studies* 18.4 (1997), pp. 445–460.

[121] Bobbie Cook Martindale, Bruce S Koch, and Stewart S Karlinsky. "Tax law complexity: The impact of style". In: *The Journal of Business Communication (1973)* 29.4 (1992), pp. 383–400.

[122] Andrea N Mahnke et al. "A rural community's involvement in the design and usability testing of a computer-based informed consent process for the personalized medicine research project". In: *American Journal of Medical Genetics Part A* 164.1 (2014), pp. 129–140.

[123] AE Fischer et al. "The readability of informed consent forms for research studies conducted in South Africa". In: *South African Medical Journal* 111.2 (2021), pp. 180–183.

[124] Giuseppe Pastore et al. "Readability of foot and ankle consent forms in Queensland". In: *ANZ Journal of Surgery* 90.12 (2020), pp. 2549–2552.

[125] Matthew R Edmunds, Robert J Barry, and Alastair K Denniston. "Readability assessment of online ophthalmic patient information". In: *JAMA ophthalmology* 131.12 (2013), pp. 1610–1616.

[126] Musa Sulaiman Umar and Natrah Saad. "Readability Assessment of Nigerian Company Income Tax Act." In: *Jurnal Pengurusan* 44 (2015).

[127] Lydia O'Sullivan et al. "Readability and understandability of clinical research patient information leaflets and consent forms in Ireland and the UK: a retrospective quantitative analysis". In: *BMJ open* 10.9 (2020), e037994.

[128] Fatma S Gargoum and Shaun T O'Keeffe. "Readability and content of patient information leaflets for endoscopic procedures". In: *Irish Journal of Medical Science (1971-)* 183.3 (2014), pp. 429–432.

[129] Luka Vučemilo and Ana Borovečki. "Readability and content assessment of informed consent forms for medical procedures in Croatia". In: *PloS one* 10.9 (2015), e0138017.

[130] Christine Grady et al. "A randomized trial comparing concise and standard consent forms in the START trial". In: *PLoS One* 12.4 (2017), e0172607.

[131] Maurice J Meade and Craig W Dreyer. "Orthodontic treatment consent forms: A readability analysis". In: *Journal of Orthodontics* 49.1 (2022), pp. 32–38.

[132] Olga Blinova and Nikita Tarasov. "A hybrid model of complexity estimation: Evidence from Russian legal texts". In: *Frontiers in Artificial Intelligence* 5 (2022), p. 248.

[133] Itamar Ashkenazi et al. "Evaluation of length and complexity of written consent forms in English and Hebrew for participation in clinical trials authorized in one medical institution in Israel: a descriptive study". In: *Accountability in Research* 27.3 (2020), pp. 138–145.

[134] Nilay Boztaş et al. "Evaluating the readibility of informed consent forms available before anaesthesia: a comparative study". In: *Turkish journal of anaesthesiology and reanimation* 42.3 (2014), p. 140.

[135] Roberto Rivera, John Shelton Reed, and Deborah Menius. "Evaluating the readability of informed consent forms used in contraceptive clinical trials". In: *International Journal of Gynecology & Obstetrics* 38.3 (1992), pp. 227–230.

[136] Timothy J Mader and Stephen J Playe. "Emergency medicine research consent form readability assessment". In: *Annals of emergency medicine* 29.4 (1997), pp. 534–539.

[137] Paul P Christopher et al. "Consent form readability and educational levels of potential participants in mental health research". In: *Psychiatric Services* 58.2 (2007), pp. 227–232.

[138] Shlomo A Koyfman et al. "Consent form heterogeneity in cancer trials: the cooperative group and institutional review board gap". In: *Journal of the National Cancer Institute* 105.13 (2013), pp. 947–953.

[139] S Michael Sharp. "Consent documents for oncology trials: does anybody read these things?" In: *American Journal of Clinical Oncology* 27.6 (2004), pp. 570–575.

[140] Alessandra Storino et al. "Assessing the accuracy and readability of online health information for patients with pancreatic cancer". In: *JAMA surgery* 151.9 (2016), pp. 831–837.

[141] Fabienne Santel et al. “Assessing readability and comprehension of informed consent materials for medical device research: A survey of informed consents from FDA’s Center for Devices and Radiological Health”. In: *Contemporary Clinical Trials* 85 (2019), p. 105831.

[142] Elisa J Gordon et al. “Are informed consent forms for organ transplantation and donation too difficult to read?” In: *Clinical transplantation* 26.2 (2012), pp. 275–283.

[143] HE Taylor and DEP Bramley. “An analysis of the readability of patient information and consent forms used in research studies in anaesthesia in Australia and New Zealand”. In: *Anaesthesia and intensive care* 40.6 (2012), pp. 995–998.

[144] Monaliza Hernandez Mamac. “Linguistic complexity of public legal information texts for young persons”. In: *Text & Talk* (2023).

