Algorithmic Fairness and Bias Mitigation in Clinical Machine Learning for Equitable Patient Outcomes

Jenny Yang

Balliol College
University of Oxford

This thesis is submitted for the degree of Doctor of Philosophy

Michaelmas 2024

Abstract

In recent years, the integration of machine learning algorithms into clinical settings has shown immense potential for improving healthcare outcomes. However, concerns regarding fairness and equity in machine learning models have garnered increasing attention, particularly in healthcare where biased algorithms can perpetuate existing disparities. This thesis investigates the role of fairness-aware algorithms in addressing these issues within clinical machine learning applications. Through case studies and empirical analyses, this research explores how biases manifest and impact model performance across diverse patient populations, highlighting the challenges and opportunities in promoting fairness within clinical machine learning. Subsequently, drawing on datasets from multiple healthcare institutions, we propose and assess the effectiveness of fairness-aware techniques in advancing equitable healthcare outcomes. Ultimately, this thesis contributes to the ongoing dialogue on fairness in machine learning, providing insights and recommendations for the development of ethically sound and socially responsible machine learning algorithms in healthcare.

Algorithmic Fairness and Bias Mitigation in Clinical Machine Learning for Equitable Patient Outcomes

Jenny Yang

Balliol College

University of Oxford

This thesis is submitted for the degree of

Doctor of Philosophy

Michaelmas 2024

This thesis is dedicated to my exceptional parents,
Hui Zhang & **Zhi Yong Yang**,
who have made so many sacrifices, and have nurtured in me
a love for learning, personal growth, and compassion.

Acknowledgements

I believe wholeheartedly that we stand on the shoulders of giants, and my DPhil journey is a reflection of the exceptional mentorship, support, and relationships I've been fortunate to have along the way. It is a privilege to express my heartfelt gratitude to some of the individuals and organizations who have made this journey possible.

First, I am deeply grateful to my supervisor, Dr. David Clifton, for his invaluable mentorship. His guidance has allowed me to explore and contribute to the fascinating field of artificial intelligence in healthcare. He has provided countless opportunities for me to grow, learn, and make meaningful contributions to this domain. His encouragement has continually pushed me to critically engage with the overarching questions and themes of my research and to rise to the rigorous standards of a DPhil. Beyond academics, his lessons in diplomacy and professionalism will inspire me throughout my career. I am committed to carrying forward his example and ensuring that the ladder remains accessible for those who follow.

I am thankful for the financial support I received during my studies from the European Union's Horizon 2020 Research and Innovation funding programme (Marie Skłodowska-Curie Grant No. 955681, "MOIRA"), the Canadian Centennial Scholarship Fund, and additional travel funding from the Department of Engineering Sciences and Balliol College. These contributions were instrumental in enabling my research and academic pursuits.

I also want to express my gratitude to the members of the CHI lab and collaborators, who have been my companions throughout this journey. In particular, I am thankful to Dr. Louise Thwaites, Dr. Lei Clifton, and Dr. Andrew Soltan, who often made time to discuss my work and significantly contributed to the development of my skills.

Beginning my DPhil at the height of the COVID-19 pandemic, I am especially grateful to all healthcare workers for their unwavering dedication, courage, and compassion during this unprecedented time. As this thesis focuses on AI developments centered around a COVID-19 case study, I extend my deepest thanks to the patients and staff from the participating NHS trusts (Oxford University Hospitals NHS Foundation Trust, University Hospitals Birmingham NHS Trust, Bedfordshire Hospitals NHS Foundation Trust, and Portsmouth Hospitals University NHS Trust) as well as the Hospital for Tropical Diseases and the National Hospital for Tropical Diseases in Vietnam.

My journey would not have been the same without the incredible friends who have supported me along the way. Moving across the pond, I've been blessed to meet amazing people whose friendships I will treasure for life.

Thank you to Alice Watson, for her endless support, kindness, and humor, which have been a constant source of strength and joy. Thank you to Felix Fu, for always looking out for me and being a wonderful friend on this doctoral journey and our European adventures. Thank you to Caroline Cuoco, for her positivity, fun energy, and incredible ability to make every moment enjoyable. Her knack for planning and discovering new experiences has made our time in the UK truly special. Finally, thank you to Sumali Bajaj and Raghav Khaund for their warmth and kindness. Their optimism and generosity of spirit are truly inspiring.

A special thank you goes to Lucy Shen and Susie Chen. Despite living in different places since our undergraduate days, we have always found time to support one another. This year marks the 10th anniversary of our friendships, and I deeply cherish the bond we have shared and continue to grow together.

Lastly, I owe everything to two people who have supported me ceaselessly for nearly three decades—my Mom and Dad. Words will never be enough to express my gratitude for their love, sacrifices, and encouragement throughout my life.

To everyone mentioned here, and those I may have unintentionally missed, my sincerest thanks. I don't for one second take for granted the support, guidance, and friendships that have shaped this incredible journey.

Abstract

In recent years, the integration of machine learning algorithms into clinical settings has shown immense potential for improving healthcare outcomes. However, concerns regarding fairness and equity in machine learning models have garnered increasing attention, particularly in healthcare where biased algorithms can perpetuate existing disparities. This thesis investigates the role of fairness-aware algorithms in addressing these issues within clinical machine learning applications. Through case studies and empirical analyses, this research explores how biases manifest and impact model performance across diverse patient populations, highlighting the challenges and opportunities in promoting fairness within clinical machine learning. Subsequently, drawing on datasets from multiple healthcare institutions, we propose and assess the effectiveness of fairness-aware techniques in advancing equitable healthcare outcomes. Ultimately, this thesis contributes to the ongoing dialogue on fairness in machine learning, providing insights and recommendations for the development of ethically sound and socially responsible machine learning algorithms in healthcare.

Lay Summary

In recent years, using machine learning in healthcare has become more common and has the potential to improve patient care. However, there are concerns about whether these algorithms are fair and treat everyone equally, especially since biased algorithms can worsen existing inequalities in healthcare. This thesis examines how we can make sure these algorithms are fair, focusing on clinical applications. By studying real-life examples and analyzing data, we explore how biases affect the accuracy of these algorithms for different groups of patients. We also propose ways to make these algorithms more fair and test them using data from various healthcare providers. Ultimately, this research aims to make sure that machine learning in healthcare is fair and helps everyone equally, providing recommendations for making algorithms that are not only accurate but also ethical and just.

Preface

All research presented in this thesis was carried out at the Institute of Biomedical Engineering, within the Department of Engineering Science, at the University of Oxford. The investigations were exclusively conducted using de-identified data obtained retrospectively. Consequently, explicit patient consent for data utilization and publication was not required, as covered by subsequent institutional approvals. All requisite consent has been acquired, and the corresponding institutional forms have been archived. Detailed declarations are provided alongside each case study.

These projects were conceived under the supervision of Professor David Clifton. I was the primary researcher, leading the study design, code development, data analysis, interpretation of findings, and manuscript preparation.

Contents

Appendices

Relevant Publications

Soltan, A. A., **Yang, J.**, Pattanshetty, R., Novak, A., Yang, Y., Rohanian, O., ... & Muthusami, V. (2022). Real-world evaluation of rapid and laboratory-free COVID-19 triage for emergency care: external validation and pilot deployment of artificial intelligence driven screening. The Lancet Digital Health, 4(4), e266-e278.

Yang, J., Soltan, A. A., & Clifton, D. A. Machine learning generalizability across healthcare settings: insights from multi-site COVID-19 screening. npj Digital Medicine, 5(1), 69. (2022).

Yang, J., Soltan, A. A., Eyre, D. W., Yang, Y., & Clifton, D. A. An adversarial training framework for mitigating algorithmic biases in clinical machine learning. npj Digital Medicine, 6(1), 55. (2023).

Yang, J., Soltan, A. A., Eyre, D. W., & Clifton, D. A. Algorithmic fairness and bias mitigation for clinical machine learning with deep reinforcement learning. Nature Machine Intelligence, 1-11. (2023).

Yang, J., El-Bouri, R., O'Donoghue, O., Lachapelle, A. S., Soltan, A. A., Eyre, D. W., ... & Clifton, D. A. (2023). Deep reinforcement learning for multi-class imbalanced training: applications in healthcare. Machine Learning, 1-20.

Yang, J., Triendl, H., Soltan, A. A. S., Prakash, M., & Clifton, D. A. (2024). Addressing label noise for electronic health records: insights from computer vision for tabular data. BMC medical informatics and decision making, 24(1), 183.

Yang, J., Dung, N. T., Thach, P. N., Phong, N. T., Phu, V. D., Phu, K. D., ... & Clifton, D. A. (2024). Generalizability assessment of AI models across hospitals in a low-middle and high income country. Nature Communications, 15(1), 8270.

Yang, J., Clifton, L., Dung, N. T., Phong, N. T., Yen, L. M., Thy, D. B. X., ... & Clifton, D. A. (2024). Mitigating machine learning bias between high income and low–middle income countries for enhanced model fairness and generalizability. Scientific Reports, 14(1), 13318.

1 Introduction

1.1 Motivation

Advancements in computational capabilities and the abundance of digital health data are fundamentally altering the landscape of patient health assessment and healthcare delivery. This paradigm shift is largely driven by the emergence of clinical artificial intelligence (AI), which encompasses cutting-edge technologies such as machine learning (ML). While ML-based tools offer numerous benefits, it is imperative to prioritize the equity of these innovations, ensuring that healthcare outcomes remain unbiased and fair for all individuals. This imperative underscores the urgent need for the development of fairness-aware ML algorithms.

In the realm of healthcare, where decisions carry profound implications, addressing biases within algorithms is paramount to safeguard against disparities in patient treatment. By integrating fairness-aware techniques into ML algorithms, we aim to mitigate potential biases in the data that could disproportionately impact specific demographic groups. This enables the development of a healthcare system that is more just, transparent, and considerate of the diverse needs and backgrounds of patient populations.

This commitment to fairness in ML aligns with broader societal aspirations of promoting inclusive and ethical practices across critical healthcare settings.

Emphasizing fairness in both the development and deployment of machine learning-driven technologies allows us to harness their transformative potential while striving to cultivate a healthcare landscape that embraces principles of equity and respect for all individuals. In doing so, we contribute to the collective goal of achieving optimal health outcomes for everyone.

1.2 Background

1.2.1 The Rise of Artificial Intelligence

According to our current understanding, the inception of the concept of "computer intelligence" can be traced back to a lecture delivered by Alan Turing in 1947 in London [1]. In this lecture, Turing articulated his vision for a machine capable of learning from experience, emphasizing the importance of a machine's ability to modify its own instructions as the fundamental mechanism for achieving this goal [1]. This early conceptualization laid the foundation for the field of artificial intelligence as we know it today.

In 1956, John McCarthy provided a formal definition of the term "artificial intelligence" as "the science and engineering of making intelligent machines" [1–4]. This seminal definition encapsulated the overarching goal of AI research: to develop machines capable of exhibiting intelligent behavior akin to that of humans.

During its nascent stages, AI primarily relied on basic "if, then" rules and focused on creating machines with the capacity for inference and decision-making, mirroring human cognitive processes [2, 5]. However, as technological advancements accelerated, AI evolved significantly, incorporating more sophisticated algorithms and branching into specialized subfields such as machine learning.

Machine learning emerged as a pivotal area of AI research, emphasizing the development of algorithms that enable computers to learn from data and improve their performance over time without explicit programming [6]. This shift towards data-driven approaches marked a turning point in AI research, unlocking new capabilities and applications across various domains.

Overall, the evolution of AI from its early conceptualizations by Turing to the establishment of formal definitions by McCarthy, and its subsequent progression into machine learning across diverse domains, highlights the dynamic and interdisciplinary nature of the field. As AI continues to advance, it holds immense potential to revolutionize industries, transform society, and address many of the most complex challenges facing humanity.

1.2.2 Clinical Machine Learning

Machine learning represents a crucial facet of artificial intelligence, focusing on the development of algorithms and statistical models capable of learning from, and adapting to, data [2, 6–9]. In essence, machine learning techniques enable computers to analyze large datasets, identify patterns, and make predictions or decisions without explicit programming instructions.

In the healthcare context, machine learning holds immense potential to revolutionize patient care by leveraging the vast amounts of data generated by individual patients and aggregated from diverse patient populations [10]. By harnessing this wealth of information, machine learning algorithms can extract valuable insights, uncover hidden correlations, and generate predictive models that facilitate more accurate diagnoses, personalized treatment plans, and improved patient outcomes.

The origins of biomedical AI can be traced to the 1950s, during which early efforts were made to develop computational models that mimicked the problem-solving approaches and expertise of biomedical scientists [11]. These involved the creation of systems for processing clinical data, interpreting information, and refining models for clinical decision-making. The advent of digital resources such as the Medical Literature Analysis and Retrieval System as well as PubMed further catalyzed these efforts [2, 11]. The initial phase of AI in medical research culminated in the mid-1970s with endeavors like the Stanford University Medical Experimental–Artificial Intelligence in Medicine laboratory (which provided advanced computing capabilities) and the inaugural National Institutes of Health–sponsored "AI in Medicine" workshop held at Rutgers University [2, 12]. However, the adoption of AI in the medical

sphere was slow, owing to constraints such as the high costs associated with building, sustaining, and updating expert digital information databases, alongside ensuring reliable performance [2, 4, 11].

Despite persistent exploration of AI in medicine between the 1970s and 1990s [2, 3, 11, 13, 14], clinical AI did not gain significant momentum until the late 1990s. This era, characterized by technological breakthroughs, set the stage of the contemporary era of clinical AI.

During the late 1990s and early 2000s, there was remarkable progress and expansion of clinical artificial intelligence [2, 4]. A crucial aspect of this advancement was the development of more sophisticated machine learning algorithms and computational models, with the emergence of deep learning marking a significant milestone [15]. Another significant driver of this transformative period was the increasing accessibility of extensive datasets for AI systems to leverage. With access to diverse and comprehensive datasets, AI systems can extract insights and patterns, facilitating better-informed clinical decisions and ultimately contributing to improved patient care outcomes. These more advanced algorithms, coupled with the increased digitization of clinical data and significant advancements in computational resources led to notable improvements in diagnostic capabilities and treatment recommendations within clinical AI applications [15].

Since then, clinical AI has continued to attract increasing attention across a spectrum of medical specialties, spanning oncology, cardiology, neurology, and beyond [3, 8, 9].

Types of Clinical Data

The utilization of clinical AI spans various data modalities, including imaging, textual data, omics, and others, each presenting distinct avenues for enhancing patient care and clinical outcomes.

Medical imaging plays a crucial role in modern healthcare, offering insights into disease diagnosis, treatment monitoring, and therapeutic guidance across a spectrum of medical conditions [4, 16–18]. Modalities such as X-rays, CT scans, MRI scans,

and ultrasound images constitute key pillars of medical imaging, providing clinicians with detailed anatomical and pathological information.

The integration of artificial intelligence into medical imaging has revolutionized the field, enabling medical professionals to harness the power of advanced algorithms to extract meaningful insights from imaging data [4, 16–18]. AI algorithms are adept at identifying subtle patterns, anomalies, and features within imaging data that may not be readily apparent to the human eye. This capability enhances diagnostic accuracy by aiding in the detection and characterization of abnormalities, such as tumors, fractures, or lesions [19, 20].

Electronic health records play a vital role as centralized repositories housing a wealth of patient information, including detailed medical histories, diagnoses, prescribed medications, laboratory results, and treatment plans [21–23]. These comprehensive records offer clinicians valuable insights into a patient's health status and medical journey over time. By harnessing AI algorithms, electronic health record data can be subjected to meticulous analysis to uncover intricate patterns and correlations that may not be readily apparent through manual review alone. These algorithms can sift through vast quantities of electronic health record data to identify trends, predict patient outcomes, and provide decision support to healthcare providers.

Omics data, encompassing genomics, transcriptomics, proteomics, and metabolomics, represents a rich source of biological information that offers insights into disease mechanisms, treatment responses, and personalized medicine [24–27]. This multidimensional data provides a comprehensive view of the molecular landscape underlying various physiological and pathological processes.

Through the integration of AI-driven analytics, omics data can be utilized to tailor treatment approaches to individual patients, taking into account their unique genetic makeup, molecular profiles, and disease characteristics [24–27]. This personalized approach to medicine holds the potential to optimize therapeutic efficacy, minimize adverse effects, and improve patient outcomes by aligning treatments with the specific molecular drivers of disease [28, 29]. Furthermore,

AI-driven analysis of omics data can contribute to the discovery of novel biomarkers and therapeutic targets, facilitating the development of innovative treatments and precision medicine interventions [30]. By elucidating the complex interplay between genes, proteins, metabolites, and disease pathways, AI enables a deeper understanding of disease pathogenesis and enables more targeted and effective interventions [31].

In addition to structured data sources like electronic health records and omics datasets, unstructured textual data such as clinical notes and physician reports contain insights into patient symptoms, diagnoses, and treatment plans [32–34]. These narrative accounts provide a nuanced understanding of patient conditions, capturing additional details that may not be captured in structured data formats. Through the application of natural language processing (NLP) and text mining techniques, meaningful information can be extracted from unstructured clinical text [35–37], as these AI algorithms can parse through vast amounts of textual data, identifying relevant clinical concepts, relationships, and patterns. Thus, by leveraging NLP and text mining, healthcare organizations can streamline documentation processes, reducing the burden on clinicians and improving workflow efficiency [36, 38]. Automated summarization and categorization of clinical notes can also enhance the accessibility and usability of patient information, facilitating more informed decision-making at the point of care.

Furthermore, the ongoing production of physiological data through medical sensors, ranging from wearable devices to monitors and implantable devices, offers avenues for real-time health surveillance [39–42]. The continuous influx of data provides a comprehensive picture of an individual's health status, allowing for proactive monitoring and timely interventions. By using AI algorithms, this wealth of physiological data can be analyzed in real time to identify deviations from normal patterns and detect potential health issues. Parameters such as heart rate, blood pressure, and glucose levels, among others, can be monitored to flag anomalies that may indicate emerging health concerns [43–45].

Recent Advancements in Clinical Machine Learning

In recent years, artificial intelligence has emerged as a powerful tool in healthcare, demonstrating notable benefits in enhancing the accuracy, consistency, and efficiency of reporting [2, 4]. Numerous studies have directly compared AI to medical professionals, showcasing the superior performance of AI algorithms across diagnostic imaging, electronic health records, and clinical decision support. Highlighting representative work across these domains illustrates both the breadth and depth of AI's clinical impact.