[145] Elena Ioriatti. “Common Contexts of Meaning in the European Legal Setting: Opening Pandora’s box?” In: *International Journal for the Semiotics of Law-Revue internationale de Sémiotique juridique* (2022), pp. 1–17.

[146] J Peter Kincaid et al. *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. Tech. rep. Naval Technical Training Command Millington TN Research Branch, 1975.

[147] Janice Redish. “Readability formulas have even more limitations than Klare discusses”. In: *ACM Journal of Computer Documentation (JCD)* 24.3 (2000), pp. 132–137.

[148] Alexis Koster and D Stott Parker. “Regulation management and logic programming”. In: *Software: Practice and Experience* 20.1 (1990), pp. 79–107.

[149] Stephen Golub. “Legal empowerment: impact and implications for the development community and the World Bank”. In: *The World Bank Legal Review, Volume 2: Law, Equity and Development*. Brill Nijhoff, 2006, pp. 167–184.

[150] Anne Rasmussen and Dimiter Toshkov. “The effect of stakeholder involvement on legislative duration: Consultation of external actors and legislative duration in the European Union”. In: *European Union Politics* 14.3 (2013), pp. 366–387.

[151] Claude E Shannon. “A mathematical theory of communication”. In: *The Bell system technical journal* 27.3 (1948), pp. 379–423.

[152] Daniel Martin Katz, Michael J Bommarito II, and Josh Blackman. “Predicting the behavior of the supreme court of the united states: A general approach”. In: *arXiv preprint arXiv:1407.6333* (2014).

[153] Jean-Edouard Colliard and Co-Pierre Georg. “Measuring regulatory complexity”. In: *HEC Paris Research Paper No. FIN-2020-1358* (2022).

[154] Yu Han, Aaron Ceross, and Jeroen Bergmann. “More than red tape: exploring complexity in medical device regulatory affairs”. In: *Frontiers in Medicine* 11 (2024), p. 1415319.

[155] Knut Blind. “15. The impact of regulation on innovation”. In: *Handbook of innovation policy impact* 450 (2016).

[156] John J Tobin and Gary Walsh. *Medical product regulatory affairs: pharmaceuticals, diagnostics, medical devices*. John Wiley & Sons, 2023.

[157] Ichiro Sakuma. “Regulatory science based approach in development of novel medical devices”. In: *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2015, pp. 1729–1732.

[158] Rosa Mayelin Guerra-Bretaña and Andrea Lucía Flórez-Rendón. "Impact of regulations on innovation in the field of medical devices". In: *Research on biomedical engineering* 34 (2018), pp. 356–367.

[159] Ilke Onur and Magnus Söderberg. "The impact of regulatory review time on incremental and radical innovation: evidence from the high-risk medical device market". In: *Journal of Regulatory Economics* 57 (2020), pp. 134–158.

[160] Ariel Dora Stern. "Innovation under regulatory uncertainty: Evidence from medical technology". In: *Journal of public economics* 145 (2017), pp. 181–200.

[161] Gregory D Curfman and Rita F Redberg. "Medical devices—balancing regulation and innovation". In: *The ethical challenges of emerging medical technologies*. Routledge, 2020, pp. 49–51.

[162] James Trevelyan. "Extending Engineering Practice Research with Shared Qualitative Data." In: *Advances in Engineering Education* 5.2 (2016), n2.

[163] Sanket S Dhruva et al. "Experts' views on FDA regulatory standards for drug and high-risk medical devices: implications for patient care". In: *Journal of General Internal Medicine* 37.16 (2022), pp. 4176–4182.

[164] Julie Polisena and Gayatri Jayaraman. "Use of real-world data and evidence for medical devices: a qualitative study of key informant interviews". In: *International Journal of Technology Assessment in Health Care* 36.6 (2020), pp. 579–584.

[165] Weiwei Zhang et al. "Factors contributing to patients' preferences for primary health care institutions in China: A qualitative study". In: *Frontiers in public health* 8 (2020), p. 414.

[166] Rana B Khoury, Jennifer Cyr, and Sara Wallace Goodman. "Sampling hard-to-reach populations". In: *Doing good qualitative research* (2024), p. 122.

[167] Greg Guest, Kathleen M MacQueen, and Emily E Namey. *Applied thematic analysis*. sage publications, 2011.

[168] Muzafar Rasool Bhat et al. "Deep LDA: A new way to topic model". In: *Journal of Information and Optimization Sciences* 41.3 (2020), pp. 823–834.

[169] Tommi Vatanen, Jaakko J Väyrynen, and Sami Virpioja. "Language Identification of Short Text Segments with N-gram Models." In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*. 2010.

[170] Joseph L Hackett and Steven I Gutman. "Introduction to the Food and Drug Administration (FDA) regulatory process". In: *Journal of proteome research* 4.4 (2005), pp. 1110–1113.

[171] AP Meshkovskiy et al. "Responsibilities of the Marketing Authorisation Holders in Respect of GMP Compliance". In: *Drug development & registration* 9.4 (2020), pp. 164–170.

[172] Lina Keutzer, Ulrika SH Simonsson, et al. "Medical device apps: an introduction to regulatory affairs for developers". In: *JMIR mHealth and uHealth* 8.6 (2020), e17567.

[173] Thomas Hatzichronoglou. *Revision of the high-technology sector and product classification*. 1997.

[174] Daniel B Kramer and Robert W Yeh. "Practical improvements for medical device evaluation". In: *JAMA* 318.4 (2017), pp. 332–334.

[175] Josue R Crespo-Varela, Lourdes A Medina, and Gül E Okudan Kremer. "An Analysis of the Relationship of Medical Device Complexity and FDA". In: *IIE Annual Conference. Proceedings*. Institute of Industrial and Systems Engineers (IISE). 2012, p. 1.

[176] Alasdair Breckenridge. "Regulatory divergence". In: *Clinical Trials* 8.4 (2011), p. 425.

[177] Zachery Matthew Thompson. "Analysis of variability in regulatory requirements for product approval, manufacturing, and post-market surveillance of human and veterinary medical devices in the United States". PhD thesis. University of Georgia, 2013.

[178] William B Cavnar, John M Trenkle, et al. "N-gram-based text categorization". In: *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*. Vol. 161175. Las Vegas, NV. 1994, p. 14.

[179] Hamed Jelodar et al. "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey". In: *Multimedia tools and applications* 78 (2019), pp. 15169–15211.

[180] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.

[181] Mahedi Hasan et al. "Normalized approach to find optimal number of topics in Latent Dirichlet Allocation (LDA)". In: *Proceedings of International Conference on Trends in Computational and Cognitive Engineering: Proceedings of TCCE 2020*. Springer. 2021, pp. 341–354.

[182] Radim Rehurek and Petr Sojka. "Gensim–python framework for vector space modelling". In: *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic* 3.2 (2011).

[183] David Easley and Maureen O'Hara. "Ambiguity and nonparticipation: The role of regulation". In: *The Review of Financial Studies* 22.5 (2009), pp. 1817–1843.

[184] Adam I Muchmore. "Uncertainty, complexity, and regulatory design". In: *Hous. L. Rev.* 53 (2015), p. 1321.

[185] Juan De Lucio and Juan S Mora-Sanguinetti. "New dimensions of regulatory complexity and their economic cost. An analysis using text mining". In: *SSRN Electronic Journal* (2021).

[186] Aaron Ceross and Jeroen Bergmann. "A machine learning approach for medical device classification". In: *Proceedings of the 14th International Conference on Theory and Practice of Electronic Governance*. 2021, pp. 285–291.

[187] Patel Drashti et al. "In-depth review on 'innovation and regulatory challenges of the drug delivering medical devices'". In: *Journal of Generic Medicines* 15.1 (2019), pp. 18–28.

[188] Kevin Crowston, Eileen E Allen, and Robert Heckman. "Using natural language processing technology for qualitative data analysis". In: *International Journal of Social Research Methodology* 15.6 (2012), pp. 523–543.

[189] J Wilkinson. *EU regulatory system brings Europeans fastest access to medical technology without compromising safety*. 2011.

[190] Deborah Cohen and Matthew Billingsley. "Europeans are left to their own devices". In: *BMJ* 342 (2011).

[191] Ana Đanić Čeko and Tunjica Petrašević. "Lack of One Common Administrative Procedure? Is Regulation on the EU Administrative Procedure Necessary in the EU?" In: *Balkan Social Science Review* 15.15 (2020).

[192] Virginia Braun and Victoria Clarke. "To saturate or not to saturate? Questioning data saturation as a useful concept for thematic analysis and sample-size rationales". In: *Qualitative research in sport, exercise and health* 13.2 (2021), pp. 201–216.

[193] Matej Ulčar and Marko Robnik-Šikonja. "Training dataset and dictionary sizes matter in bert models: the case of baltic languages". In: *International Conference on Analysis of Images, Social Networks and Texts*. Springer. 2021, pp. 162–172.

[194] Jeroen HM Bergmann. *The Emerging Field of Medical Regulatory Technology and Data Science*. 2022.

[195] Arun James Thirunavukarasu et al. "Large language models in medicine". In: *Nature medicine* 29.8 (2023), pp. 1930–1940.

[196] Yu Han et al. "Evaluation of large language models for the classification of medical device software". In: *Bio-Design and Manufacturing* 7 (2024), pp. 819–822.

[197] Yu Han and Jingwen Tao. "Revolutionizing Pharma: Unveiling the AI and LLM Trends in the Pharmaceutical Industry". In: *arXiv preprint arXiv:2401.10273* (2024).