Diagnostic imaging has been one of the most visible areas of progress. For instance, a study conducted in 2020 in the UK and the USA introduced an AI system capable of outperforming human experts in predicting breast cancer using mammograms [46]. The study revealed a significant reduction in false positives by 5.7% and 1.2% (USA and UK, respectively), as well as reductions in false negatives by 9.4% and 2.7% (USA and UK, respectively) compared to human experts. Additionally, through an independent study involving six radiologists, they determined that the AI system surpassed all human readers, achieving an area under the receiver operating characteristic curve (AUROC) more than 11.5% higher than the average radiologist. Similarly, another study in 2020, utilizing data from institutions in the USA, UK, and South Korea, found that an AI algorithm trained on mammography examinations exhibited significantly better performance than 14 radiologists [47]. The AI algorithm demonstrated higher sensitivity in diagnosing breast cancer with mass (90% vs. 78% compared to radiologists) or architectural distortion/asymmetry (90% vs. 50% compared to radiologists). Additionally, AI proved more adept at detecting early breast cancer (91% vs. 74% compared to radiologists).

Beyond imaging, electronic health records have also served as a rich substrate for clinical ML applications. For example, research has indicated the effectiveness of ML algorithms utilizing electronic health record data in predicting antibiotic resistance in urinary tract infection cases. According to findings, these algorithms improved the predictability of resistance, thereby reducing the incidence of mismatched

treatments when utilizing drug recommendations suggested by the algorithm [48]. Another study employed electronic health record data to forecast the likelihood of antibiotic resistance in uncomplicated UTIs [49]. The models were found to reduce inappropriate antibiotic therapy by 18% relative to clinicians and corrected 47% of inappropriate first-line drugs chosen by a clinician. The approach also led to a 67% reduction in the recommendation of second-line drugs compared to clinicians.

Recent work has also demonstrated the role of AI in decision support across diverse clinical specialties. For example, in 2024, researchers presented findings from a large teledermatology experiment focused on diagnosing skin diseases [50]. In this experiment, dermatologists and primary-care physicians achieved diagnostic accuracies of 38% and 19%, respectively, when tasked with diagnosing 46 skin diseases from 364 images. In contrast, the AI decision support system significantly improved the diagnostic accuracy of both specialists and generalists by more than 33%.

Taken together, these instances illustrate how clinical ML has advanced along multiple fronts: from imaging to EHR-based predictions to decision support across specialties. By harnessing AI algorithms, healthcare professionals can leverage the wealth of physiological data available to them to detect anomalies, diagnose conditions, and provide timely interventions, ultimately enhancing patient care and health outcomes.

The evolution of AI since Turing's pioneering vision has been nothing short of remarkable, especially within the healthcare domain. The integration of machine learning into healthcare has been motivated by its demonstrated efficacy and the promise of notable progress across diverse domains, ushering in a transformative era marked by personalized and precision medicine. However, amidst the enthusiasm and aspirations to achieve this objective, it is vital to recognize the challenges and considerations inherent in the adoption of machine learning in healthcare. These challenges encompass a spectrum of issues, including data privacy concerns, algorithm bias, interpretability challenges, and ethical implications [51–54]. Addressing these complexities demands a multidisciplinary approach, necessitating

collaboration among clinicians, data scientists, ethicists, policymakers, and other stakeholders. This collaborative effort is vital to ensure the responsible and ethical utilization of machine learning technologies in healthcare [4].

1.2.3 Sources of Bias in Clinical Data

The focal point of this thesis lies in addressing algorithmic bias, a critical concern within the realm of healthcare. The recognition of bias in clinical data, and consequently, in the machine learning models trained on such data, has garnered increasing attention in both healthcare research and practical application. Clinical data, which encompasses a diverse array of information sourced from patients, medical records, and healthcare systems, is vulnerable to various forms of bias originating from factors such as data collection methodologies, institutional protocols, and societal disparities. These biases have the potential to significantly impact the accuracy, dependability, and equity of analyses and insights derived from clinical data, thus exerting a profound influence on healthcare decision-making processes and patient outcomes. Particularly, when developing machine learning algorithms utilizing clinical data, further layers of complexity and challenges may emerge regarding bias. These algorithms are vulnerable to inheriting biases present in the data, potentially exacerbating them inadvertently. Thus, understanding the presence of bias in clinical data and recognizing its implications for machine learning are essential stages in the development process. Through this awareness, effective strategies can be devised to mitigate bias and foster data-driven healthcare practices that are equitable and dependable for all individuals.

Bias in clinical data refers to systematic errors or imbalances that can occur during the collection, processing, or analysis of healthcare-related information, stemming from various sources.

Selection Bias

One common source of bias in clinical data is selection bias, which arises when the individuals included in a study or dataset are not representative of the broader

population [55, 56]. For example, randomized trials assess treatment effects for a trial population; however, participants in clinical trials often do not reflect the demographic diversity of the patient population that ultimately receives the treatment [57, 58]. This could be due to a variety of issues such as resource constraints, regional biases, and other factors. Consequently, if a model determines who receives a specific drug or intervention, minority groups such as ethnic minorities, women, and obese patients might be adversely affected, perpetuating demographic inequities in healthcare.

Measurement Bias

Another type of bias is measurement bias. This refers to inaccuracies or inconsistencies introduced into research or clinical data due to flawed or inconsistent methods of data collection or measurement [59]. For example, diagnostic tests may be prone to errors, such as false positives or false negatives, which can skew the data; or, healthcare providers may have different approaches or criteria for diagnosing conditions or assessing patient outcomes, leading to variability in the data collected. Moreover, documentation standards may vary among healthcare institutions, resulting in inconsistencies in the recorded information. These discrepancies in data collection methods can compromise the validity of research findings and clinical decision-making. If the data are not accurately collected or measured, the conclusions drawn from them may not reflect the true relationships or effects being studied.

Confounding Bias

Measurement bias can also interact with confounding bias, where external factors influence both the exposure and outcome of interest, making it difficult to establish causal relationships [60, 61]. For example, socioeconomic factors like income or education level may confound the association between a specific treatment and patient outcomes. If these factors are not properly accounted for or controlled, biased estimates of treatment effects may result.

Implicit Bias

Implicit bias represents a significant source of bias in various aspects of society, including healthcare. These biases are unconscious and often stem from societal stereotypes, historical disparities, or systematic inequalities. In healthcare, implicit biases can manifest in several ways, ultimately leading to skewed or incomplete datasets and perpetuating disparities in patient care.

One prominent example of implicit bias in healthcare is gender bias. Studies have demonstrated unconscious biases among physicians regarding the presentation and diagnosis of certain conditions in women. For instance, one study highlighted how physicians may attribute symptoms of coronary heart disease to other conditions in women, potentially leading to underdiagnosis or misdiagnosis [62]. Furthermore, an observational study revealed discrepancies in the rates of referral to cardiology specialists, indicating that women were 2.5 times less likely to receive referrals for additional assessment, despite presenting with similar symptoms of chest pain when compared to men [63]. Additionally, another study has shown that physicians tended to ask fewer diagnostic questions and prescribe fewer coronary heart disease-related medications to middle-aged women [64], further contributing to gender-based disparities in healthcare delivery.

Similarly, ethnicity-based bias has been documented in healthcare settings [65]. For example, one study found disparities in the administration of pain medication in emergency rooms, with black patients receiving pain medication at a 40% lower rate compared to white patients [66]. These disparities underscore how implicit biases can influence clinical decision-making and patient care, leading to unequal treatment based on race or ethnicity. Thus, if a model intended to inform the prescription of medications for coronary heart disease primarily draws from data featuring male patients receiving medication, it may unfairly favor men, thus deepening disparities in care.

Location-based Bias

In the realm of healthcare, it is well-documented that clinical outcomes and medical practices can significantly vary not only across different geographic regions but also between hospitals within the same region. This variability encompasses a broad spectrum of factors, including disease prevalence, mortality rates, quality of healthcare services, and even the specific medical devices utilized, such as various brands of blood analysis equipment. This location bias has been acknowledged worldwide and has been examined for a range of medical conditions and diseases[67–69], as well as different drivers of healthcare quality[67, 70].

This wide-ranging heterogeneity in healthcare settings poses significant challenges, particularly in the context of data-driven approaches such as machine learning. Models trained on data from one hospital may not effectively generalize when applied to data from a different hospital. This lack of generalizability stems from the inherent biases encoded within the data collected, processed, and organized at each specific site. Moreover, a recent study has highlighted how state-of-the-art machine learning methods can systematically underdiagnose under-served patient populations [71]. These findings underscore the critical need for addressing site-specific biases in healthcare data to ensure equitable and accurate healthcare delivery, particularly for vulnerable or marginalized populations.

Overall, the existence of biases within clinical data can carry significant consequences, especially concerning the development of machine learning models for healthcare purposes. Given that these models heavily depend on the quality and representativeness of their training data, datasets suffering from unintentional biases can lead to the creation of models that sustain and exacerbate prevailing healthcare inequalities.

1.2.4 Bias in Machine Learning Models

In the realm of machine learning research, it is widely recognized that models can develop biases based on the specific samples utilized during the training process. This phenomenon can result in decreased predictive accuracy and the

potential for unfair decision-making. In this context, bias refers to variations in performance across different subgroups when performing predictive tasks [71]. Similarly, an unfair decision is defined as any outcome that exhibits a bias towards a particular group or population [72, 73]. For example, machine learning models have previously demonstrated susceptibility to demographic biases. One study found that a recidivism prediction model exhibited bias against black defendants, incorrectly classifying them as future criminals at nearly double the rate of white defendants [74]. Similarly, another study found that a ML model used for predicting juvenile recidivism tended to discriminate against male defendants, foreigners, or people of specific national groups [75].

Formally, consider a scenario where a classifier Y is trained to predict labels y_i based on features x_i for various samples i. Bias occurs when there is a noticeable difference in the statistical properties of the distribution of predicted labels $\{y_i, i \in Z\}$ compared to $\{y_i, i \in Z'\}$, where Z represents a sensitive subgroup. A sensitive subgroup is a group that the model might exhibit bias against, while Z' represents its non-sensitive complement.

In simpler terms, bias arises when the model's predictions disproportionately favor or disadvantage certain groups over others. This bias can lead to unfair decision-making, where individuals from sensitive subgroups may be systematically disadvantaged or marginalized by the model's predictions. Understanding and addressing these biases are critical steps in ensuring that machine learning models produce fair and equitable outcomes across all demographic groups.

Various statistical measures have been proposed and studied in the literature to assess and mitigate biases present in predictive models. Some of these measures include demographic parity [76, 77], equality of odds and equality of opportunity [76, 78–81], statistical parity [80–83], and disparate impact [80–82, 84]. Details and formal definitions of these measures are presented in Chapter 4.2.5.

When biases are ingrained within the data used for training machine learning models and subsequently internalized during the model training process, the resulting clinical ML models may inadvertently generate unfair treatment or divergent

outcomes for specific patient demographics [57, 85]. This phenomenon has significant implications in sensitive domains like healthcare for several key reasons:

1. Inaccurate predictions for critical decisions: Biased models can lead to inaccurate predictions for crucial decisions in healthcare. These decisions may have life-altering consequences for patients, such as treatment plans, diagnoses, or interventions. If the model is biased, it may provide recommendations that are not optimal or even harmful to certain demographic groups, potentially jeopardizing patient well-being.

2. Poorer care for disadvantaged groups: Bias against a particular demographic group can result in those patients receiving suboptimal care compared to individuals from other groups. This disparity in treatment may stem from the model's tendency to prioritize or allocate resources differently based on demographic factors, rather than clinical need. Consequently, disadvantaged groups may face barriers to accessing appropriate care, leading to poorer health outcomes and exacerbating existing health disparities.

3. Exacerbation of inequities: Biased models have the potential to exacerbate and perpetuate existing inequities in healthcare and society. By reinforcing discriminatory patterns present in the training data, these models can amplify disparities in access to care, health outcomes, and overall well-being among different demographic groups. This not only undermines the principles of fairness and justice but also undermines public trust in healthcare systems and exacerbates social inequalities.

Recognizing and addressing biases during the development of machine learning tools in healthcare is crucial for fostering more equitable and reliable healthcare systems. By actively mitigating biases in both the data and the model architecture, developers can ensure that ML tools better meet the diverse needs of all patients, regardless of their demographic characteristics. This proactive approach is essential for building trust in AI-driven healthcare solutions and advancing towards a future

where healthcare is truly equitable and accessible to all. This thesis will primarily focus on mitigating location-based biases, specifically those arising from hospital differences. Additionally, it will briefly address ethnicity-based bias.

1.2.5 Towards Fairness-Aware Machine Learning

Fairness-aware machine learning is an evolving research field with the goal of mitigating biases and promoting fairness within ML models and algorithms [80, 86, 87]. The increasing integration of ML systems into various societal domains like finance, hiring, healthcare, and criminal justice has emphasized the importance of ensuring that these systems do not unintentionally perpetuate or worsen existing biases and inequalities. Fairness-aware ML addresses this concern by explicitly integrating fairness considerations into the design, development, and deployment phases of ML models. This entails creating algorithms that minimize biases and promote fairness across diverse demographic groups, ensuring that individuals or groups are not unfairly treated based on sensitive attributes such as race, gender, ethnicity, or socioeconomic status [80].

A common strategy in fairness-aware ML involves establishing fairness metrics that quantify the level of fairness or discrimination within a model's predictions [72, 80]. These metrics, which may include measures like disparate impact, demographic parity, and equal opportunity, are then integrated into the optimization process of the model. By evaluating models using these fairness metrics, developers can gauge the presence of bias or discrimination and make appropriate adjustments to enhance fairness in their ML systems.

Equity Versus Fairness

Although the terms are often used interchangeably, *equity* and *fairness* represent distinct but related concepts. In machine learning, fairness usually refers to formalized definitions—such as demographic parity, equalized odds, or predictive parity—that assess whether model predictions or errors are distributed evenly across

groups [88, 89]. These criteria provide quantifiable and interpretable measures, making them widely adopted in algorithmic fairness research.

Equity, by contrast, emphasizes justice in outcomes rather than equal treatment. An equity-oriented perspective acknowledges that different groups may begin from unequal starting points due to structural or historical disadvantages [90, 91]. In such cases, treating all groups identically may perpetuate disparities rather than reduce them. Instead, equity seeks to adjust allocations or decision thresholds in ways that address pre-existing imbalances and promote more just outcomes [88].

A salient example comes from estimated glomerular filtration rate (eGFR) calculations in nephrology. Historically, eGFR formulas incorporated a race-based adjustment for Black patients, which led to systematically higher reported kidney function estimates compared to non-Black patients with identical clinical markers. While this adjustment was intended to improve predictive "fairness" in calibration across populations, it raised equity concerns because it delayed referral of Black patients for specialist care or transplant eligibility [92, 93]. This illustrates that fairness defined as statistical parity or calibration does not always align with equity understood as redressing structural inequities in healthcare outcomes.

In this thesis, we primarily evaluate fairness as defined through group fairness metrics, given their operationalizability in machine learning (this is further introduced in Chapter 4). Nonetheless, we recognize equity as a complementary but distinct perspective that highlights the ethical implications of model deployment in clinical contexts.

Fairness-Aware Machine Learning Methods

Various techniques have been proposed within the ML literature to achieve fairness goals. This includes methods for processing data to remove bias, modifying the training process to penalize unfair outcomes, or post-processing model outputs to ensure fairness [80]. Additionally, researchers explore the trade-offs between fairness and other desirable properties of ML models, such as accuracy or computational efficiency, to develop balanced and effective solutions [86, 94].

In general, fairness-aware ML techniques can be broadly categorized into pre-processing, in-processing, and post-processing methods [80, 95]. Pre-processing techniques involve modifying the training data to remove or mitigate biases before training the model, for example through reweighting, resampling, or generating synthetic instances to balance subgroup representation [96–99]. In-processing methods integrate fairness considerations during model training by altering the optimization process itself, for example through regularization, constraints, cost-sensitive weighting, or adversarial learning[76, 77, 100–103]. Post-processing techniques involve *post-hoc* adjustments to the model's predictions to ensure fairness, typically by applying algorithms that reassign or reweight predictions to achieve fairness criteria [104–109].

These fairness-aware methods have been applied across diverse data modalities, including natural language processing (e.g., reducing gender bias in sentiment and occupation classification [110, 111]), computer vision (e.g., mitigating racial and gender bias in face recognition [112, 113]), and tabular socio-economic data such as income and recidivism prediction [101, 114, 115]. In healthcare, fairness research has highlighted disparities in electronic health record–based risk prediction [116, 117], medical imaging pipelines [118], and cross-institutional generalization, where models often underperform on underrepresented demographics or patients from specific hospitals [117, 119].

Research in this area highlights that sensitive attributes are not limited to demographics but can also include institutional or site-level factors, such as the hospital at which a patient presents. Yet, in contrast to the extensive fairness-focused ML research on demographic bias, site-related disparities remain comparatively under-explored. This thesis addresses this gap by systematically examining in-processing bias mitigation methods—also known as fairness-aware algorithms—which embed fairness considerations directly into the training process. Our investigation applies these methods to structured clinical data, with a particular focus on ensuring fairness across hospitals in the context of COVID-19 diagnosis.

Fairness-aware algorithms present significant advantages over pre-processing techniques. By directly tackling biases within the model without altering the original data, they prioritize maintaining the integrity of information, thus averting potential distortions and upholding the original dataset's integrity. In contrast, adjusting training data poses the risk of introducing new biases or altering the original information.

Furthermore, fairness-aware algorithms exhibit adaptability to evolving data dynamics. Given that training data may evolve over time, leading to shifts or emergence of biases, these algorithms can be adapted to continuously monitor and adjust for biases during model deployment, ensuring robust fairness considerations over time.

Moreover, fairness-aware algorithms are preferred over post-processing methods due to their proactive stance in addressing bias early in the model training process. By embedding fairness considerations directly into the optimization process, they can identify and mitigate bias before it becomes entrenched in the model's predictions. This integrated approach enables balancing fairness with other performance metrics (e.g., accuracy or predictive power) during model training, resulting in more effective models that are both equitable and resilient.

Additionally, fairness-aware algorithms offer efficiency benefits by streamlining the bias mitigation process. Unlike modifying training data or making *post-hoc* adjustments, which can be labor-intensive and time-consuming, fairness-aware algorithms seamlessly integrate fairness considerations into the model's design and optimization process, enhancing overall efficiency.

In summary, fairness-aware algorithms provide a principled and scalable approach to mitigating bias in machine learning models, enhancing efficiency, flexibility, and adaptability. They represent a proactive and integrated solution to addressing bias in machine learning models, thereby promoting fairness and social responsibility in AI systems. By integrating fairness considerations into the model training process, developers can create models that prioritize fairness alongside other performance metrics, contributing to equitable outcomes across various domains.

1.2.6 Limitations of Current Clinical Machine Learning Deployment

Despite the rapid growth of machine learning research in healthcare and the proliferation of proof-of-concept studies, the real-world clinical impact of these technologies remains limited [120]. To date, most work in fairness-aware ML has focused on developing theoretical frameworks, algorithmic innovations, or retrospective validations on historical datasets. While these contributions are indispensable for advancing the field, they provide little direct evidence that such systems improve patient outcomes or reduce healthcare costs when deployed in practice.

Some prospective or real-world implementation studies that do exist have also yielded disappointing results. A prominent example is the widely deployed EPIC sepsis prediction tool, which was designed to provide early warnings of sepsis onset in hospitalized patients. Despite its broad clinical rollout across multiple healthcare systems, independent evaluations revealed that the model had poor sensitivity and generated a high number of false alarms, leading to limited clinical utility and increased burden on healthcare staff [121]. Such cases highlight the critical gap between algorithmic performance in retrospective validation and tangible benefits in real-world clinical environments.

This gap underscores the importance of not only developing fairness-aware and high-performing models, but also rigorously evaluating their clinical value and utility in deployment. Demonstrating improvements in patient outcomes, clinician decision-making, or health system efficiency remains an essential and underexplored frontier in clinical machine learning. The present thesis responds to this challenge by focusing on fairness in clinical ML through the lens of location-based bias, with an emphasis on approaches that move the field closer to models that are not only theoretically fair but also practically useful and equitable in real-world healthcare contexts.

1.3 Research Aims and Thesis Overview

This thesis is dedicated to exploring the origins and ramifications of data-level biases within healthcare contexts. It aims to illustrate how these inadvertent

biases can emerge within clinical data and, consequently, impact the fairness of outcomes produced by machine learning models trained on such datasets. Moreover, it underscores the pivotal role and effectiveness of fairness-aware algorithms in fostering the development of more equitable models, emphasizing their capacity to mitigate biases from influencing model decisions. By focusing on biases associated with geographic disparities—particularly differences among hospital sites—this thesis provides frameworks for analyzing bias effects and devising robust bias mitigation strategies.

The rationale for focusing on hospital site differences stems from a clear recognition of the variability among healthcare institutions, and the ability to identify and categorize different types of biases that may arise. Thus, this approach allows for a nuanced exploration of how location-related disparities can influence model training and performance, providing insights into the ways biases manifest in clinical settings.

Moreover, the pursuit of fairness in machine learning outcomes across diverse geographic locations is crucial for achieving the broader objective of model generalizability. Generalizability remains a paramount goal among machine learning practitioners, as it ensures that models are not only effective in narrow or specific contexts but also across a wide range of scenarios and populations. This dissertation thus aims to bridge the gap between theoretical fairness (i.e. formal, mathematical definitions of fairness) and practical application in clinical machine learning (i.e. how fairness is actually achieved or interpreted in clinical environments), advocating for the development and deployment of algorithms that are not only high-performing but also equitable and inclusive across different hospital environments. In doing so, it seeks to contribute to the advancement of machine learning practices that are ethically grounded and socially responsible, capable of serving diverse patient populations equitably.

I hypothesized that biases stemming from disparities and imbalances among different hospital sites might be apparent in the data and subsequently perpetuated through machine learning models trained on such data, resulting in unfair decision-making. Additionally, I posited that addressing these biases during model training

could lead to outcomes that exhibit greater equity when applied across diverse sites. The implication of this hypothesis suggests that proactive measures and fairness-aware techniques, implemented during the model development phase, can effectively mitigate biases and promote equitable decision-making. Validating these hypotheses would not only advance our comprehension of the impact of data-level biases in clinical machine learning but also provide practical strategies for developing more equitable and inclusive machine learning models in healthcare.

Chapter 2 delves into a COVID-19 screening case study, which serves as a focal point throughout this thesis. Specifically, I elaborate on a multicenter validation study conducted in a UK emergency department, where an AI-based screening model was employed. This study demonstrates the practical application of an AI algorithm, and will be used to assess various machine learning approaches, including fairness-aware algorithms.

Chapter 3 outlines the particular datasets and processing methodologies employed in the development and assessment of the COVID-19 screening models within this thesis. Additionally, it examines potential biases inherent in the data that could influence the results of machine learning models trained on said data.

Chapter 4 tests the effectiveness of advanced fairness-aware algorithms in reducing biases, particularly in the context of supervised model training. It underscores their effectiveness in generating machine learning models that are more equitable.

Following this, Chapter 5 presents an innovative bias mitigation approach within reinforcement learning, marking a novel contribution to bias reduction in a distinct machine learning paradigm.

In Chapter 6, the scope of our investigation expands to include applications of fairness-aware algorithms in both UK hospital groups and two Vietnamese hospitals, illustrating the potential of these algorithms to improve fairness across different socioeconomic conditions.

Chapter 7 provides additional validation of the effectiveness of fairness-aware techniques through two supplementary case studies. These studies focus on

mitigating ethnicity bias in COVID-19 screening and predicting patient discharge status, showcasing the adaptability and efficacy of these methods in promoting fairness across diverse applications.

Chapter 8 concludes the thesis by summarizing key findings, discussing the limitations of the conducted studies, and suggesting directions for future research in promoting fairness in machine learning within healthcare and beyond.

This body of work not only addresses the urgent challenge of biases originating from data in machine learning, but also provides insights and resources for fostering fairness in the ever-evolving domain of clinical machine learning.

2

Patient Triage During the COVID-19 Pandemic

2.1 Introduction

2.1.1 Overview

This chapter delves into the central case study of this thesis: the diagnosis of COVID-19. It sets the stage by exploring the challenges and nuances of patient management during the COVID-19 pandemic, focusing on the critical role of diagnostic testing for the virus, and highlighting the urgent need for innovative strategies to bolster the resilience and preparedness of healthcare systems against current and future pandemics.

The discussion focuses on the healthcare infrastructure in England, with a particular emphasis on the NHS, which represents the cornerstone of the publicly funded healthcare system in England. The NHS's guidelines and protocols during the pandemic have been instrumental in shaping patient care pathways in emergency departments across the country. These protocols cover a broad spectrum, from the management of patients suspected or confirmed to have COVID-19 to the care of individuals seeking medical attention for other health issues.

Through an exploration of these admission processes and the broader strategies implemented by the NHS amid the pandemic, this chapter aims to provide a

thorough overview of the operational hurdles and response mechanisms during that period. This context is pivotal for understanding the driving force behind the pursuit of innovative diagnostic and management strategies for COVID-19, notably in the realm of AI-driven triage.

2.1.2 The COVID-19 Pandemic

The COVID-19 pandemic, triggered by the emergence of the novel coronavirus SARS-CoV-2, emerged as a pivotal global health emergency, exerting a profound and multifaceted impact on societies, economies, and healthcare infrastructures across the globe [122–124]. The pandemic's dynamic and unpredictable nature underscored the urgent need for AI-driven tools to enhance response efforts. Machine learning and AI have shown immense potential in tackling key challenges, including improving diagnostic accuracy, optimizing resource allocation, and predicting disease progression. Their ability to process vast datasets, uncover patterns, and provide actionable insights in real time is invaluable in managing the rapid spread and clinical variability of such a global crisis.

The scientific consensus holds that the virus was first identified in December 2019 in Wuhan, a city in China's Hubei province. Rapidly spreading beyond geographical borders, this virulent pathogen prompted the World Health Organization (WHO) to declare a Public Health Emergency of International Concern by January 2020, and later, in March 2020, to classify it as a pandemic [125].

The pandemic is distinguished by its rapid spread and a wide spectrum of clinical outcomes, from individuals remaining asymptomatic to cases of severe and critical respiratory ailments [126, 127]. This variability has significantly complicated efforts to diagnose, treat, and contain the virus. Moreover, the pandemic placed unparalleled strain on global healthcare systems, prompted widespread economic disruptions, and necessitated profound adjustments to the daily lives of billions of individuals worldwide.

In the face of these challenges, an international coalition of governments, healthcare workers, researchers, and communities was mobilized to confront the

pandemic's wide-ranging impacts. This collective effort encompassed the development and distribution of vaccines, the implementation of public health measures to control the virus's spread, and the exploration of therapeutic strategies to treat infected individuals.

Understanding the intricate dynamics of the COVID-19 pandemic is essential for devising comprehensive strategies geared towards an effective response, mitigation, and eventual recovery. The insights gained helped shape response endeavors at the time, but will also enhance global readiness for future pandemics. This underscores the vital significance of resilience, innovation, and international collaboration in addressing emerging infectious diseases.

In this thesis, the focus will center on COVID-19 screening and diagnosis, delving into specific aspects surrounding these critical areas.

2.2 COVID-19 Triage in Hospital Emergency Departments

During the COVID-19 pandemic, efficient front door triaging emerged as a crucial strategy in managing healthcare facilities. This process refers to the initial point of contact for patients seeking medical assistance, enabling the swift identification and separation of potential COVID-19 cases. Through the implementation of meticulous triage protocols, healthcare institutions managed to effectively reduce the risk of virus transmission within their premises, optimize the allocation of resources, and ensure timely care delivery to both COVID-19 and non-COVID-19 patients. This overview delves into the significance of front door triaging during the pandemic, highlighting its pivotal role in safeguarding public health, improving healthcare delivery, and contributing to broader containment efforts.

At the core of front door triaging lies the principle that upon arrival at the hospital, patients are promptly directed to a designated treatment area or service where appropriate healthcare professionals are available to address their medical needs [128]. The initial evaluation, commonly referred to as triage, is performed by trained healthcare providers, such as nurses or doctors [128]. It is noteworthy

that a significant portion of patients arriving at emergency departments do so on their own accord, without prior referral [128]. Thus, triage plays a central role in prioritizing patient treatment to ensure that those with the most critical conditions receive immediate attention. Especially during periods of heightened demand in emergency departments, triage serves as a valuable tool for clinicians in determining the sequence in which patients should be attended to. Moreover, beyond its role in managing patient flow, triage is also a routine practice aimed at facilitating the efficient delivery of healthcare services. By systematically evaluating and categorizing patients based on the severity of their conditions, triage contributes to the overall effectiveness and functionality of the emergency department. During this time, additional assessments including vital sign measurements or blood tests, may be conducted to further inform treatment decisions.

With respect to triage, the NHS offers same-day emergency care services with the objective of reducing waiting times and unnecessary hospital admissions, as appropriate [129, 130]. This care model enables patients presenting at hospital emergency departments with relevant conditions to undergo rapid assessment, diagnosis, and treatment without requiring admission to a ward. If deemed clinically safe, these patients can receive care and return home on the same day. This approach not only enhances the patient experience by ensuring swift access to care but also contributes to the reduction of hospital admissions.

The same-day emergency care model holds several advantages. Firstly, it facilitates prompt assessment, diagnosis, and initiation of treatment, thereby improving patient outcomes and satisfaction. Moreover, it minimizes the need for unplanned and prolonged hospital stays, reducing the risk of hospital-acquired infections and preventing patient deconditioning. By optimizing resource utilization and streamlining care delivery processes, this model enhances overall healthcare efficiency and effectiveness.

During the COVID-19 pandemic, hospitals implemented modifications to their patient care protocols and access procedures, primarily due to heightened infection prevention and control (IPC) measures [128, 130, 131]. These adjustments included

the rapid testing of patients for COVID-19 to promptly identify and manage cases. Additionally, NHS trusts adopted a color-coded "green–amber–blue" categorization system. Under this system, patients are categorized based on their COVID-19 status: green denotes patients with no features of COVID-19, amber indicates symptoms potentially indicative of COVID-19, and blue signifies laboratory-confirmed COVID-19 infection [132, 133]. This categorization aided healthcare providers in prioritizing and managing patient care effectively, ensuring appropriate infection control measures were implemented while delivering timely and tailored treatment.

2.3 Diagnosis of COVID-19

Upon arrival at hospital emergency departments, the diagnostic tests patients undergo are tailored to their presenting clinical symptoms, medical history, suspected or potential diagnoses, and the urgency of the situation. During the pandemic, healthcare providers often included COVID-19 diagnostic testing options in their assessment. These options included symptom-guided evaluation, rapid lateral flow antigen device tests (LFD), or real-time polymerase chain reaction (PCR) tests. These tests were crucial for confirming the presence of the virus and aiding in infection prevention and control measures. Healthcare professionals determined the most appropriate testing approach for each patient depending on the severity of symptoms, risk factors, and local guidelines at the time.

2.3.1 Clinician-Assessed Symptom-Guided Evaluation

Clinician-assessed symptom-guided evaluation during hospital emergency department triage plays a crucial role in swiftly identifying and managing patients suspected of having infectious diseases like COVID-19. Upon arrival, patients undergo an initial assessment performed by triage nurses or clinicians. This comprehensive evaluation involved measuring vital signs, collecting essential demographic information, and assessing the patient's primary complaint. Subsequently, patients are meticulously screened for symptoms typically linked with infectious diseases, such as fever, cough, shortness of breath, fatigue, and others [134].

Following the initial assessment, clinicians use a combination of gathered information and clinical judgment to determine whether diagnostic testing is necessary and, if so, which type of test is most appropriate for the patient's circumstances. This decision-making process involves considering several factors, such as the patient's presenting symptoms, medical history, and during the COVID-19 pandemic, any potential exposure to the virus. Clinicians carefully evaluate the benefits and limitations of each testing method to ensure accurate diagnosis and optimal patient management.

For COVID-19 testing, clinicians would typically consider whether to administer an LFD test or a PCR test. The choice between these tests depended on various factors, including the urgency of obtaining results, the patient's clinical condition, and the healthcare setting's testing capabilities.

For instance, the NHS recommended using rapid antigen tests (LFD) for patients entering the emergency department with a possibility of admission, as these tests can facilitate early decision-making and help identify COVID-19 cases promptly [131, 135]. However, PCR tests were generally preferred for patients requiring critical care or those with severe symptoms, as they offer higher sensitivity and specificity in detecting the virus [136]. Additionally, all symptomatic or asymptomatic patients requiring emergency or unplanned admission via emergency departments were given PCR testing to ensure appropriate patient placement and infection control measures [135].

By carefully selecting the most appropriate testing method based on individual patient needs and clinical circumstances, clinicians were able to ensure timely diagnosis, appropriate patient management, and effective implementation of infection control measures, ultimately contributing to better outcomes for both patients and public health.

2.3.2 Real-Time Polymerase Chain Reaction

Real-time PCR testing was vital in the fight against the COVID-19 pandemic, especially in dynamic hospital settings where accurate diagnosis is paramount. High

sensitivity and specificity rendered PCR testing very dependable for identifying individuals infected with the virus, effectively distinguishing them from those who are not. This level of accuracy holds immense significance in clinical practice, guiding clinicians in tailoring patient management strategies, implementing targeted infection control measures, and curbing the transmission of viruses within communities.

The reliability of PCR testing was instrumental in several critical aspects of pandemic response. Firstly, it enabled clinicians to diagnose and isolate COVID-19 cases, preventing further transmission within healthcare facilities and the broader community. Additionally, accurate diagnosis facilitated initiation of appropriate medical interventions, thereby improving patient outcomes and reducing the strain on healthcare resources. Furthermore, PCR testing served as a cornerstone for epidemiological surveillance efforts, providing essential data for tracking the spread of the virus, identifying emerging hotspots, and informing public health policies and interventions.

Throughout the pandemic, Public Health England placed a high priority on bolstering public health testing initiatives. Their primary focus was on facilitating widespread screening and testing to swiftly identify and contain cases of COVID-19. The preferred method of screening/testing recommended by Public Health England involved molecular diagnosis using real-time PCR assays conducted on oral swab samples [137].

To achieve this level of screening, Public Health England collaborated closely with NHS England and NHS Improvement to scale up testing capacity across the healthcare system. By expanding testing capabilities, the aim was to promptly identify individuals infected with the virus, such that the necessary containment measures could be implemented to limit its spread within communities. Additionally, the concerted efforts aimed to delay and mitigate the transmission of the virus, providing crucial time for healthcare systems to prepare and respond effectively to the evolving situation.

The coordinated approach between Public Health England and NHS entities underscored a proactive strategy to confront the challenges posed by the pandemic.

By prioritizing robust testing infrastructure and capacity expansion, the aim was to establish a comprehensive framework for identifying and managing COVID-19 cases, ultimately contributing to the broader public health efforts to curb the spread of the virus and safeguard community health and well-being.

2.3.3 Lateral Flow Antigen Device Testing

Amidst the challenges posed by the COVID-19 pandemic, the urgency of promptly identifying and containing cases quickly became paramount in combating the spread of the virus. In meeting this critical need, rapid antigen testing, commonly known as LFD testing, emerged as a pivotal diagnostic tool. LFD testing offered a swift and accessible means of detecting the presence of COVID-19, providing rapid results within minutes and eliminating the necessity for specialized laboratory equipment [131, 138].

As nations worldwide contended with the complexities of mass testing and contact tracing, the incorporation of LFD testing attracted considerable attention for its potential to streamline case identification processes. By delivering rapid results, LFD testing enabled healthcare authorities to promptly isolate confirmed cases, initiate necessary treatment protocols, and implement targeted containment measures to curb transmission rates. Thus, the integration of rapid testing into broader testing strategies holds promise for enhancing the efficiency and effectiveness of public health responses.

The NHS implemented rapid testing for patients admitted via the emergency department using LFD on nasal swab samples, aiming to swiftly identify COVID-19 positive patients and reduce the risk of hospital-acquired infections [131].

Despite offering rapid results (typically within 15-30 minutes), LFD testing is less sensitive compared to PCR tests [138]. As a result, the NHS recommended using LFD testing as a supplementary measure to PCR testing, which typically takes several hours from swab collection to result [131]. The suggested approach involved responding exclusively to positive LFD results, with LFD testing not intended to replace mandatory PCR testing [131].

Consequently, in December 2020, the NHS issued a recommendation for emergency department settings, stipulating that a trained staff member should conduct both the LFD test and the PCR test simultaneously upon identifying a patient likely to be admitted [131]. Upon receiving a positive LFD result, patients were promptly classified as COVID-19 positive, eliminating the need to await PCR results. In cases where LFD results were indeterminate, patients were managed as potentially positive or negative, requiring isolation until confirmation through PCR testing. Similarly, patients with negative LFD results were isolated pending confirmation from PCR testing, ensuring cautious management and preventing potential transmission risks within the hospital environment.