[198] Yu Han, Aaron Ceross, and Jeroen HM Bergmann. "Uncovering Regulatory Affairs Complexity in Medical Products: A Qualitative Assessment Utilizing Open Coding and Natural Language Processing (NLP)". In: *arXiv preprint arXiv:2401.02975* (2023).

[199] Thomas P Quinn et al. "The three ghosts of medical AI: Can the black-box present deliver?" In: *Artificial intelligence in medicine* 124 (2022), p. 102158.

[200] Kicky G van Leeuwen et al. "Artificial intelligence in radiology: 100 commercially available products and their scientific evidence". In: *European radiology* 31 (2021), pp. 3797–3804.

[201] Dan Hendrycks et al. "Measuring massive multitask language understanding". In: *arXiv preprint arXiv:2009.03300* (2020).

[202] Gemini Team et al. "Gemini: a family of highly capable multimodal models". In: *arXiv preprint arXiv:2312.11805* (2023).

[203] Jinze Bai et al. "Qwen technical report". In: *arXiv preprint arXiv:2309.16609* (2023).

[204] Aiyuan Yang et al. "Baichuan 2: Open large-scale language models". In: *arXiv preprint arXiv:2309.10305* (2023).

[205] Aohan Zeng et al. "Glm-130b: An open bilingual pre-trained model". In: *arXiv preprint arXiv:2210.02414* (2022).

[206] Yu Sun et al. "Ernie: Enhanced representation through knowledge integration". In: *arXiv preprint arXiv:1904.09223* (2019).

[207] Albert Q Jiang et al. "Mistral 7B". In: *arXiv preprint arXiv:2310.06825* (2023).

[208] Fiona Fui-Hoon Nah et al. *Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration*. 2023.

[209] Yu Han, Aaron Ceross, and Jeroen Bergmann. "Governing AI in Medical Device Regulation: A Framework for Risk-Stratified Accuracy, Explainability, and Lifecycle Feasibility". Manuscript under review. 2024.

[210] U.S. Food and Drug Administration. *FDA announces completion of first AI-assisted scientific review pilot and aggressive agency-wide AI rollout.* Accessed: 2025-05-10. May 2025. URL: https://www.fda.gov/news-events/press-announcements/fda-announces-completion-first-ai-assisted-scientific-review-pilot-and-aggressive-agency-wide-ai (visited on 2024).

[211] Will Douglas Heaven. "OpenAI and the FDA are talking about AI for drug evaluation". In: *WIRED* (May 2025). Accessed: 2025-05-10. URL: https://www.wired.com/story/openai-fda-doge-ai-drug-evaluation/ (visited on 2024).

[212] David Mantus and Douglas J Pisano. *FDA regulatory affairs*. CRC Press, 2014.

[213] B Sai Kumari et al. "Current Regulatory Requirements for Registration of Medicines, Compilation and Submission of Dossier in Australian Therapeutic goods Administration". In: *International Journal of Advanced Scientific and Technical Research, ISSN* (2016), pp. 2249–9954.

[214] Urs J Muehlematter, Christian Bluethgen, and Kerstin N Vokinger. "FDA-cleared artificial intelligence and machine learning-based medical devices and their 510 (k) predicate networks". In: *The Lancet Digital Health* 5.9 (2023), e618–e626.

[215] Jeffrey K Aronson, Carl Heneghan, and Robin E Ferner. "Medical devices: definition, classification, and regulatory implications". In: *Drug safety* 43.2 (2020), pp. 83–93.

[216] Jennifer L Pomeranz. "A comprehensive strategy to overhaul FDA authority for misleading food labels". In: *American journal of law & medicine* 39.4 (2013), pp. 617–647.

[217] Anne Herman et al. "Position statement: the need for EU legislation to require disclosure and labelling of the composition of medical devices". In: *Journal of the European Academy of Dermatology and Venereology* 35.7 (2021), pp. 1444–1448.

[218] Jiamin Yin, Kee Yuan Ngiam, and Hock Hai Teo. "Role of artificial intelligence applications in real-life clinical practice: systematic review". In: *Journal of medical Internet research* 23.4 (2021), e25759.

[219] Isabella Castiglioni et al. "AI applications to medical images: From machine learning to deep learning". In: *Physica medica* 83 (2021), pp. 9–24.

[220] Osamu Imaichi, Toshihiko Yanase, and Yoshiki Niwa. "A comparison of rule-based and machine learning methods for medical information extraction". In: *The first workshop on natural language processing for medical and healthcare fields*. 2013, pp. 38–42.

[221] Ruchika S Patil, Samruddhi B Kulkarni, and Vinod L Gaikwad. "Artificial intelligence in pharmaceutical regulatory affairs". In: *Drug Discovery Today* (2023), p. 103700.

[222] Tushar Khinvasara, Nikolaos Tzenios, and Abhishek Shanker. "Post-Market Surveillance of Medical Devices Using AI". In: *Journal of Complementary and Alternative Medical Research* 25.7 (2024), pp. 108–122.