2.4 Challenges of Diagnostic Methods for COVID-19 Patient Triage

While symptom-guided evaluation, rapid testing with LFDs, and PCR testing are essential for enhancing patient care and reducing transmission risks, they do come with their respective challenges.

Symptom-guided evaluation conducted by healthcare professionals for COVID-19 triaging encountered several challenges amidst the intricate landscape of the pandemic. Firstly, the broad array of symptoms linked to COVID-19, spanning from mild respiratory illness to severe respiratory distress and multi-organ dysfunction, introduced complexities in accurate diagnosis and triage [126, 127]. This spectrum of symptoms complicated the prompt identification of COVID-19 cases, which may have potentially led to delays in diagnosis and inadequate patient management, thus heightening the risk of transmission within healthcare settings.

Moreover, the characteristic symptoms of COVID-19 overlaps with that of other respiratory illnesses, such as influenza and the common cold, further complicating the differentiation of COVID-19 cases from other respiratory infections [139, 140]. This diagnostic ambiguity can result in excessive testing, overuse of resources, and heightened patient anxiety, especially during peak flu seasons when healthcare facilities are already under strain.

Although LFD testing had become a valuable tool in the fight against COVID-19, a notable drawback was its lower sensitivity compared to PCR testing, especially in low prevelance areas, or when detecting asymptomatic individuals or those with low viral loads [133, 141, 142]. For example, one study found that the LFD tests detected 46.2% of the positives detected by RT-PCR[143]. Another study found that LFD sensitivity versus RT-PCR was 63.2% [142]. Decreased sensitivity can increase the risk of false-negative results, potentially leading to undetected cases and further spread of the virus.

This issue becomes particularly concerning in scenarios where patients undergo assessment in the emergency department and await confirmatory PCR results in segregated areas. For example, placing individuals who tested negative for COVID-19 via LFD testing in close proximity to those confirmed positive or vice versa raises concerns about inadvertent exposure and virus transmission. This concern is pertinent to the application of NHS trusts' triage system, categorized as "green–amber–blue," and their protocol for LFD testing. As per this protocol, patients with indeterminate or negative LFD results were isolated until PCR results became available.

Despite its reputation for high sensitivity and specificity, PCR testing also faced challenges, one of which was the turnaround time for test results [144]. This timespan can fluctuate considerably based on various factors, including testing capacity, laboratory workload, and logistical constraints. Delays in receiving PCR test results posed a notable obstacle to timely diagnosis and subsequent patient management. Such delays could impede efforts to swiftly implement effective triaging and isolation measures, potentially allowing for further transmission of the virus within communities and healthcare facilities. As a result, strategies to minimize turnaround times and streamline the PCR testing process were crucial for enhancing the efficiency and efficacy of COVID-19 diagnostic efforts [145].

Overall, one fundamental challenge persists in emergency department triage: the need to swiftly provide relevant results to healthcare providers in a practical manner. In the case of COVID-19 diagnosis, the resulting delays in diagnosing

infections using PCR tests led to obstacles in promptly identifying positive cases. And, while LFDs offered expedited time-to-results, their sensitivity tended to be suboptimal and variable.

Tackling this challenge necessitates continuous efforts to refine testing protocols, scale up testing capacity, and enhance turnaround times, all without imposing significant burdens or expense. Moreover, exploring alternative and innovative testing methods could enhance our diagnostic capabilities and mitigate some of these challenges. Broadening our range of testing options holds promise for enhancing our capacity to detect and manage virus transmission, such as with COVID-19, more effectively. Specifically, this thesis explores the deployment of an AI-powered COVID-19 screening tool to aid in triage within hospital emergency departments.

2.5 Machine Learning for Rapid COVID-19 Patient Triage

During the pandemic, emergency departments faced a significant influx of patients exhibiting symptoms suggestive of COVID-19 infection, leading to concerns about the potential spread of the virus among other patients and healthcare staff. A key challenge was determining an effective method to identify and segregate patients suspected of having COVID-19 from those with other medical conditions, thereby minimizing the risk of transmission within the hospital setting. This challenge was exacerbated by numerous factors, including the limited availability of suitable laboratory screening tests, their suboptimal sensitivity in detecting the virus, the associated high costs, and the resulting delays in obtaining confirmatory test results. As a result, emergency departments were under pressure to quickly implement strategies to address these challenges and enhance their capacity for timely and accurate identification of COVID-19 cases while mitigating the risk of in-hospital contamination.

Machine learning presents a promising avenue for enhancing COVID-19 triage protocols in emergency departments by harnessing its capacity for sophisticated data analysis, pattern recognition, and predictive modeling. In the fast-paced environment

of emergency settings, where copious amounts of patient data are generated, machine learning algorithms offer the capability to efficiently parse through this vast collection of data. They can discern subtle patterns and correlations that may not be immediately apparent to human clinicians, thereby aiding in the early identification of potential COVID-19 cases.

Through the analysis of varied datasets comprising clinical symptoms, laboratory results, imaging data, and patient demographics, machine learning models can deliver precise predictions regarding the probability of COVID-19 infection. This functionality enables healthcare professionals to select and implement suitable interventions and patient management protocols. Furthermore, machine learning algorithms can accelerate the identification of COVID-19 cases, facilitating swift isolation and treatment measures to mitigate further transmission, both within the emergency department and in the wider community.

Rapid triaging facilitated by machine learning also holds the potential to optimize resource allocation in emergency departments. By prioritizing patients based on their predicted risk of COVID-19 infection, healthcare resources can be allocated more efficiently, ensuring that critical interventions are promptly administered to those most in need. This streamlined approach not only enhances patient care but also contributes to better overall outcomes by minimizing delays in diagnosis and treatment initiation.

The existing body of literature on COVID-19 triage systems demonstrates a variety of tools offering swift and effective decision-support mechanisms. These systems incorporate a variety of predictors, ranging from clinical data encompassing symptoms and vital signs to laboratory tests and radiological findings [144, 146–149]. Through the integration of these diverse data sources, these triage systems aim to provide comprehensive assessments that aid healthcare professionals in making timely and informed decisions regarding patient management and care.

2.5.1 Real-World Evaluation of AI-Driven COVID-19 Triage in Emergency Department

This section provides a concise overview of the findings from our investigation into validating an AI-driven COVID-19 triage model. The results of this study build upon a previous study [144], and have been documented and published in Lancet Digital Health [133].

The process of COVID-19 triaging involves classifying individuals based on their likelihood of being infected with the virus, making it a supervised learning classification task (a detailed description of supervised learning is provided in Chapter 4.1.2). In our particular study, we investigated a cohort comprising 72,223 patients from four distinct hospital groups - Oxford University Hospitals NHS Foundation Trust (OUH), Portsmouth Hospitals University NHS Trust (PUH), University Hospitals Birmingham NHS Trust (UHB), and Bedfordshire Hospitals NHS Foundations Trust (BH), covering the period from December 1, 2019, to March 31, 2021. Acronyms for these hospital groups are provided for consistency throughout this thesis.

We introduced two distinct models: CURIAL-Lab, which incorporates vital signs and commonly available blood tests (including full blood count, urea, creatinine, electrolytes, liver function tests, and C-reactive protein), and CURIAL-Rapide, which relies solely on vital signs and full blood count data. These models are developed using Extreme Gradient Boosting (XGBoost) (further explained in Chapter 4.2), and underwent external validation and prospective evaluation for emergency admissions across four UK NHS trusts.

Our analysis indicated consistent performance across various healthcare trusts for both CURIAL-Lab and CURIAL-Rapide, as evidenced by AUROC values ranging from 0.858 to 0.881 (95% CI 0.838–0.912) for CURIAL-Lab and from 0.836 to 0.854 (0.814–0.889) for CURIAL-Rapide, as shown in Table 2.1. This matched the gold standard benchmark set by real-time PCR, which has estimated sensitivities of approximately 80%-90% [150, 151]. The table presents sensitivity, specificity,

Table 2.1: Performance of CURIAL-Rapide during prospective and external validation. Results are optimized to sensitivities of 0.9. OUH results are reported alongside SD across imputation methods. PUH, UHB, and BH results reported alongside 95% confidence intervals. These results are taken directly from [133]

Test Set	Sensitivity	Specificity	PPV	NPV	AUROC
OUH	85.6 (0.6)	59.1 (0.3)	9.46 (0.0)	98.8 (0.0)	0.843 (0.002)
PUH	83.5 (81.8-85.1)	63.6 (63.1-64.1)	11.4 (10.9-11.9)	98.6 (98.4-98.7)	0.842 (0.832-0.852)
UHB	82.2 (78.4-85.5)	65.4 (64.5-66.3)	9.6 (8.7-10.6)	98.8 (98.5-99.0)	0.836 (0.814-0.858)
BH	74.3 (66.6-80.7)	81.8 (79.3-84.0)	36.3 (31.0-41.9)	95.8 (94.3-96.9)	0.854 (0.819-0.889)

positive and negative predictive values (PPV and NPV), and the area under receiver operator characteristic curve (AUROC).

Incorporating these models with LFDs led to enhanced triage sensitivity. Specifically, the model helped increase sensitivity from 56.9% (51.7–62.0) with LFDs alone to 85.6% (81.6–88.9) with CURIAL-Lab and 88.2% (84.4–91.1) with CURIAL-Rapide.

Following this, CURIAL-Rapide was implemented for prospective use at the John Radcliffe Hospital, where 520 patients underwent point-of-care full blood count analysis between February 18 and May 10, 2021. An OUH-approved service evaluation (Ulysses ID 6907) oversaw the implementation of two OLO rapid haematology analysers in the hospital's emergency department in Oxford. The aim was to evaluate the operational and predictive performance of CURIAL-Rapide in a laboratory-free setting. Of the enrolled participants, 436 underwent confirmatory PCR testing, revealing that ten individuals (2.3%) tested positive for COVID-19.

Figure 2.1 illustrates the patient flow upon arrival at the hospital emergency department. Initially, upon arrival, patients undergo routine blood tests and vital signs recordings. This process occurs either through rapid point-of-care haematology analysers, taking approximately 10 minutes, or via the existing laboratory pathway, which typically takes about 1 hour. As depicted, real-time algorithmic analysis enables the early and confident identification of patients who test negative, allowing for their safe triage to COVID-19-free (green) clinical areas. Patients who receive positive CURIAL results are directed to enhanced precautions (amber) areas,

pending confirmatory PCR testing. Those who test positive using a LFD test are immediately directed to COVID-19 (red) clinical areas.

This real-time AI-driven approach showcased a considerable reduction in time-to-result, with the median time from arrival to obtaining a CURIAL-Rapide result being 45 minutes. This marked a 16-minute (26.3%) improvement compared to LFDs, which took a median time of 61 minutes, and a significant advancement of 6 hours and 52 minutes (90.2%) compared to PCR, which required a median time of 7 hours and 37 minutes. The classification accuracy of CURIAL-Rapide was notably high, boasting a sensitivity of 87.5% (52.9–97.8), specificity of 85.4% (81.3–88.7), and negative predictive value of 99.7% (98.2–99.9). Moreover, CURIAL-Rapide effectively ruled out infection for 31 (58.5%) of the 53 patients initially suspected of COVID-19 by a physician but who later tested negative via PCR.

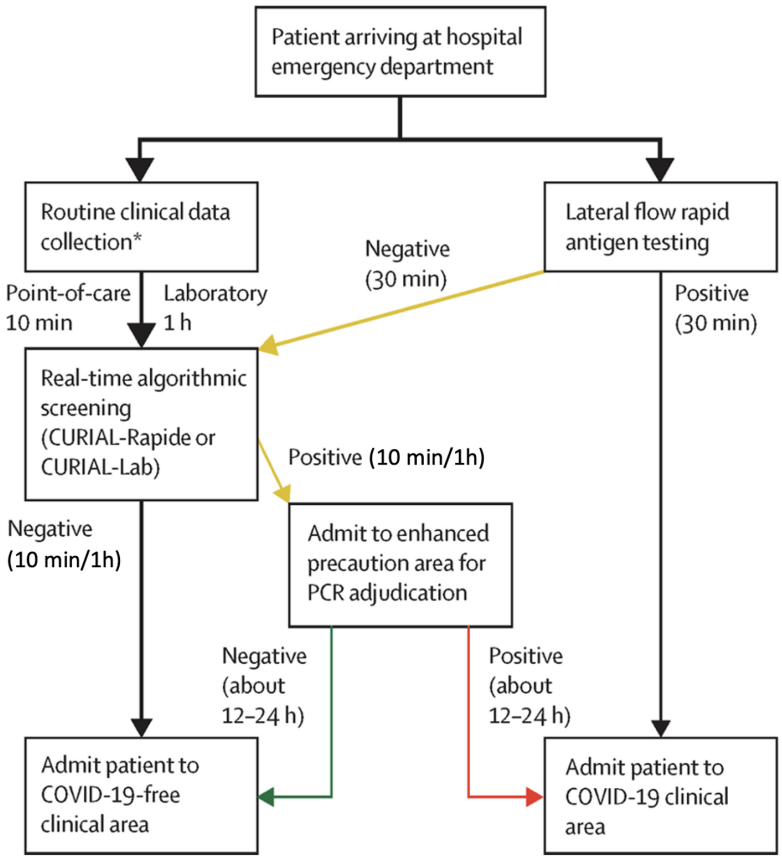

Figure 2.1: Patient flow upon arrival to the hospital emergency department. Figure is taken from [133].

This study emphasizes a clear and straightforward message: incorporating an AI-driven model into the triage process upon admission to the emergency department can provide significant decision support. This integration considers various factors such as the model's accessibility, ease of use, and its ability to provide rapid diagnostic outcomes.

Furthermore, the notably high negative predictive value highlighted in the study emerges as a crucial asset for frontline triage. This indicates the model's effectiveness in accurately identifying patients who are unlikely to have COVID-19. Such accuracy could aid in the effective management of the influx of patients, particularly during periods of heightened demand or surges in COVID-19 cases. By swiftly ruling out infection in individuals deemed low-risk, healthcare providers can better allocate resources and prioritize care for those most in need, thereby enhancing the overall efficiency and effectiveness of the triage process.

The strength of this study lies in its rigorous validation process and real-time evaluation of an AI-driven model within the operational setting of an emergency department. By subjecting the model to external validation, we were able to assess its performance across diverse patient populations and healthcare settings, bolstering confidence in its reliability and generalizability. Moreover, conducting real-time assessments provided valuable insights into the model's practical utility and feasibility in a dynamic and fast-paced clinical environment, further substantiating its potential relevance in emergency scenarios.

Despite these encouraging findings, the results indicate that the model does not achieve perfect generalizability, as evidenced by varying sensitivities and specificities across different hospital sites (Table 2.1). Consequently, concerns remain regarding the potential presence of site-specific biases among hospitals and the fairness of the model's deployment, along with potential biases in its resultant outcomes. Recognizing these concerns, subsequent chapters of this thesis aim to tackle issues related to fairness in machine learning algorithms. In particular, the focus shifts towards the development, deployment, and assessment

of machine learning techniques that prioritize fairness, with a specific emphasis on statistical outcome fairness.

By prioritizing fairness, the goal is to ensure that AI-driven models produce equitable outcomes for all patients, regardless of demographic or socioeconomic factors. The proposed methods aim to mitigate any biases that may exist within the model's decision-making process, thereby promoting greater equity in patient care and treatment outcomes. Ultimately, the following chapters aim to enhance the overall fairness and effectiveness of the AI-driven triage system introduced, contributing to more equitable healthcare delivery in emergency settings.

3

Identification of Data-Level Bias Between Hospitals

3.1 Introduction

3.1.1 Overview

In this chapter, we explore the impact of bias originating from the geographic origins of clinical data, specifically in healthcare contexts. We provide an overview of the four main datasets central to this thesis, outlining their structure, characteristics, and the methodologies used for preprocessing in the development of AI-based COVID-19 screening models. Additionally, we examine the potential biases inherently present within these datasets, underscoring the pivotal necessity of upholding fairness and equity in AI-driven strategies. This assessment serves as the foundation for subsequent chapters, where we delve deeper into the development, implementation, and evaluation of fairness-aware AI-driven screening algorithms. By scrutinizing the datasets and acknowledging potential sources of bias, we aim to foster transparency, reliability, and equity in the subsequent COVID-19 screening methodologies presented.

3.1.2 Location-Based Bias in Healthcare

Bias originating from individual hospitals can arise from various complex factors inherent in the healthcare system and the diverse patient populations served by different hospitals.

Firstly, bias across datasets sourced from distinct hospitals may be influenced by the ethnic and genetic demographic of the local population. Ethnicity has been found to influence the prevalence of certain diseases and health conditions due to genetic predispositions, cultural practices, socioeconomic factors, and access to healthcare services [152–155]. For example, certain genetic variations may predispose individuals to specific diseases or affect their response to medications [154]. Consequently, the disease profiles, treatment outcomes, and physiological characteristics can and will vary between hospitals.

Additionally, disparities in healthcare utilization among ethnic groups can further exacerbate bias across datasets sourced from various hospitals [156, 157]. Elements like language barriers, cultural preferences, distrust of the healthcare system, and discrepancies in access to care can significantly influence healthcare-seeking behaviors among different ethnicities. As a result, hospitals serving communities with diverse ethnic backgrounds may encounter differences in patient demographics, patterns of healthcare utilization, and disease manifestation.

Ethnicity also frequently intersects with socioeconomic status [158], which can, in turn, lead to imbalanced representation within clinical data. For example, medical centers serving populations with higher socioeconomic status or improved access to healthcare resources may exhibit superior outcomes. Conversely, hospitals catering to marginalized or underserved communities may lack essential elements of high quality care, such as advanced medical equipment and sufficient staffing, leading to discrepancies in patient outcomes[159].

Moreover, variations in the implementation of advanced technologies and innovations can also contribute to disparities in clinical data quality [160]. Hospitals equipped with cutting-edge tools and technologies may generate more thorough and precise clinical data compared to those with limited resources or technological

infrastructure. In contrast, healthcare facilities facing resource constraints may rely on less advanced diagnostic equipment or manual methods for data recording, resulting in incomplete or less accurate documentation of patient information.

Finally, the influence of research affiliations and academic partnerships can impact the quality and types of clinical data collected by hospitals. Hospitals affiliated with research institutions may participate in research studies or clinical trials, where rigorous data collection standards are enforced to meet research requirements. As a result, these hospitals may prioritize data accuracy, completeness, and standardization, contributing to higher-quality clinical data. Conversely, hospitals without research affiliations may not have the same level of emphasis on data quality or standardization, potentially leading to inconsistencies or inaccuracies in the recorded clinical data.

Overall, bias based on the hospital attended by a patient manifests itself in clinical data through differences in patient demographics, healthcare practices, and resource availability across hospitals [161–163]. These biases can lead to disparities in data quality, completeness, and accuracy, affecting the reliability and generalizability of clinical research findings, healthcare interventions, and machine learning models trained on clinical data. Thus, addressing location-based biases requires concerted efforts to improve data collection practices, standardize clinical protocols, and promote equity in healthcare delivery to ensure fair and accurate representation of patient populations across different hospitals.

3.2 Methods

3.2.1 Dataset

To train and validate the machine learning models proposed in this thesis, we use clinical data with linked, deidentified demographic information for patients presenting to emergency departments across four independent UK NHS Trusts – OUH, PUH, UHB, and BH. Each hospital operates within its own distinct IT infrastructure. However, in general, laboratory data is managed within a system referred to as LIMS (Laboratory Information Management System). The data

extraction process for these datasets typically involved sourcing data from either a LIMS mirror, a trust integration system that interfaces with LIMS, or a direct extraction from the LIMS system itself. With respect to these datasets, UK NHS approval via the national oversight/regulatory body, the HRA, has been granted for development and validation of artificial intelligence models to detect COVID-19 (CURIAL; NHS HRA IRAS ID: 281832). The studies are limited to working with deidentified data, and extracted retrospectively; thus, explicit patient consent for use of the data was deemed to not be required, and is covered within the HRA approval. All necessary consent has been obtained and the appropriate institutional forms have been archived.

These four UK datasets used are identical to those used in the validation study discussed in Chapter 2.5.1, allowing for direct comparison. Specifically, for OUH, we included all patients presenting and admitted to the emergency department. As for PUH, UHB, and BH, the inclusion criteria comprised all patients admitted to the emergency department. The full inclusion and exclusion criteria for patient cohorts can be found in Appendix A.1.

3.2.2 Training, Continuous Validation, and Test Dataset Partitioning

For each of the models, a training set, $\mathcal{D}_{train}$, was used for model development, hyperparameter selection, and model training; a continuous validation set, $\mathcal{D}_{val}$, was used for continuous model validation during development and threshold adjustment; and after successful development and training, the held-out test set, $\mathcal{D}_{test}$, was used to evaluate the performance of the final models.

From OUH, we curated two data extracts corresponding to distinct periods: the first "wave" of the COVID-19 pandemic in the UK (December 1, 2019, to June 30, 2020), and the second "wave" (October 1, 2020, to March 6, 2021). A plot of OUH positive cases, showing the first and the second waves is shown in Figure 3.1.

During the initial wave, challenges such as incomplete testing and the imperfect sensitivity of the PCR swab test led to uncertainties in determining the viral status

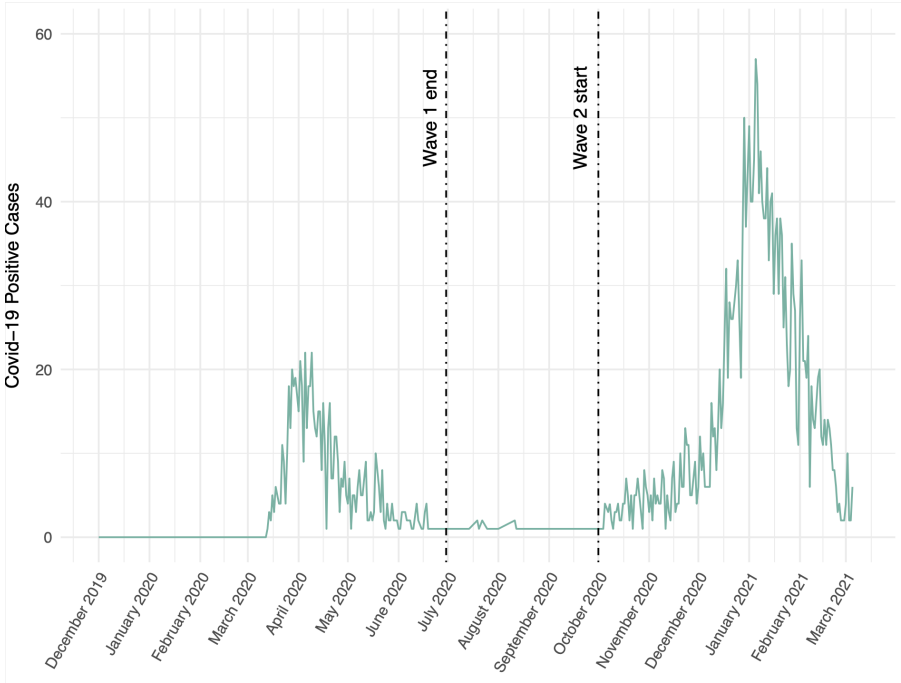

Figure 3.1: Plot of OUH positive cases, showing the first "wave" of the COVID-19 pandemic in the UK from December 1, 2019 to June 30, 2020; and the second "wave" from October 1, 2020 – March 6, 2021.

of patients who were either untested or tested negative [144]. Thus, from the "wave one" dataset, we only included the positive cases (as determined through PCR tests) in training; and from the "wave two" dataset, we included both positive COVID-19 cases (by PCR) and negative controls. This was done to ensure that the label of COVID-19 status was correct during training. This resulted in a prevalence of 11.1% used during training, which is within the spatial and temporal range of prevalences observed across the UK trusts used in our study (prevalences between 4.27%-12.2%, as shown in Table 3.1).

We divided the OUH data into training, continuous validation, and test sets through a random split, allocated at 60%, 20%, and 20%, respectively.

Since UHB, PUH, and BH each provided a single dataset from a defined period of the pandemic, we further divided these into training, continuous validation, and test sets through a random split, allocated at 60%, 20%, and 20%, respectively. This division was stratified based on the COVID-19 status, which was determined through confirmatory PCR testing.

Table 3.1: Summary population characteristics and COVID-19 prevalence for OUH (wave one and wave two), PUH, UHB, and BH cohorts. * indicates merging for statistical disclosure control.

	OUH (wave one cases)	OUH (wave two cases)	PUH	UHB	BH
Total Patients	701	22,857	37,896	10,293	1177
COVID positive (%)	701 (100%)	2,012 (8.80%)	2,005 (5.29%)	439 (4.27%)	144 (12.2%)
Sex:					
- Male (%)	376 (53.64)	11409 (49.91)	20839 (54.99)	4831 (46.93)	627 (53.27)
- Female (%)	325 (46.36)	11448 (50.09)	17054 (45.0)	5462 (53.07)	549 (46.64)
Age, yr (IQR)	72 (55-82)	67 (49-80)	69 (48-82)	63 (42-79)	68.0 (48-82)
Ethnicity:					
-White (%)	480 (68.47)	17387 (76.07)	28704 (75.74)	6848 (66.53)	1024 (87.0)
-Not Stated (%)	128 (18.26)	4127 (18.06)	8389 (22.14)	1061 (10.31)	≤10
-South Asian (%)	22 (3.14)	441 (1.93)	170 (0.45)	1357 (13.18)	71 (6.03)
-Chinese (%)	*	51 (0.22)	42 (0.11)	41 (0.4)	≤10
-Black (%)	25 (3.57)	279 (1.22)	187 (0.49)	484 (4.7)	36 (3.06)
-Other (%)	34 (4.85)*	410 (1.79)	269 (0.71)	333 (3.24)	29 (2.46)
-Mixed (%)	12 (1.71)	162 (0.71)	135 (0.36)	169 (1.64)	13 (1.1)

Finally, we merged data from all sites, yielding final training, continuous validation, and held-out test sets comprising of 58,339 presentations used in training and continuous validation (including 4,245 of which were COVID-19 positive), and 14,585 presentations in the held-out test set (including 1,056 of which were COVID-19 positive). A summary of each dataset is provided in Table 3.2. It is important to note that the same individual may appear in multiple datasets if they had multiple visits to the emergency department. Each presentation was treated as an independent case, reflecting real-world scenarios.

Table 3.2: Summary of number of patients and COVID-19 positive cases for training, continous validation, and test sets.

	Training ($\mathcal{D}_{train}$)	Continuous Validation ($\mathcal{D}_{val}$)	Test ($\mathcal{D}_{test}$)
Total Patients	43,754	14,585	14,585
COVID-19 positive (PCR)	3,171	1,074	1,056
Hospital:			
OUH (%)	14,104 (32.2)	4,694 (32.2)	4,760 (32.6)
PUH (%)	22,750 (52.0)	7,596 (52.1)	7,550 (51.8)
UHB (%)	6,186 (14.1)	2,062 (14.1)	2,045 (14.0)
BH (%)	714 (1.6)	233 (1.6)	230 (1.6)

Since we removed patients from "wave one" who were labeled as "COVID-19 negative," we performed sensitivity analysis to account for this. We trained a COVID-19 prediction model and evaluated this on the continuous validation set (as to ensure that the test sets were not used until a final model is developed), achieving AUROC scores of 0.836 (95% CI 0.811-0.860) and 0.857 (0.833-0.880) for the original and

adjusted training sets, respectively. The comparable results (overlapping confidence intervals) demonstrate model stability across these cohorts. Additionally, the AUROCs achieved are consistent with those of CURIAL-Lab and CURIAL-Rapide in the original validation study discussed in Chapter 2, demonstrating that the models trained are robust and align with the established benchmarks.

3.2.3 Clinical Features Used for Diagnosis

As described in Chapter 2.5, our focus is on rapid patient triaging [133, 144], acting as a preliminary measure during the period when confirmatory laboratory testing is awaiting results or when access to definitive molecular testing for COVID-19 is constrained. With an emphasis on scalability, the datasets used encompass a segment of regularly acquired clinical data, comprising initial blood tests, vital signs, and the confirmation of COVID-19 diagnosis through a PCR swab test. This approach was chosen due to the widespread and standardized practice of collecting such data within the initial hour of a patient's arrival in emergency care pathways. This practice was consistent across hospitals situated in middle- to high-income countries [133]. By using these commonly collected metrics, our aim was to ensure the applicability and practicality of our triaging models across a broad spectrum of healthcare settings, facilitating efficient and effective patient assessment and management. To ensure a meaningful comparison with prior studies, the features included are similar to those used in the study described in Section 2.5.1, which references the relevant published works [133, 144]). Table 3.3 summarizes the final features included in training.

Regarding the data structure, the clinical features employed are presented in a tabular layout, organized into rows and columns. This structured format systematically arranges the data, where each row corresponds to information pertaining to an individual patient, and each column signifies a specific attribute or measurement. Thus, in our case, these columns depict laboratory blood test outcomes and vital sign readings.

Table 3.3: Clinical predictors considered for COVID-19 status prediction.

Category	Features
Vital Signs	Heart rate, respiratory rate, systolic blood pressure, diastolic blood pressure, temperature
Blood Tests	Haemoglobin, haematocrit, mean cell volume, white cell count, neutrophil count, lymphocyte count, monocyte count, eosinophil count, basophil count, platelets
Liver Function Tests & C-reactive protein	Albumin, alkaline phosphatase, alanine aminotransferase, bilirubin, C-reactive protein
Urea & Electrolytes	Sodium, potassium, creatinine, urea, estimated glomerular filtration rate

3.2.4 Data Preprocessing

To start, we ensured consistency in the units used for identical features. Here, all features included are continuous variables, and their units can be found in Table A.1 in Appendix A.

In managing missing values within the dataset, we employed population median imputation. This method involves replacing missing values with the median of the corresponding feature across the entire dataset, preserving the overall distribution and variability of the data. During the study highlighted in Chapter 2.5.1, various imputation strategies, including population median, population mean, and age-based imputation, were initially used to address missing data. As part of a sensitivity analysis examining the impact of imputation strategies on model performance, we conducted a prospective evaluation—assessing models in a forward-looking, real-world context that accounts for their influence on future data collection and distribution [164]—of models trained with each imputation method in all patients presenting to OUH emergency departments during the second wave of the COVID-19 pandemic. Performance metrics, such as mean values and standard deviations (SD), are detailed in Table A.2 in Appendix A.2. Narrow standard deviations observed across all performance metrics indicate robustness to the imputation method used. For example, for AUROC, the means and SDs were 0.843 (0.002) and 0.878 (0.001)

for CURIAL-Rapide and CURIAL-Lab, respectively. Thus, we proceeded to assess models trained with missing data imputed using population median. Additionally, a comprehensive summary of data completeness is provided in Table A.1.

Finally, all features underwent standardization to achieve a mean of 0 and a standard deviation of 1. This standardization process ensures that all features have similar scales and distributions, preventing certain features from dominating others during the training process. Additionally, standardization aids in optimization and convergence when training neural network (NN) models, which we use in our investigations [165–167].

3.2.5 Bias Identification Methods

In our examination of the four distinct UK datasets, we use established methodologies commonly employed in data science and machine learning to assess data-level bias. While we acknowledge the importance of identifying and quantifying biases within datasets, our primary emphasis in this thesis lies in the development and implementation of fairness-aware methods. Thus, instead of extensively exploring bias identification methods, our focus is on devising strategies to mitigate bias and foster equitable outcomes in AI-driven decision-making processes.

To examine potential sources of bias across different hospital sites, we start by using the Kruskal-Wallis test, a statistical method employed to assess covariate shift. This test allows us to discern any variations in data distributions, particularly differences in the medians, among various hospitals. By employing this test, we aim to identify and quantify any disparities that may exist in the data collected from different hospital settings.

At the sample level, we employ t-Stochastic Neighbor Embedding (t-SNE) to create low-dimensional visualizations encompassing all positive COVID-19 cases obtained from the four NHS sites (Figure 3.2). Through t-SNE, we generate plots that condense the complex, high-dimensional data into a more manageable and interpretable form. By examining the clusters within the t-SNE plot, we can discern patterns and similarities among the data points that may not be immediately evident

in the original high-dimensional space. This visualization technique provides valuable insight into the underlying structure of the data, facilitating the identification of clusters or groupings that may indicate commonalities or differences among the samples. Moreover, these visualizations help in uncovering potential biases or site-specific influences that may impact the interpretation of the data. By observing the distribution of clusters, we can gain a deeper understanding of the data and make informed decisions regarding subsequent analyses or interventions aimed at addressing any identified biases.

3.3 Results

On a population-level, Table 3.2 illustrates varying sample sizes across different hospital groups, suggesting potential bias due to unequal representation of each group in the dataset. Such discrepancies can distort analysis, outcomes, and conclusions, particularly in machine learning tasks, where they may unduly impact results, potentially resulting in misleading interpretations.

At the feature level, across all UK cohorts, every matched feature exhibited a notable difference in population median (Kruskal-Wallis test, $p < 0.0001$, determined as statistically significant using a threshold value of 0.05), except for platelets, where the population median appeared comparable ($p = 0.127$). These statistically significant variations in feature distributions among different locations imply the existence of biases across these areas. Supplementary Table A.3 provides comprehensive summary statistics, including median and interquartile ranges, for vital signs and blood tests across all patient cohorts.

Figure 3.2 displays the t-SNE plot obtained when applied on the dataset. Here, the presence of an isolated green cluster corresponding exclusively to a subset of presentations from one NHS site - OUH - suggests that the data from OUH has distinct features or characteristics that separates it from data collected at other sites. This underscores the importance of accommodating site-specific biases during model development.

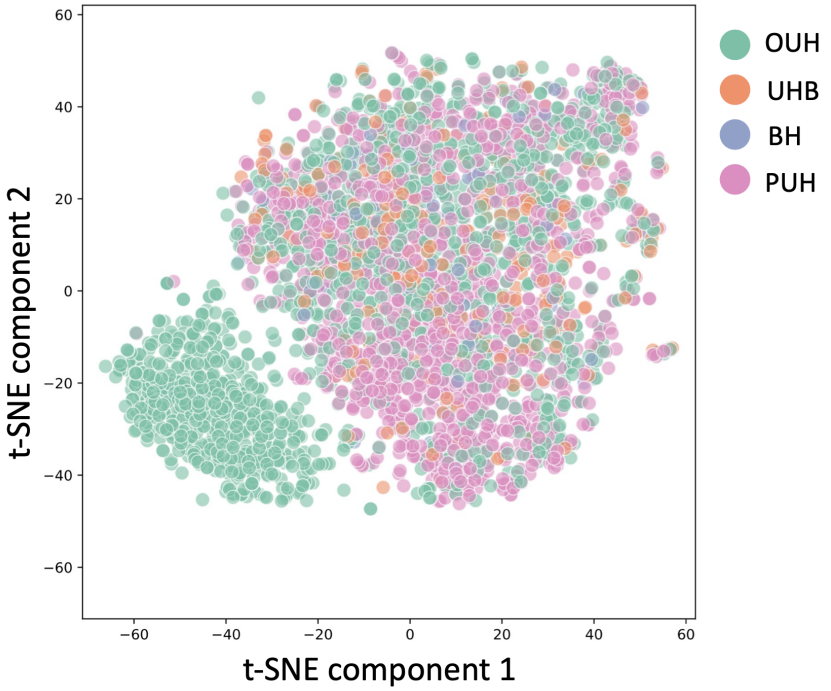

Figure 3.2: Visualization via t-SNE representation of datasets used in the study, including all positive COVID-19 cases across four NHS trusts (OUH, PUH, UHB, BH). This figure has been published in [168]

3.4 Discussion

Although PCR is considered the diagnostic benchmark, it is important to recognize the limitations of using PCR as the gold standard for COVID-19 diagnosis. Its sensitivity is imperfect and can vary with factors such as viral load, sample collection technique, and timing of infection [169, 170]. Consequently, some patients labeled as "COVID-19 negative" may in fact have been positive, introducing noise into the ground-truth labels. This label uncertainty may propagate into model training and evaluation, and must be considered when interpreting performance metrics.

At a population level, it is evident that there exists unequal representation concerning the quantity of available data across individual hospital sites. When soliciting data from diverse hospitals, numerous factors can lead to the reception of disparate amounts of data from each site. These factors include discrepancies in policies and agreements regarding data sharing and collaboration with external parties among hospitals, divergent capabilities in granting access to data owing to

variations in data storage systems, data sharing agreements, data governance policies, as well as the tendency of hospitals with limited resources to prioritize clinical operations over data sharing initiatives, potentially causing delays or limitations in providing data to external researchers.

When considering downstream analyses and such a dataset, this unequal subgroup representation can profoundly influence the performance and fairness of machine learning models. When certain subgroups are underrepresented or overrepresented in the training data, it can lead to biased predictions and undermine the model's ability to generalize effectively. One consequence of unequal subgroup representation is biased predictions, where the model may learn patterns primarily from the majority subgroup and struggle to accurately predict outcomes for underrepresented groups. This bias arises because the model's training process prioritizes patterns that are more prevalent in the training data, potentially leading to inaccurate or unfair predictions for minority subgroups. For example, if certain subgroups are consistently underrepresented in the training data, the model may systematically misclassify or disadvantage those groups, resulting in unfair outcomes. Furthermore, unequal subgroup representation can hinder the model's ability to generalize to new, unseen data from diverse populations. This limitation can undermine the model's utility in real-world applications, where it may encounter diverse populations with varying characteristics and needs.