[223] D Maghesh Kumar. "Automatic induction of rule based text categorization". In: *International Journal of Computer Science & Information Technology (IJCSIT)* 2.6 (2010).

[224] Aurangzeb Khan et al. "A review of machine learning algorithms for text-documents classification". In: *Journal of advances in information technology* 1.1 (2010), pp. 4–20.

[225] Andreas Korger and Joachim Baumeister. "Rule-based Semantic Relation Extraction in Regulatory Documents." In: *Proceedings of the Conference on Lernen, Wissen, Daten, Analysen (LWDA)*. 2021, pp. 26–37.

[226] Mariette Awad and Rahul Khanna. "Support vector machines for classification". In: *Efficient learning machines: Theories, concepts, and applications for engineers and system designers* (2015), pp. 39–66.

[227] Geoffrey I Webb, Eamonn Keogh, and Risto Miikkulainen. "Naıve Bayes." In: *Encyclopedia of machine learning* 15.1 (2010), pp. 713–714.

[228] Dareen M Salama and Nora M El-Gohary. "Semantic text classification for supporting automated compliance checking in construction". In: *Journal of Computing in Civil Engineering* 30.1 (2016), p. 04014106.

[229] Jiansong Zhang and Nora M El-Gohary. "Semantic NLP-based information extraction from construction regulatory documents for automated compliance checking". In: *Journal of Computing in Civil Engineering* 30.2 (2016), p. 04015014.

[230] Sri Banerjee et al. "Large language modeling and classical AI methods for the future of healthcare". In: *Journal of Medicine, Surgery, and Public Health* 1 (2023), p. 100026.

[231] Kevin Wu et al. "Regulating AI Adaptation: An Analysis of AI Medical Device Updates". In: *arXiv preprint arXiv:2407.16900* (2024). URL: https://arxiv.org/abs/2407.16900 (visited on 2024).

[232] Xudong Shen et al. "Towards Regulatable AI Systems: Technical Gaps and Policy Recommendations". In: *arXiv preprint arXiv:2306.12609* (2023). URL: https://arxiv.org/abs/2306.12609 (visited on 2024).

[233] Xubin Wang et al. "Empowering Edge Intelligence: A Comprehensive Survey on On-Device AI Models". In: *arXiv preprint arXiv:2503.06027* (2025). URL: https://arxiv.org/abs/2503.06027 (visited on 2024).

[234] Shun-ichi Amari and Si Wu. "Improving support vector machine classifiers by modifying kernel functions". In: *Neural networks* 12.6 (1999), pp. 783–789.

[235] Tianqi Chen and Carlos Guestrin. "XGBoost: A scalable tree boosting system". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016), pp. 785–794.

[236] Scott M Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *Advances in Neural Information Processing Systems (NeurIPS)* 30 (2017).

[237] Yoon Kim. "Convolutional neural networks for sentence classification". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2014, pp. 1746–1751.

[238] Wenpeng Yin et al. "Comparative study of CNN and RNN for natural language processing". In: *arXiv preprint arXiv:1702.01923* (2017).

[239] Rie Johnson and Tong Zhang. "Deep pyramid convolutional neural networks for text categorization". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2017, pp. 562–570.

[240] Siwei Lai et al. "Recurrent convolutional neural networks for text classification". In: *Proceedings of the 29th AAAI Conference on Artificial Intelligence*. 2015, pp. 2267–2273.

[241] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. “Axiomatic attribution for deep networks”. In: *International Conference on Machine Learning (ICML)* (2017), pp. 3319–3328.

[242] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “"Why should I trust you?": Explaining the predictions of any classifier”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 1135–1144.

[243] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems*. 2017, pp. 5998–6008.

[244] Jacob Devlin et al. “BERT: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).

[245] Yiming Cui et al. “Pre-training with whole word masking for Chinese BERT”. In: *arXiv preprint arXiv:1906.08101* (2019).

[246] Hugo Touvron et al. “LLaMA: Open and efficient foundation language models”. In: *arXiv preprint arXiv:2302.13971* (2023).

[247] DeepSeek AI. “DeepSeek-VL: Scaling Vision-Language Alignment with Decoupled Pretraining”. In: *arXiv preprint arXiv:2401.14185* (2024).

[248] Alon Jacovi and Yoav Goldberg. “Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?” In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)* (2020), pp. 4198–4205.

[249] Cynthia Rudin. “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead”. In: *Nature Machine Intelligence* 1.5 (2019), pp. 206–215.

[250] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “Anchors: High-precision model-agnostic explanations”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 32.1 (2018).

[251] Scott M Lundberg et al. “Explainable machine-learning predictions for the prevention of hypoxaemia during surgery”. In: *Nature Biomedical Engineering* 2.10 (2018), pp. 749–760.

[252] Scott M Lundberg et al. “Local explanation methods for tree-based models: A unified approach”. In: *Nature Machine Intelligence* 2.1 (2020), pp. 56–67.

[253] Emma Strubell, Ananya Ganesh, and Andrew McCallum. “Energy and policy considerations for deep learning in NLP”. In: *arXiv preprint arXiv:1906.02243* (2019).

[254] Zihan Wu et al. “A survey on efficient training of transformers”. In: *arXiv preprint arXiv:2202.06930* (2022).

[255] Victor Sanh et al. “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter”. In: *arXiv preprint arXiv:1910.01108* (2019).

[256] Chuxin Zhang et al. *Towards Practical Trade-offs Between Interpretability and Performance: A Composite Interpretability Metric for Trustworthy AI*. 2024. arXiv: 2503.07914 [cs.LG].

[257] Harini Kaur et al. “Trust and Transparency in Human-AI Interaction: A Survey of Trust Calibration, User Understanding, and Interpretability”. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. 2022, pp. 487–498.

[258] Sitao Chen, Tongshuang Wu, and Q. Vera Lee. "A Taxonomy of Interpretability in Human-Centered AI: From Explanations to User Experience". In: *IEEE Transactions on Visualization and Computer Graphics* (2024).

[259] Abhisek Ukil and Abhisek Ukil. "Support vector machine". In: *Intelligent systems and signal processing in power engineering* (2007), pp. 161–226.

[260] Emily Reif et al. "Visualizing and measuring the geometry of BERT". In: *Advances in neural information processing systems* 32 (2019).

[261] National Medical Products Administration. *UDI Database*. `https://udi.nmpa.gov.cn/download.html`. Accessed: 2024-09-25. n.d.

[262] Tasos. *Split the data between the Training Data and Test Data using sklearn*. Accessed: 2025-05-04. 2019. URL: `https://datascience.stackexchange.com/questions/51236/split-the-data-between-the-training-data-and-test-data-using-sklearn` (visited on 2024).

[263] Trevor Hastie, Robert Tibshirani, Jerome Friedman, et al. *The elements of statistical learning*. 2009.

[264] Fabian Pedregosa et al. "Scikit-learn: Machine learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[265] Miles Turpin et al. "Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 74952–74965.

[266] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. "A training algorithm for optimal margin classifiers". In: *Proceedings of the fifth annual workshop on Computational learning theory* (1992), pp. 144–152.

[267] David Alvarez-Melis and Tommi S Jaakkola. "On the robustness of interpretability methods". In: *arXiv preprint arXiv:1806.08049* (2018).

[268] Michael Veale and Frederik Zuiderveen Borgesius. "Demystifying the Draft EU Artificial Intelligence Act". In: *Computer Law Review International* 22.4 (2021), pp. 97–112.

[269] Martin Ebers et al. "Standardizing AI: The case of the European Commission's proposal for an Artificial Intelligence Act". In: *SSRN Electronic Journal* (2021).

[270] Thomas J Hwang, Aaron S Kesselheim, and Kerstin N Vokinger. "Lifecycle regulation of artificial intelligence- and machine learning-based medical devices: a review". In: *JAMA* 328.9 (2022), pp. 882–883.

[271] International Organization for Standardization. *Medical device software – Software life cycle processes*. Standard IEC 62304:2006+AMD1:2015. Geneva, Switzerland, 2015.

[272] National Medical Products Administration (NMPA). *Medical Device Classification Rules*. `https://www.nmpa.gov.cn/`. China Food and Drug Administration Order No. 15. 2017.

[273] State Council of the People's Republic of China. *Regulations on the Supervision and Administration of Medical Devices*. `http://www.gov.cn/`. State Council Order No. 739, Revised 2021. 2021.

[274] Center for Medical Device Evaluation, NMPA. *Guiding Principles for Registration Review of Artificial Intelligence Medical Software Products.* https://www.cmde.org.cn/. CMDE Technical Guidelines. 2022.

[275] European Commission. *Proposal for a Regulation on a European Approach for Artificial Intelligence.* Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206. 2021.

[276] U.S. Food and Drug Administration (FDA). *Good Machine Learning Practice for Medical Device Development: Guiding Principles.* Available at: https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles. 2021.

[277] Yu Han, Aaron Ceross, and Jeroen H. M. Bergmann. "Toward Automated Regulatory Decision-Making: AI-Assisted Medical Device Risk Classification with Multimodal Transformers and Self-Training". In: *Expert Systems with Applications* 318 (2026), p. 131669.

[278] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. "Multimodal machine learning: A survey and taxonomy". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.2 (2019), pp. 423–443.

[279] Bo Jing, Pengtao Xie, and Eric Xing. "On the automatic generation of medical imaging reports". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* 2018, pp. 2577–2586.

[280] Pablo Messina et al. "A survey on deep learning and explainability for automatic report generation from medical images". In: *ACM Computing Surveys (CSUR)* 54.10s (2022), pp. 1–40.