The Kruskal-Wallis test, conducted to compare features across four distinct hospitals, revealed a statistically significant difference in the distribution of feature values among these hospitals. Specifically, the test findings suggest that there are variations in the medians of the feature values across the hospitals, signifying a potential influence exerted by the hospital site on the features being investigated. As previously discussed, this difference may originate from a multitude of factors, such as differences in patient demographics, healthcare practices, or other contextual elements inherent to each hospital setting. These distinctions underscore the importance of considering the hospital site as a relevant factor when analyzing clinical data, as it may contribute to significant variations in the distribution and

characteristics of the features being investigated. Understanding and accounting for these site-specific influences are crucial for ensuring the accuracy and reliability of any subsequent analyses or predictive models built upon the data from these hospitals.

In the context of COVID-19 cases across multiple NHS sites, the visualization of distinct clusters in the t-SNE plot provides valuable insights into the shared characteristics among data points originating from specific sites. These clusters represent patterns or similarities in the data that are inherent to each NHS site and are indicative of various factors influencing data collection, processing, and representation. Beyond distribution differences due to patient demographics and healthcare practices, a potential factor contributing to the formation of distinct clusters can be differences in annotation methods used at each site. Annotation methods refer to the process of labeling or categorizing data points based on specific criteria or attributes. Variations in annotation methods across NHS sites can result in differences in how features are classified or categorized, leading to the emergence of distinct clusters in the t-SNE plot. Similarly, discrepancies in data truncation, which involves limiting or cutting off data beyond a certain threshold or range, can also contribute to the formation of distinct clusters. Data truncation may occur due to various reasons, such as data storage limitations, data preprocessing procedures, or data quality control measures implemented at each site. These differences in data truncation practices can result in variations in the distribution and representation of COVID-19 cases within each hospital site. Furthermore, variations in measurement devices or data collection tools used at different NHS sites can also contribute to the presence of distinct clusters in the t-SNE plot. Differences in the type, accuracy, or calibration of measurement devices used to collect clinical data, such as temperature sensors, respiratory monitors, or laboratory analyzers, can lead to variations in the recorded data and subsequently influence clustering patterns. Similarly, variations in data collection and processing tools, such as electronic health record systems or data management software, can also impact the representation of COVID-19 cases within each NHS site and contribute to the formation of distinct clusters.

Finally, it is important to acknowledge that bias may persist even after addressing missing data through imputation techniques. While imputation helps fill in missing values, the missingness itself could carry significant information or reflect underlying biases in data collection processes. For example, disparities in clinical practices or recording standards among hospitals or healthcare facilities can lead to discrepancies in missing data patterns. This missing data, if systematically affecting specific subgroups or variables, can introduce bias. Therefore, it is essential for future studies to explore alternative methods to quantify and address missing data comprehensively. Simply imputing missing values may not fully capture the complexities and nuances of the underlying biases present in the data. Researchers may need to employ more sophisticated approaches, such as pattern recognition algorithms [171] or causal inference methods [172, 173], to identify and mitigate biases associated with missing data.

Understanding the significance of unintended biases and disparities that may emerge among various hospital sites, potentially influencing the data, is essential for accurately interpreting and applying analyses or models incorporating these factors. Additional investigation is required to uncover the underlying causes of these discrepancies and to assess their potential impact on the precision and applicability of findings across different hospital settings. This deeper comprehension is critical for validating and ensuring the relevance of research outcomes across a range of healthcare environments.

Therefore, regarding subsequent analyses, it is imperative to actively recognize and alleviate any potential biases linked to hospital location. This is especially critical in the creation of machine learning models, given that their performance hinges entirely on the quality of the data they are trained on. By acknowledging and rectifying these biases, we can uphold fairness and accuracy in predictions across a wide range of healthcare contexts. This proactive approach is vital for ensuring that machine learning models provide equitable and reliable insights that can benefit patients across different healthcare settings.

4

Bias Mitigation for Supervised Learning

4.1 Introduction

4.1.1 Overview

As highlighted earlier, recent years have seen a growing emphasis on ensuring fairness and equity in machine learning outcomes, spurring the development of innovative methodologies in fairness-aware algorithms. In this chapter, we will explore in detail the specific machine learning techniques designed to address this critical challenge by focusing on bias mitigation at the algorithmic level. These techniques have gained significant traction within the machine learning community, as demonstrated by a wealth of research studies dedicated to advancing fairness in AI systems [76, 79, 81, 101, 102, 174–176], establishing themselves as state-of-the-art strategies for addressing bias in machine learning models [80].

This heightened focus on fairness and equity in machine learning is driven by the increasing influence of algorithms in critical sectors such as healthcare, finance, and criminal justice [94, 177, 178]. As algorithms play a more significant role in decision-making processes, there is a growing imperative to ensure that these algorithms are fair and free from biases. The consequences of biased algorithms can be far-reaching, impacting individuals' quality of care, access to opportunities, resources, and even

their fundamental rights. Therefore, the need to mitigate biases and promote fairness in machine learning outcomes has become more urgent than ever before.

This chapter explores the principles involved in the development, utilization, and assessment of fairness-aware algorithms. We showcase practical applications of various supervised learning techniques designed to alleviate unintentional biases inherent in training data and enhance fairness in machine learning outcomes. We illustrate these methods using the COVID-19 screening task discussed in earlier chapters. Through empirical analysis, we examine the effectiveness and constraints of these techniques in promoting fairness and reducing disparities in algorithmic decision-making.

4.1.2 The Three Branches of Machine Learning

Machine learning, a subset of artificial intelligence, encompasses various techniques designed to enable systems to learn and improve from experience without explicit programming. One possible way to categorize these methods is into three branches, each addressing specific learning scenarios and objectives: supervised learning, unsupervised learning, and reinforcement learning (RL) [179, 180].

Supervised learning involves training a model to learn the mapping between input data and corresponding output labels based on training examples [179, 181]. One of the distinguishing characteristics of supervised learning is the presence of labeled data, which provides explicit feedback to the model about the correctness of its predictions. This feedback mechanism allows the model to learn from its mistakes and gradually improve its performance over time. Subsequently, the goal of supervised learning is to enable the model to generalize its learning from the training data to make accurate predictions or classifications on unseen or new data.

Unlike supervised learning, where the model learns from labeled examples to predict or classify new data, unsupervised learning algorithms aim to uncover underlying patterns, structures, or relationships within the data without explicit guidance. In unsupervised learning, the model is presented with a dataset consisting only of input data, without corresponding output labels [181, 182]. The goal is to

extract meaningful insights, discover hidden patterns, or identify intrinsic structures in the data. This process is often described as "learning without a teacher," as the model must autonomously infer the underlying structure of the data based solely on the input features [183]. Overall, unsupervised learning plays a crucial role in exploratory data analysis, pattern recognition, and feature extraction tasks, enabling the discovery of hidden structures and insights within unlabeled datasets.

Reinforcement learning is a powerful paradigm in machine learning that enables agents to learn optimal behavior by interacting with an environment and receiving feedback in the form of rewards or penalties. Unlike supervised and unsupervised learning, where the model learns from labeled or unlabeled data, reinforcement learning is based on trial-and-error learning, where the agent learns to make sequential decisions to maximize cumulative rewards over time [179, 184–186]. In reinforcement learning, the agent operates within an environment, which can be any system with defined states, actions, and rewards. The goal of the agent is to learn a policy—a mapping from states to actions—that maximizes the cumulative reward it receives over time. At each time step, the agent observes the current state of the environment, selects an action to perform, and receives feedback from the environment in the form of a reward signal [184–186].

When dealing with prediction tasks where the objective is to predict a label or outcome, as seen with our case study for COVID-19 diagnosis, supervised learning and reinforcement learning are viable approaches. In supervised learning, the model learns from labeled data to make predictions, whereas in reinforcement learning, the agent learns optimal actions through trial and error to maximize rewards and come to the correct prediction. However, unsupervised learning is not suitable for such prediction tasks as it isn't typically used for making predictions or decisions based on labeled outcomes.

In this chapter, we focus on supervised learning, as this is where the majority of bias mitigation techniques have been developed. Here, algorithms are trained on a dataset containing individuals' features (including vital signs and blood tests) labeled with their COVID-19 status (positive or negative), as discussed in previous chapters.

It should be noted that previous studies have primarily focused on debiasing binary sensitive attributes [76, 80, 101, 102, 176]. However, many real-world scenarios necessitate preserving a higher level of granularity. Binning may not accurately represent biological distinctions and can heavily reflect biases present in the sample population. Therefore, in our study, where we aim to address bias across four hospital sites, we seek to promote and illustrate the efficacy of bias mitigation methods across a broader spectrum of prediction tasks and demographic characteristics.

4.1.3 Bias Mitigation Methods

In this section, we introduce commonly employed fairness-aware algorithms, specifically focusing on in-processing bias mitigation techniques. These methods aim to address bias during the algorithm's training phase. In general, in-processing bias mitigation techniques, such as regularization/constraints [77, 103, 113, 187], cost-adjusted weighting [80, 81, 96, 114], and adversarial debiasing [76, 80, 81, 174], have emerged as promising strategies for addressing biases inherent in the learning process. Compositional techniques [188–191], which involve training multiple classification models independently for each sensitive class, also fall under the category of in-processing bias mitigation methods. However, for the purposes of this thesis, we focus on bias mitigation within a single algorithm, thus we exclude compositional methods from our discussion.

Regularization and Constraints

Regularization and constraints are essential techniques used to modify the loss function of a learning algorithm, allowing for the incorporation of fairness considerations into the training process. Regularization involves adding an additional term to the loss function, which penalizes discriminatory behavior exhibited by the machine learning algorithm [80, 81, 113, 192–194]. While the primary loss function typically focuses on optimizing classification accuracy metrics, the regularization term aims to discourage the model from making decisions that result in unfair outcomes.

For instance, studies have demonstrated that incorporating a discrimination-aware regularization term into the learning objective can effectively diminish gender bias in income prediction compared to conventional methods [193]. Similarly, research has shown promising results with a convex framework where fairness is enforced by the regularizer [115], as well as with a regularizer inspired by the Hilbert-Schmidt Independence Criteria, a statistical metric used to gauge the independence between two variables in a dataset [195, 196]. Recently, a study proposed a regularization technique aimed at refining feature representations of sensitive subgroups to enhance fairness in machine learning models [113].

Constraints, on the other hand, establish predefined bias thresholds based on the loss functions that must not be exceeded during training. Unlike regularization, which imposes penalties for discriminatory patterns indirectly through the loss function, constraints directly enforce specific levels of bias tolerance throughout the training process, thereby limiting a model's ability to discriminate based on sensitive attributes.

With respect to constraints, a recent study introduced a differential privacy mechanism that dynamically adjusts instance influence in each class based on theoretical bias-variance bounds [194]. The authors demonstrated that by incorporating differential privacy into the training process, the model becomes more robust to biases present in the data and demonstrated improved fairness across various benchmark datasets and scenarios, ranging from text to vision tasks. Furthermore, another research effort proposed a method using a Coefficient of Determination, which measures the predictive power of features to a target variable, as a constraint. This approach allows for rigorous control of fairness by treating the level of fairness as an explicit constraint [197].

Cost-Adjusted Weighting

In cost-adjusted weighting for bias-mitigation, the weight assigned to each instance in the dataset is determined based on its sensitive attribute [80, 81, 96, 114]. This weighting strategy aims to address the imbalance in the dataset by assigning

higher weights to instances belonging to sensitive and/or rare groups. For example, instances from an underrepresented sensitive group are given higher weights due to their relative scarcity in the dataset. In our investigations, where the sensitive attribute pertains to the hospital of patient presentation/attendance, higher weights would be designated to patient samples from BH, given its lower representation in the dataset.

During the training phase of classification models, these instance weights play a crucial role. Instances with higher weights contribute more to the loss function, meaning that misclassifying them incurs a greater penalty. This mechanism ensures that the model pays more attention to the underrepresented or disadvantaged groups during training, thereby improving its performance in predicting outcomes for these groups.

It is worth noting that while cost-adjusted weighting is often categorized as a preprocessing method, it can also be considered an in-processing technique. This distinction arises from the fact that cost-adjusted weighting operates within the training process of the machine learning algorithm itself. By modifying the loss function through the assignment of varied weights to examples, it directly influences how the algorithm updates its parameters to minimize loss. Importantly, this adjustment occurs without altering the actual training set used for model development, making it an integral part of the training process rather than a separate preprocessing step. This is also distinct to resampling, including upsampling and downsampling, whereby the frequencies of samples in the training set are directly altered.

Prior research has showcased the efficacy of reweighing techniques in addressing biased labels within the initial training dataset [73, 96, 114, 198]. Additionally, in another investigation, instead of assigning uniform weights to data instances from identical population subgroups, researchers assigned personalized weights to each instance in the training dataset [199].

Adversarial Debiasing

Unlike the previously introduced bias mitigation methods which operate on one task network, adversarial debiasing seeks to rectify the presence of biases by simultaneously learning to optimize a primary task network, while being adversarially trained to mitigate biases using another network [76, 80, 81, 174].

At its core, this technique operates on the principle of adversarial learning, a concept inspired by game theory. The model comprises two components: a primary task network and a fairness adversary. The primary task network learns to perform the intended prediction or classification task, such as diagnosing a disease, while the fairness adversary aims to detect and mitigate biases within the model's decision-making process. In the case study presented in this thesis, the primary task is diagnosing COVID-19, and the aim is to counteract any biases that may stem from the specific hospital attended by a patient.

The process involves three key components - an adversary network, a primary task model, and an objective function.

The adversary network is trained to predict the sensitive attribute (which in our case is the hospital attended by a patient) from the representations learned by the primary task model. During training, the adversary network strategically perturbs the primary model's parameters to combat discriminatory patterns and biases in the dataset which can be learned by the primary task model. Essentially, it functions as a "debiasing" mechanism by identifying and mitigating the impact of sensitive attributes on the primary task.

The primary task model encompasses the standard machine learning model designed to perform the primary task, such as classification or regression. This network strives to minimize its prediction error while simultaneously thwarting the adversarial attacks from the fairness adversary. Here, COVID-19 prediction is a classification task, whereby the primary task model predicts whether an individual is COVID-19 positive or negative based on the chosen clinical features.

Lastly, the training process revolves around optimizing a composite objective function that balances the primary task's performance with bias minimization.

Typically, this objective function integrates the loss function associated with the primary task model and a term penalizing the adversary network's ability to predict the sensitive attribute. By jointly optimizing these components via adversarial interplay, the primary task network is encouraged to learn representations that are both predictive and fair, thereby diminishing the influence of biased factors in decision-making.

Overall, adversarial debiasing has emerged as one of the most used in-processing bias mitigation techniques [80], and has been applied in various domains with the goal of promoting more equitable outcomes across diverse populations. It has previously been shown to be successful in reducing gender (male versus female) bias in salary prediction [76, 102] and ethnicity (black vs white) bias in recidivism prediction [101, 176].

4.1.4 Fairness Metrics

Fairness metrics serve as crucial components of bias mitigation strategies due to their pivotal role in assessing the presence of bias within machine learning models [80]. These metrics provide quantifiable measures to evaluate the fairness of algorithmic decision-making processes across different demographic groups or sensitive attributes. By systematically measuring the disparities in model predictions or outcomes, fairness metrics enable researchers and practitioners to identify areas of bias and develop targeted interventions to mitigate them.

Moreover, fairness metrics facilitate the comparison and selection of appropriate bias mitigation techniques by providing objective criteria for evaluating their effectiveness. By quantifying the degree of fairness achieved before and after applying mitigation methods, these metrics allow for evidence-based decision-making in the development and deployment of machine learning systems.

In this section, we primarily focus on *group fairness* metrics, which evaluate disparities between predefined subpopulations (e.g., based on race, gender, or institutional affiliation). While alternative notions of fairness exist, group fairness

remains the most widely studied and applied in healthcare machine learning, providing interpretable criteria for assessing and mitigating bias.

Recent advancements in fairness literature have introduced a diverse range of fairness metrics, each focusing on distinct aspects of classification performance [80, 81, 200]. These metrics offer nuanced insights into various dimensions of bias, enabling a comprehensive evaluation of model fairness.

In categorical prediction tasks, like binary classification of COVID-19 status, fairness metrics evaluating outcome fairness generally categorize into two groups according to their definitions: those focusing solely on the predicted outcome and those incorporating both the predicted and actual outcomes. Additionally, there exist fairness metrics derived from dataset labels, predicted probabilities and actual outcomes, similarity, and causal reasoning; however, these metrics are beyond the scope of this thesis.

To provide equations of fairness metrics, we use the following notation:

- Z: The sensitive attribute
- Y: The actual (ground-truth) label
- $\hat{Y}$: The predicted label

Metrics Based on the Predicted Label

Metrics based on the predicted outcome, often referred to as "parity-based" metrics, focus on assessing whether different demographic groups or sensitive attributes receive equitable treatment in the model's predictions [200, 201]. Unlike other fairness metrics that rely on additional information such as actual labels, parity-based metrics only necessitate knowledge of the predicted classification outcomes [80].

Parity-based metrics are particularly useful in scenarios where access to actual labels may be limited or unavailable. Their reliance solely on predicted outcomes makes them practical and applicable in a wide range of real-world settings, facilitating the assessment of fairness without compromising privacy or data confidentiality.

A commonly employed metric for gauging fairness based on predicted outcomes is Demographic Parity [76, 77, 80, 202–205]. This is a principle of fairness which asserts that the distribution of outcomes (such as positive predictions or decisions) should be equal across different demographic groups or sensitive attributes, regardless of their membership in these groups. In other words, demographic parity implies that individuals from different demographic groups should have an equal chance of receiving a positive outcome. Mathematically, demographic parity states that $P(\hat{Y} = 1|Z = 0) = P(\hat{Y} = 1|Z = 1)$. Thus, the metric can be calculated by determining the difference between subgroups, which is known as Statistical Parity Difference (SPD) [80], or by computing their ratio, which is known as Disparate Impact (DI) [80, 82, 206].

$$M_{SPD} = P(\hat{Y} = 1|Z = 0) - P(\hat{Y} = 1|Z = 1) \tag{4.1}$$

$$M_{DI} = \frac{P(\hat{Y} = 1|Z = 0)}{P(\hat{Y} = 1|Z = 1)} \tag{4.2}$$

Like Disparate Impact, the P-rule assesses two ratios of positive labels ($\frac{group_1}{group_2}, \frac{group_2}{group_1}$), and selects the minimum value from these ratios.

$$M_{P-rule} = min\left(\frac{P(\hat{Y} = 1|Z = 0)}{P(\hat{Y} = 1|Z = 1)}, \frac{P(\hat{Y} = 1|Z = 1)}{P(\hat{Y} = 1|Z = 0)}\right) \tag{4.3}$$

Ideally, we want to achieve M_{SPD}=0, M_{DI}=1, and M_{P-rule}=1.