[281] Zhiyong Lu and W John Wilbur. "Overview of BioCreative III gene normalization". In: *Proceedings of the BioCreative III Workshop.* 2009, pp. 1–10.

[282] Sicen Liu et al. "Multimodal data matters: Language model pre-training over structured and unstructured electronic health records". In: *IEEE Journal of Biomedical and Health Informatics* 27.1 (2022), pp. 504–514.

[283] Liunian Harold Li et al. "VisualBERT: A simple and performant baseline for vision and language". In: *arXiv preprint arXiv:1908.03557* (2019).

[284] Hao Tan and Mohit Bansal. "LXMERT: Learning cross-modality encoder representations from transformers". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).* Association for Computational Linguistics. 2019, pp. 5100–5111.

[285] Qizhe Xie et al. "Self-training with noisy student improves imagenet classification". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2020, pp. 10687–10698.

[286] World Health Organization. *Ethics and governance of artificial intelligence for health: large multi-modal models. WHO guidance.* World Health Organization, 2024.

[287] European Commission. *Proposal for a Regulation laying down harmonised rules on artificial intelligence*. COM(2021) 206 final; last updated 20 June 2023. 2021. URL: https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence (visited on 2025).

[288] International Organization for Standardization (ISO) / International Electrotechnical Commission (IEC). *Information technology – Artificial intelligence (AI) – Bias in AI systems and AI-aided decision making (ISO/IEC TR 24027:2021)*. Technical report; uploaded 3 July 2024 to OECD AI Tools Metrics catalogue. 2021. URL: https://aistandardshub.org/ai-standards/information-technology-artificial-intelligence-ai-bias-in-ai-systems-and-ai-aided-decision-making/ (visited on 2025).

[289] Rosita Guido et al. "An overview on the advancements of support vector machine models in healthcare applications: a review". In: *Information* 15.4 (2024), p. 235.

[290] Li-Fang Cheng et al. "Sparse multi-output Gaussian processes for online medical time series prediction". In: *BMC medical informatics and decision making* 20.1 (2020), p. 152.

[291] Joseph Futoma, Sanjay Hariharan, and Katherine Heller. "Learning to detect sepsis with a multitask Gaussian process RNN classifier". In: *International conference on machine learning*. PMLR. 2017, pp. 1174–1182.

[292] Elia Bruni, Gemma Boleda Tran, and Marco Baroni. "Distributional semantics from text and images". In: *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*. Association for Computational Linguistics. 2011, pp. 22–32.

[293] Douwe Kiela and Léon Bottou. "Learning image embeddings using convolutional neural networks for improved multi-modal semantics". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics. 2014, pp. 36–45.

[294] Lior Rokach. "Ensemble-based classifiers". In: *Artificial Intelligence Review* 33.1 (2010), pp. 1–39.

[295] Ting Chen et al. "A simple framework for contrastive learning of visual representations". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 1597–1607.

[296] Alec Radford et al. "Learning Transferable Visual Models From Natural Language Supervision". In: *Proceedings of the 38th International Conference on Machine Learning (ICML)*. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 8748–8763. URL: https://proceedings.mlr.press/v139/radford21a.html (visited on 2025).

[297] Lars Heiliger et al. "Beyond medical imaging-a review of multimodal deep learning in radiology". In: *TechRxiv* 19103432 (2022).

[298] Tao Zhang, Ying Wu, and Yaqiong Liu. "A Text Classification System for Chinese Medical Device Registration Based on BERT and Rule Fusion". In: *IEEE Access* 9 (2021), pp. 144758–144768.

[299] Thomas Hartung et al. "Is regulatory science ready for artificial intelligence?" In: *npj Digital Medicine* 8.1 (2025), p. 200.

[300] Yen-Chun Chen et al. “UNITER: Universal image-text representation learning”. In: *European conference on computer vision*. Springer. 2020, pp. 104–120.

[301] Feifei Liu et al. “Automatic knowledge extraction from Chinese electronic medical records and rheumatoid arthritis knowledge graph construction”. In: *Quantitative imaging in medicine and surgery* 13.6 (2023), p. 3873.

[302] Sun Junyi. *Jieba: Chinese text segmentation*. `https://github.com/fxsjy/jieba`. Accessed: 2025-04-30. 2012.

[303] Jason Wei and Kai Zou. “EDA: Easy data augmentation techniques for boosting performance on text classification tasks”. In: *arXiv preprint arXiv:1901.11196* (2019).

[304] Sosuke Kobayashi. “Contextual augmentation: Data augmentation by words with paradigmatic relations”. In: *arXiv preprint arXiv:1805.06201* (2018).

[305] Rico Sennrich, Barry Haddow, and Alexandra Birch. “Improving neural machine translation models with monolingual data”. In: *arXiv preprint arXiv:1511.06709* (2015).

[306] Gerard Salton and Christopher Buckley. “Term-weighting approaches in automatic text retrieval”. In: *Information Processing & Management* 24.5 (1988), pp. 513–523.

[307] Scott Deerwester et al. “Indexing by Latent Semantic Analysis”. In: *Journal of the American Society for Information Science* 41.6 (1990), pp. 391–407.

[308] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2019, pp. 4171–4186. URL: `https://aclanthology.org/N19-1423` (visited on 2025).

[309] Nils Reimers and Iryna Gurevych. “Sentence-BERT: Sentence embeddings using Siamese BERT-networks”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. 2019, pp. 3982–3992.

[310] Fabian Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[311] Mingxing Tan and Quoc Le. “EfficientNet: Rethinking model scaling for convolutional neural networks”. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 6105–6114.

[312] Adam Paszke et al. “PyTorch: An imperative style, high-performance deep learning library”. In: *Advances in neural information processing systems* 32 (2019).

[313] Diederik P Kingma and Jimmy Ba. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014).

[314] David Yarowsky. “Unsupervised word sense disambiguation rivaling supervised methods”. In: *33rd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. 1995, pp. 189–196.

[315] Ramprasaath R Selvaraju et al. “Grad-cam: Visual explanations from deep networks via gradient-based localization”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 618–626.

[316] Jiquan Ngiam et al. “Multimodal deep learning.” In: *ICML*. Vol. 11. 2011, pp. 689–696.

[317] Faraat Ali, Anam Ilyas, and Shaima Ahmadeen. “Regulations in Japan”. In: *Global Regulations of Medicinal, Pharmaceutical, and Food Products*. CRC Press, 2024, pp. 153–167.

[318] Ninareh Mehrabi et al. “A survey on bias and fairness in machine learning”. In: *ACM computing surveys (CSUR)* 54.6 (2021), pp. 1–35.

[319] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “" Why should i trust you?" Explaining the predictions of any classifier”. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144.

[320] Finale Doshi-Velez and Been Kim. “Towards a rigorous science of interpretable machine learning”. In: *arXiv preprint arXiv:1702.08608* (2017).

[321] Nathalie A Smuha. “Regulation 2024/1689 of the Eur. Parl. & Council of June 13, 2024 (EU Artificial Intelligence Act)”. In: *International Legal Materials* (2025), pp. 1–148.

[322] Vasiliki Rahimzadeh. “US regulation of medical artificial intelligence and machine learning (AI/ML) research and development”. In: *Research Handbook on Health, AI and the Law* (2024), pp. 273–290.

[323] Meseret N Teferra. “ISO 14971-medical device risk management standard”. In: *International Journal of Latest Research in Engineering and Technology (IJLRET)* 3.3 (2017), pp. 83–87.

[324] Peter Jordan. “Standard IEC 62304-Medical device software-Software lifecycle processes”. In: *2006 IET Seminar on Software for Medical devices*. IET. 2006, pp. 41–47.

[325] Alessandro Simonetta and Maria Cristina Paoletti. “ISO/IEC Standards and Design of an Artificial Intelligence System”. In: *CEUR Workshop Proceedings*. 2024.

[326] Juan Ignacio Torres, Ariel Cristian Pasini, and Patricia Mabel Pesado. “Ensuring Quality in the OECD AI Lifecycle Through ISO/IEC Standards”. In: *XIII Jornadas de Cloud Computing, Big Data & Emerging Topics (La Plata, 24 al 26 de junio de 2025)*. 2025.

[327] Qizhe Xie et al. “Self-training with noisy student improves ImageNet classification”. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 10687–10698.

[328] Kihyuk Sohn et al. “FixMatch: Simplifying semi-supervised learning with consistency and confidence”. In: *Advances in Neural Information Processing Systems (NeurIPS)* 33 (2020).

[329] Antti Tarvainen and Harri Valpola. “Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results”. In: *Advances in Neural Information Processing Systems (NeurIPS)* 30 (2017).

[330] Junnan Li et al. “BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models”. In: *International Conference on Machine Learning (ICML)*. 2023.

[331] Jean-Baptiste Alayrac et al. “Flamingo: A visual language model for few-shot learning”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2022.

[332] An Yang et al. “Chinese CLIP: Contrastive vision-language pretraining in Chinese”. In: *arXiv preprint arXiv:2211.01335* (2022).

[333] U.S. Food and Drug Administration. *Clinical decision support software: Guidance for industry and Food and Drug Administration staff*. Tech. rep. U.S. Department of Health and Human Services, 2022.

[334] I. Y. Chen et al. "Ethical machine learning in healthcare". In: *Annual Review of Biomedical Data Science* 4 (2021), pp. 123–144.

[335] D. W. Shimabukuro et al. "Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial". In: *JAMIA Open* 24.1 (2017), pp. 3–10.

[336] Chao Jia et al. "Scaling up visual and vision-language representation learning with noisy text supervision". In: *International Conference on Machine Learning (ICML)*. Vol. 139. 2021, pp. 4904–4916.