Metrics Based on Predicted and Actual Labels

Definitions considering both predicted and actual outcomes are used to assess the predictive accuracy across different subgroups (e.g., does the classification model exhibit a tendency to make errors more frequently when dealing with underrepresented groups?) [80, 204]. These metrics entail comparing the rates of classification accuracy between different groups. Often, these are derived from combinations of true positive rate (TPR), false positive rate (FPR), true negative rate (TNR), and false negative rate (FNR).

$$TPR = \frac{TP}{TP + FN} \tag{4.4}$$

$$FPR = \frac{FP}{FP + TN} \tag{4.5}$$

$$TNR = \frac{TN}{TN + FP} \tag{4.6}$$

$$FNR = \frac{FN}{FN + TP} \tag{4.7}$$

Some commonly used metrics within this category include Equality of Opportunity, Predictive Equality, and Equalized Odds [78, 203–205, 207].

Equality of Opportunity describes the difference in TPR across subgroups [78, 80], whilst Predictive Equality describes the difference in FPR across subgroups [204, 207]. The equations used to calculate these metrics are as follows:

$$\text{Equality of Opportunity} = M_{EO(TP)} = TPR_{z=0} - TPR_{z=1} \tag{4.8}$$

$$\text{Predictive Equality} = M_{EO(FP)} = FPR_{z=0} - FPR_{z=1} \tag{4.9}$$

In the ideal case where these metrics take the value 0, we say that Equality of Opportunity is satisfied and Predictive Equality is satisfied. Equality of Opportunity is often referred to as the true positive Equalized Odds, or $M_{EO(TP)}$, and Predictive Equality is often referred to as the false positive Equalized Odds, or $M_{EO(FP)}$.

We say that Equalized Odds is satisfied when the true positive rate and the false positive rate are equivalent across all subgroups. Thus, Equalized Odds being satisfied requires that both Equality of Opportunity and Predictive Equality are satisfied. In other words, the model should make predictions equally well for all groups, regardless of their demographic characteristics. Formally, this states that a classifier is fair if $\hat{Y}$ and Z are conditionally independent given Y [72, 73, 76]. For binary classification, this is equivalent to $P(\hat{Y} = 1|Y = y, Z = 0) = P(\hat{Y} = 1|Y = y, Z = 1)$, with $y \in \{0, 1\}$.

Another similar metric is Average Odds Difference (AOD), which is determined by the average score of Equality of Opportunity and Predictive Equality [80].

$$M_{AOD} = \frac{1}{2}\Big(\text{Equality of Opportunity} + \text{Predictive Equality}\Big) \tag{4.10}$$

Limitations of Group Fairness and Alternative Fairness Concepts

While group fairness metrics such as demographic parity and equalized odds are widely adopted, they also present important limitations. First, they assume that sensitive attributes (e.g., race, gender, institutional affiliation) are clearly defined and reliably recorded. In practice, these attributes are often socially constructed, may be poorly documented, or entirely absent from datasets. Moreover, group fairness frameworks generally require a priori knowledge of which subgroups are relevant to fairness assessments, which risks overlooking underrepresented or intersectional populations[208].

Another challenge lies in the potential tension between group fairness metrics and other desirable goals. For example, enforcing demographic parity may reduce overall predictive accuracy or inadvertently mask disparities in error rates across groups [209, 210]. Furthermore, satisfying one group fairness criterion can make it impossible to satisfy another simultaneously, reflecting the well-known trade-offs highlighted in the fairness literature [211].

In response to these limitations, several alternative fairness concepts have been proposed. *Fairness through unawareness* stipulates that sensitive attributes should not be used in the prediction process, although this approach is often insufficient because bias can still enter through correlated variables [212, 213]. *Counterfactual fairness* defines fairness in terms of individual-level comparisons: a decision is fair if the predicted outcome would remain unchanged had the individual's sensitive attribute been different in a counterfactual world [214, 215]. Other perspectives, such as *equity*-based approaches emphasize allocating resources in a way that accounts for pre-existing disparities rather than simply treating groups identically [216].

Taken together, these considerations highlight that fairness is a multifaceted and contested concept, with no single definition universally applicable across domains. This thesis primarily evaluates group fairness metrics due to their interpretability and widespread use in healthcare ML research, while acknowledging the broader landscape of fairness definitions.

4.2 Methods

4.2.1 Baseline Model Comparators

In order to evaluate the influence of incorporating a bias-mitigating element during model training, we start by training two baseline models: XGBoost and a fully-connected neural network. These baseline models serve as benchmarks against which we can gauge the efficacy of subsequent fairness-aware models. By establishing these baselines, we gain an understanding of model performance without any bias mitigation techniques applied. This comparison enables us to assess the effectiveness of our proposed bias-mitigating strategies, examining their performance in predictive accuracy and their potential contribution to enhancing model fairness.

XGBoost

The XGBoost models previously trained and used in Chapter 2.5.1 serve as valuable benchmarks for evaluating COVID-19 prediction performance [133, 144]. In this section, we aim to provide a concise overview of the core principles of XGBoost, rather than an exhaustive analysis, as we primarily use XGBoost as a benchmark model. For a detailed description of its full architecture, readers can refer to the original publication [217].

XGBoost is a highly effective ensemble learning method that has garnered widespread recognition for its performance in a range of machine learning tasks [217]. It operates within the gradient boosting framework, where weak learners, typically decision trees, are combined sequentially to correct errors from previous iterations, creating a strong predictive model [218]. By optimizing an objective function that balances accuracy and simplicity - through the use of a loss function to measure prediction errors and regularization terms to control model complexity - XGBoost achieves high performance while mitigating overfitting [217].

During each training iteration, XGBoost calculates the residuals (the differences between predicted and actual outcomes) along with their gradients and second-order gradients (the first and second derivatives of the loss function). A new decision tree is then built to predict these residuals, and its output is incorporated into the

ensemble, with its contribution modulated by a learning rate [217]. Additionally, the algorithm employs various system-level optimizations, including parallel processing to enhance computational speed, sparsity-aware split finding for efficient handling of sparse data (a feature particularly useful for clinical records), and out-of-core computing to process large datasets that exceed memory limitations.

The final XGBoost model is composed of an ensemble of decision trees whose outputs are combined to generate predictions. For regression tasks, these outputs are summed, while for classification tasks (as addressed in this thesis), they are converted into probabilities using a logistic function. XGBoost's flexibility, scalability, and customizable hyperparameters (e.g., tree depth, learning rate, and number of estimators) make it highly effective for structured data tasks such as classification and regression. Renowned for its cutting-edge performance, XGBoost has become a favored tool in both academic research and real-world applications, as well as a staple in machine learning competitions [217].

As a reliable and efficient benchmark model, XGBoost serves as a valuable reference point for evaluating the performance of newer classifiers, including those based on neural network architectures, which are the primary focus of this thesis.

Standard Neural Network

A standard neural network forms the fundamental building block of deep learning models, serving as the basis for evaluating various bias mitigation techniques. By developing a predictor network without any debiasing components, we create a baseline to measure the relative impact of adding bias mitigation methods and assess their effectiveness.

For this purpose, we train a fully-connected neural network featuring rectified linear unit (ReLU) activation functions in the hidden layers and the sigmoid activation function in the output layer. This architecture is selected due to its ability to capture intricate patterns and correlations within the data. Throughout the training phase, model weights are iteratively adjusted using binary cross-entropy

(CE) loss and the Adam optimizer, commonly used for binary classification tasks like COVID-19 classification [219].

Full details of the architecture can be found in Table B.1 in Appendix B.

4.2.2 Cost-Adjusted Weighting

The initial bias mitigation technique under assessment is cost-adjusted weighting, applied in accordance with the sensitive attribute, which in this case is the hospital attended by each patient. In implementing this technique, weights are allocated to each patient sample based on the hospital attended. Employing a straightforward weighting principle, the weight assigned to each sample is inversely proportional to the frequency of the subgroup within the training dataset. We implement this weighting for both XGBoost and standard neural network models.

4.2.3 Regularization

Regarding regularization-/constraints-based methods, we opt for a consistency regularization technique [113, 220] to enhance fairness outcomes and consequently mitigate bias. In a recent study, authors introduced a bias mitigation approach which uses the concept of semantic-preserving augmentations at both image and feature levels, integrated within a self-consistency framework aimed at gender classification [113]. The core premise posited by the authors suggests that refining the feature representation for each subgroup could be instrumental in enhancing fairness, without compromising performance across different subgroups. This aligns with other studies which have indicated that enhancing feature representation not only improves classifier generalizability, but also reduces variance within the most underrepresented groups, thereby effectively mitigating bias [113, 221]. By using a gender-annotated facial image dataset, the study observed that employing a consistency regularizer during mitigation not only boosted overall gender classification accuracy but also lessened bias across all gender-racial categories compared to other state-of-the-art bias mitigation methods [113].

Based on insights from this research, our methodology also includes using a consistency regularizer during the training process. This is aimed at refining the feature representation of clinical data for COVID-19 screening, consequently minimizing bias across various groups. The primary objective is twofold: firstly, to diminish variability among underrepresented subgroups, particularly patient samples from distinct hospital sites, and secondly, to address bias in model outcomes.

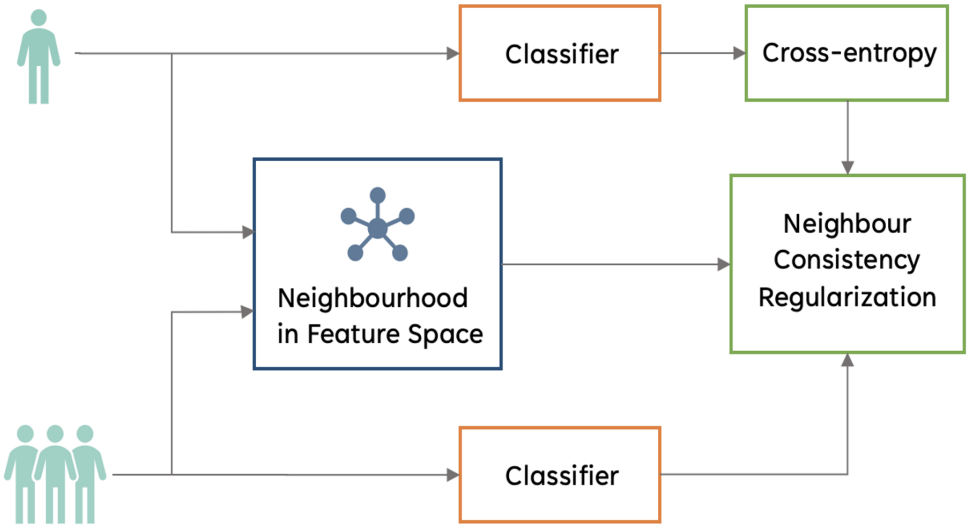

Figure 4.1: Neighbour consistency regularization.

In our specific implementation, we adopt Neighbour Consistency Regularization (NCR) [220], a regularization method predicated on the principle that instances belonging to the same class should exhibit comparable latent representations, regardless of potentially noisy or disparate labels. This concept holds particular relevance in the domain of COVID-19 screening, where diagnostic tools do not reach 100% sensitivity and specificity, however their outcomes are still relied on as ground-truth training labels during model development.

We define the similarity between two examples by the cosine similarity of their feature representations [220]:

$$s_{i,j} = \cos(v_i, v_j) = \frac{v_i^T v_j}{\|v_i\| \, \|v_j\|} \tag{4.11}$$

Here, cosine similarity is bounded between [0,1], and the feature representations are non-negative values (obtained after a ReLU transformation) from a specific

hidden layer. If v_i and v_j have high cosine similarity $s_{i,j}$, then a classifier f, is encouraged to predict the same label for $f(v_i)$ and $f(v_j)$, regardless of their true labels y_i and y_j. This discourages the model from overfitting to any incorrect mapping, if either (or both) of y_i and y_j are noisy.

To enforce NCR, the objective function is formulated to minimize the distance between logits z_i and z_j, when their corresponding feature representations v_i and v_j are similar. Using Equation 4.11, the consistency regularizer can be written as:

$$L_{NCR} := \frac{1}{m}\sum_{i=1}^{m} D_{KL}\left(\sigma(z_i)\middle\| \sum_{j\in NN_k(v_i)} \frac{s_{i,j}}{\sum\limits_k s_{i,k}}\sigma(z_j)\right) \tag{4.12}$$

In this formulation, D_{KL} represents the Kullback-Leibler (KL) divergence loss used to measure the dissimilarity between two distributions. The term $NN_k(v_i)$ refers to the set of k nearest neighbours of v_i in the feature space.

To ensure that the similarity values form a probability distribution, we normalize them. Additionally, we set the self-similarity $s_{i,i}$ to zero to avoid it dominating the normalized similarity. Gradients are propagated back to all inputs.

Thus, the NCR loss term encourages the output of a classifier to classify x_i in a way which aligns to its latent space neighbours, regardless of the potentially noisy label y_i. Regarding the COVID-19 screening task, this regularizer prompts the classifier to categorize a patient sample in a manner that closely aligns with similar samples, irrespective of the recorded diagnosis label.

We combine this regularizer with the standard supervised classification loss function, namely cross-entropy, to form the final objective function that is minimized during training. Thus, the final loss function is:

$$L_{total} := L_{CE} + \alpha L_{NCR}, \tag{4.13}$$

Here, the hyperparameter α controls the strength of the NCR term. This differs slightly from the implementation in [220], where the authors implement a linear interpolation between the NCR and CE contributions (however, both implementations adjust the relative contributions of the CE and NCR loss terms).

Therefore, through the incorporation of this consistency regularizer, our objective is to enhance the feature representation of clinical attributes and diminish the variability among patient samples from different hospital sites, thereby alleviating any potential bias.

In addition to KL divergence, during hyperparamter optimization, we also evaluated the Jensen-Shannon Divergence and Mean Absolute Error losses. The equations for these loss functions can be found in Appendix B.

4.2.4 Adversarial Debiasing

Finally, we evaluate adversarial debiasing. The adversarial debiasing architecture we implement consists of two individual networks – a primary task (predictor) network, P, and an adversary network, A, as shown in Figure 4.2. P and A are each a multilayer perceptron – the simplest form of a neural network. Here, P is trained to predict COVID-19 status, Y, given a set of clinical features, without being biased by Z (the sensitive feature). For our purposes, Z is the hospital location.

Because we are training a classifier, P, to accurately predict Y while satisfying an equality constraint, namely, Equalized Odds, defined in Section 4.2.5, we must consider this in our training of an adversary model. As previously mentioned, Equalized Odds states that a classifier, P, is fair if $\hat{Y}$ and Z are conditionally independent given Y. Following this definition, we provide the adversary model, A, access to both the true label, Y, and the predicted label, $\hat{Y}$; thus, limiting the information provided to A to those features contained in the definition of Equalized Odds (which is defined using both the true and predicted labels). In other words, the classifier's raw output, $\hat{Y}$ – the predicted probability score, and the true label, Y, are used as the input to A, which tries to predict Z. Although we chose to use Equalized Odds, this method can be extended to other definitions as well. For example, if one wanted to train a classifier to satisfy demographic parity (which states that a classifier is fair if $\hat{Y}$ and Z are independent), the adversary would be trained to predict Z, solely given $\hat{Y}$.

Our goal is to train P to predict Y effectively, regardless of the demographic membership of Z. Thus, we want P to be able to accurately predict Y, and A to poorly predict Z, as this suggests that P has been trained in such a way that debiases $\hat{Y}$ with respect to Z. We use cross-entropy loss (and binary cross-entropy loss when the feature is binary), where L_P represents the loss for P , and L_A represents the loss for A.

For P to be good at predicting Y while being unbiased towards Z, P is typically trained to balance the trade-off between the two losses. This is achieved using the combined loss function:

$$L = L_P - \alpha L_A, \tag{4.14}$$

where α is an adjustable hyperparameter that signifies the importance of debiasing with respect to the sensitive feature, z. This combined function encourages P to minimize L_P while maximizing L_A. However, to ensure that P propagates in the correct direction at the beginning of training, we modified the combined loss function to include an attenuating correction term, such that the loss function for P becomes:

$$L = L_P + \frac{L_P}{L_A} - \alpha L_A \tag{4.15}$$

Under the assumption that L_A starts small (A is able to accurately predict z at the beginning of training), the correction term, $\frac{L_P}{L_A}$, ensures that L is large at the beginning of the training process (as $L_P + \frac{L_P}{L_A}$ is large), encouraging the adversary to increase its loss, L_A. During training, as L_P becomes smaller (P becomes better at predicting y) and L_A becomes larger (A becomes unable to accurately predict z), $\frac{L_P}{L_A} \to 0$, converging to the original adversarial loss function, $L_P - \alpha L_A$.

For P, the sigmoid activation function is used in the output layer (since COVID-19 prediction is a binary task); and for A, the softmax activation function is used instead (as the prediction of the label of Z is a multiclass task).

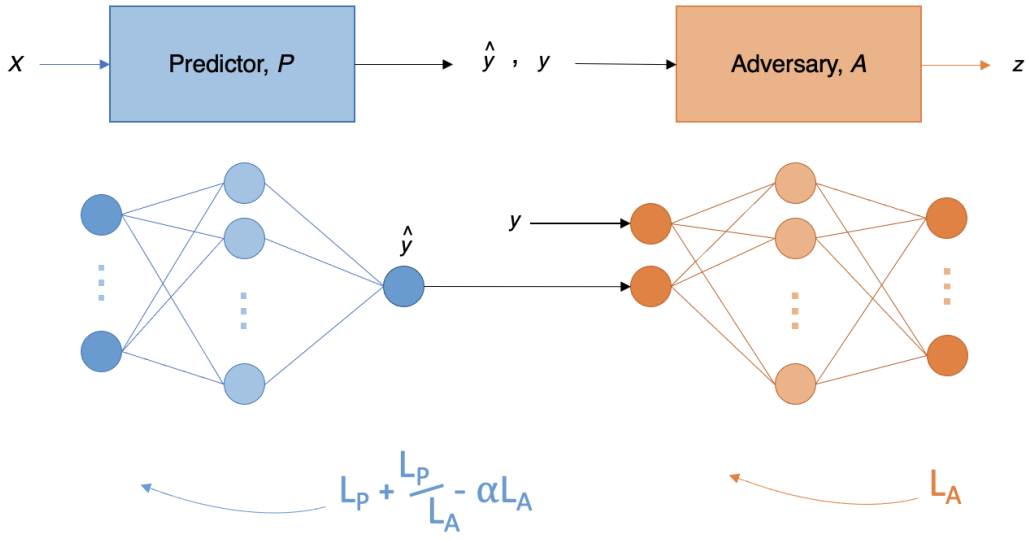

Figure 4.2: Adversarial debiasing architecture.

4.2.5 Evaluation Metrics

To evaluate the performance of COVID-19 prediction, we report sensitivity, specificity, PPV, NPV, and AUROC. It should be noted that PPV and NPV are prevalence-dependent, and thus, they are calculated based on the actual prevalence within each test set. These metrics are accompanied by 95% confidence intervals (CIs), computed using 1000 bootstrapped samples drawn from the test set. Tests of significance, indicated by p-values, involve evaluating how often one model outperforms another across 1000 pairs of bootstrapped iterations. We use 0.05 as the threshold value for determining statistical significance. Results are based on the evaluation of final, held-out test sets.

In our framework, our primary aim is to develop models that demonstrate fairness concerning sensitive attributes. Thus, to evaluate fairness, we employ Equalized Odds. We choose this metric because it considers both predicted and actual outcomes, allowing us to evaluate predictive accuracy across various subgroups effectively. Additionally, Equalized Odds offers a comprehensive evaluation of both true positive (TP) and false positive (FP) rates, which are both important with respect to COVID-19 diagnosis.

Thus, evaluation based on Equalized Odds ensures that our models achieve precise and reliable detection of COVID-19 cases while minimizing the occurrence of false alarms and their subsequent implications. Striking a balance between TP and FP rates is crucial for effective public health management and the equitable distribution of resources during a pandemic. Therefore, Equalized Odds serves as a critical measure for assessing the fairness and effectiveness of our models in supporting COVID-19 screening efforts.

Recall that by definition, Equalized Odds is defined based on two subgroups (a sensitive group and a non-sensitive complement), and is measured based on the difference between TP rate and FP rate between these groups. However, in this work, there are more than two subgroups within the sensitive category. Specifically, there are four different hospitals. Thus, to assess multiple labels (i.e., >2), we used the standard deviation (SD) of true positive and false positive scores. Thus, SD scores closer to zero suggest greater outcome fairness. The equations used to calculate TP and FP SD scores are as follows:

$$\begin{aligned} M_{EO(TP)} =& SD\Bigg(\Big\{P(\hat{Y}=1|Y=1,Z=z_i), P(\hat{Y}=1|Y=1,Z=z_{i+1}), \\ & ..., P(\hat{Y}=1|Y=1,Z=z_N)\Big\}\Bigg) \\ =& SD\left(\left\{\frac{TP_i}{TP_i+FN_i}, \frac{TP_{i+1}}{TP_{i+1}+FN_{i+1}}, \ldots, \frac{TP_N}{TP_N+FN_N}\right\}\right), \end{aligned} \tag{4.16}$$

$$\begin{aligned} M_{EO(FP)} =& SD\Bigg(\Big\{P(\hat{Y}=1|Y=0,Z=z_i), P(\hat{Y}=1|Y=0,Z=z_{i+1}), \\ & ..., P(\hat{Y}=1|Y=0,Z=z_N)\Big\}\Bigg) \\ =& SD\left(\left\{\frac{FP_i}{TP_i+FN_i}, \frac{FP_{i+1}}{TP_{i+1}+FN_{i+1}}, \ldots, \frac{FP_N}{TP_N+FN_N}\right\}\right) \end{aligned} \tag{4.17}$$

4.2.6 Hyperparameter Optimization

In the process of determining hyperparameter values, we employed grid search in conjunction with standard five-fold cross-validation, using the training set. This method involves systematically evaluating various combinations of hyperparameters

across different ranges to identify the optimal configuration for our models. Through this iterative process, we aimed to find the set of hyperparameters that yielded the best performance on the training data while avoiding overfitting. By using cross-validation, we ensured robustness in our hyperparameter selection process, as it allowed us to assess the generalizability of our models across multiple subsets of the training data.

Grid search was used for all neural network models to identify the optimal configurations, including the number of hidden layers, the number of nodes within each layer, and the learning rate. In the context of adversarial debiasing, this search was conducted separately for both the predictor and adversary networks, as well as for the α hyperparameter. For XGBoost, a range of parameters such as learning rate, depth, and the number of trees were tested. In the case of NCR, the evaluation involved assessing the number of hidden layers, the specific hidden layer used to compute the NCR loss, the chosen loss function, the number of nearest neighbours, the weight assigned to the NCR term, and the starting epoch for NCR.

Detailed information regarding the software used, implementation, and final hyperparameter values selected for each model can be found in Appendix B.

4.2.7 Threshold Optimization

It is important to note that many of the definitions and methods introduced thus far work with both regression and classification models. For our purposes, rather than working with a continuous probability score, we chose to perform thresholding and use a binary classification (COVID-19 positive or negative), to correspond to the "green–amber–blue" categorization system used by NHS Trust policy [132, 133] (detailed in Chapter 2). Here, green represented an illness with no symptoms of COVID-19, amber represented an illness with symptoms potentially characteristic of COVID-19, and blue represented a laboratory-confirmed COVID-19 infection. Therefore, using classification is consistent with performing rapid triage into either a green or amber pathway.

For classification tasks, the raw output of a machine learning algorithm is typically a probability score indicating the likelihood that an instance belongs to a particular class. This probability is then mapped to a discrete class label using a predefined threshold (example shown in Figure 4.3). In binary classification tasks, the default threshold is often set at 0.5, meaning instances with predicted probabilities equal to or greater than 0.5 are assigned to one class, while those below 0.5 are assigned to the other [222]. While straightforward, this default threshold may not always yield optimal results, particularly in scenarios with significant class imbalances. For example, in our task, where COVID-19 negative cases vastly outnumber positive ones, relying on the default threshold risks poor sensitivity, potentially failing to detect a substantial number of positive cases.

To address this issue, we performed a grid search on the continuous validation set to optimize the threshold for prediction. This approach allowed us to systematically evaluate different threshold values and their impact on classification performance, particularly sensitivity and specificity. The goal was to identify a threshold that improves detection rates without excessively compromising other performance metrics.

For our specific application, we prioritized achieving a sensitivity of 0.9, as this ensures clinically acceptable performance in identifying positive COVID-19 cases. High sensitivity is crucial in this context to minimize the risk of missing positive cases, which could have serious public health implications. By optimizing the threshold, we tailored the model's predictions to the task's specific requirements and enhanced its clinical utility. This threshold adjustment underscores the importance of tuning model parameters to align with the practical and ethical considerations of real-world applications. This chosen threshold also exceeds the sensitivity of lateral flow device tests, which achieved a sensitivity of 56.9% (95% CI 51.7%-62.0%) for OUH admissions between December 23, 2021 and March 6, 2021 [133]. Additionally, as previously mentioned, the gold standard for COVID-19 diagnosis is by real-time PCR, which has estimated sensitivities of approximately 80%-90% [150, 151]. Therefore, setting the threshold at 0.9 enables the models to effectively identify COVID-19 positive cases while achieving sensitivities that surpass those

of current diagnostic tests (noting that the training datasets use PCR as the gold standard for COVID-19 diagnosis).

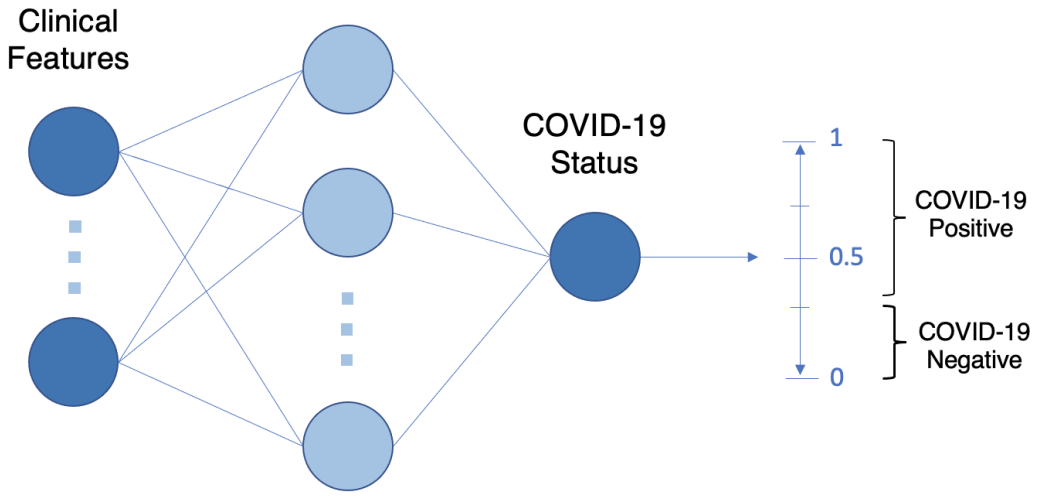

Figure 4.3: Figure shows neural network architecture used for classification, including a visual of the decision boundary used to determine COVID-19 positive or negative status.

4.3 Results

After hyperparameter optimization and model fitting, we evaluated all models using the held-out set. In terms of the AUROC metric, model performances exhibited relative consistency, with AUROC values ranging from 0.882 to 0.901. The weighted XGBoost model achieved the highest AUROC score of 0.901 (95% CI ±0.12), while the adversarial model achieved the lowest at 0.882 (±0.14), although only marginally lower. When using a sensitivity configuration of 0.9, we also observed

Table 4.1: Equalized Odds evaluation for hospital bias and COVID-19 status prediction test results across different models, optimized to sensitivities of 0.9. Bolded values denoting the the best (underlined) and second best Equalized Odds scores. Classification metrics are reported alongside 95% CIs, with bolded values denoting best scores achieved on the test set.

Model	$M_{EO(TP)}$	$M_{EO(FP)}$	Sensitivity	Specificity	PPV	NPV	AUROC
XGBoost	0.024	0.057	0.875 (± 0.020)	**0.720 (± 0.008)**	**0.196 (± 0.011)**	0.987 (± 0.003)	0.900 (± 0.12)
Neural Net.	**0.022**	0.065	0.879 (± 0.019)	0.676 (± 0.008)	0.175 (± 0.010)	0.986 (± 0.002)	0.891 (± 0.13)
Neural Net. (weighted)	0.033	0.055	0.883 (± 0.019)	0.686 (± 0.008)	0.180 (± 0.011)	0.987 (± 0.002)	0.894 (± 0.13)
XGBoost (weighted)	**0.014**	0.057	**0.892 (± 0.019)**	0.681 (± 0.008)	0.179 (± 0.01)	**0.988 (± 0.002)**	**0.901 (± 0.12)**
Neural Net. (reg.)	**0.022**	**0.046**	0.865 (± 0.021)	0.714 (± 0.008)	0.191 (± 0.012)	0.985 (± 0.003)	0.892 (± 0.13)
Adversarial	**0.022**	**0.045**	0.882 (± 0.019)	0.642 (± 0.008)	0.161 (± 0.009)	0.986 (± 0.003)	0.882 (± 0.14)

consistent sensitivity scores across all models, ranging from 0.865 to 0.892. The neural network regularized using NCR achieved the lowest score of 0.865 ($\pm$0.021), while the weighted XGBoost model achieved the highest score of 0.882 ($\pm$0.019). It is noteworthy that although the neural network regularized with NCR achieved slightly lower sensitivity, it attained the highest specificity of 0.714 ($\pm$0.008), highlighting the tradeoff between sensitivity and specificity metrics. All models demonstrated high prevalence-dependent NPV scores exceeding 0.985, indicating their capability to confidently exclude COVID-19 cases. Despite the apparent similarity in overall predictive performance among models, the difference in AUROC values was found to be statistically significant ($p < 0.0001$, based on 1,000 bootstrapped samples).

In terms of bias mitigation, the impact of cost-adjusted weighting on Equalized Odds varied across models. For the neural network, the cost-adjusted weighting led to a slight improvement in Equalized Odds concerning the false positive rate, but conversely worsened it with respect to the true positive rate. However, for XGBoost, the weighted XGBoost model showed improvement compared to the standard XGBoost baseline. It achieved the best Equalized Odds concerning the true positive rate across all methods, with a value of 0.014. Here, true positive rates across hospitals had a small range, with values of 0.888, 0.886, 0.875, and 0.850 for OUH, UHB, BH, and PUH, respectively. However, Equalized Odds for the false positive rate remained unchanged and overall ranked as the second worst across all methods, at 0.057, with false positive rates of 0.349, 0.264, 0.273, and 0.215, for OUH, UHB, BH, and PUH, respectively.

On the other hand, the neural network regularized using NCR obtained the second-best Equalized Odds for both true positive and false positive rates, with values of 0.022 and 0.046, respectively. It should be noted that the score for true positive rate is the same for the standard neural network without any bias mitigation technique. When looking at individual true positive and false positive rates across hospitals, the standard neural network generally achieved higher true positive rates than the NCR-regularized model (true positive rate range 0.863-0.920, compared to 0.849-0.906, respectively), however, the NCR-regularized model

generally showed improved false positive rates compared to the standard neural network (false positive rate range 0.207-0.338, compared to 0.192-0.372, respectively). Similarly, the adversarial model achieved a score of 0.022 for the true positive rate, while displaying slightly superior performance for the false positive rate, with a score of 0.045. Consequently, overall, the adversarial model demonstrated the fairest performance, with the neural network regularized with NCR achieving the second-best fairness performance. Full hospital subgroup analysis of true positive and false positive rates can be found in Tables C.2 and C.3 in the Appendix.

This emphasizes the significant improvement in Equalized Odds achieved by integrating some form of debiasing capability within models, with only a slight compromise in predictive performance. Specifically, when comparing the regularized neural network and adversarial models to both the neural network and the XGBoost implementations (with and without cost-adjusted weighting), there was a reduction in the AUROC score of less than 0.02.

4.4 Discussion

We observed that neural network-based methods (with and without bias mitigating components) consistently achieved AUROC scores comparable to those using XGBoost. One notable advantage of employing neural networks is their applicability to a wide range of problems beyond the scope of traditional methods like XGBoost. Neural networks are well-suited for tasks such as image recognition and Natural Language Processing (NLP), where XGBoost and similar algorithms may not be as effective. Furthermore, these methods can be adapted to different model architectures and incorporated into transfer learning frameworks to enhance model performance. Unlike tree-based algorithms like XGBoost, which rely on the entire dataset during training, neural networks offer greater flexibility for transfer learning, enabling the utilization of pre-trained models and using knowledge from related tasks or domains.

The bias mitigation strategies we applied in our study yielded generally improved outcomes compared to models without any bias mitigation component. Despite

this progress, our models fell short of fully satisfying the Equalized Odds criteria, as indicated by their failure to achieve TP and FP standard deviations of zero. One potential explanation for this discrepancy lies in the inherent imbalance present in our training datasets, particularly concerning sensitive features. Because we used neural network-based models, skewed distributions within our data can have a notable impact on classification outcomes. This observation aligns with findings from previous studies [102], which have highlighted the significant influence of data balance on the effectiveness of bias mitigation methods. Moving forward, it would be interesting for future experiments to address the imbalance in training data to enhance model performance and foster greater fairness in predictions. One possible approach involves using sampling techniques, such as over- or under-sampling, to create more balanced training datasets. However, it is important to consider the implications of altering the dataset's prevalence, especially in scenarios where maintaining the true sample population is crucial for accurate representation of real-world conditions. Thus, achieving a balance between data subgroup frequency and preserving the true prevalence is paramount in ensuring the effectiveness and applicability of bias mitigation techniques in practical settings.

Furthermore, our analysis revealed inconsistent performance of cost-adjusted weighting in terms of improving Equalized Odds across both neural network and XGBoost models. While it exhibited enhancements in either the true positive or true negative rate, it did not consistently improve both simultaneously. In the standard supervised learning setup observed in these models, the cross-entropy loss function provides a learning signal regardless of the input data, potentially leading to model skewness or bias based on the majority class within the batch. Despite the implementation of cost-adjusted weights to mitigate class imbalances to some extent, models still rely on the cross-entropy loss function. As a result, the effectiveness of using cost-sensitive weighting, which is based on sensitive attribute frequency, may have been constrained.

These observations suggest that while cost-adjusted weighting can partially address class imbalances, it may not fully address the underlying biases present in

the dataset, particularly when the loss function itself does not differentiate between different classes based on their significance or sensitivity. Therefore, alternative approaches or combinations of bias mitigation techniques may be necessary to achieve more consistent and effective improvements in model fairness.

In this chapter, we also observed that a neural network trained with both the NCR and the adversarial framework exhibited superior fairness performance compared to other models. Specifically, the model trained with the adversarial framework achieved the best fairness outcomes, while the one trained with the NCR regularizer performed the second best. Furthermore, despite the incorporation of bias mitigation components, both the NCR regularizer and adversarial framework models maintained consistent AUROC scores when compared to a standard neural network trained without any bias mitigation component and XGBoost, with very little trade-off in accuracy. Previous studies have acknowledged the trade-off between fairness and accuracy [86, 94], yet some research has suggested that it might be feasible to enhance fairness without sacrificing accuracy [223, 224], aligning with our findings. These results demonstrate promise in the effectiveness of bias mitigation techniques in advancing fairness within machine learning models.

It should also be noted that the techniques employed can alter decision boundaries and model outcomes, potentially impacting the perceived significance of features in classification. For instance, NCR promotes smoothness and consistency in predictions among neighboring data points. Consequently, features that contribute to stable predictions and smooth decision boundaries within local neighborhoods are likely to be prioritized by the model. However, overemphasizing feature significance within local neighborhoods may result in a focus on local rather than global feature importance.

In summary, the findings presented in this chapter emphasize the efficacy of implementing bias mitigation techniques, including the NCR regularizer and adversarial framework, to improve fairness while maintaining predictive performance in machine learning models, particularly within supervised learning contexts. Moving forward, the subsequent chapter introduces an innovative bias mitigation approach

within reinforcement learning, representing a novel contribution to bias reduction within a distinct machine learning paradigm. This expansion into reinforcement learning not only broadens the scope of bias mitigation efforts but also demonstrates a commitment to advancing fairness across various machine learning domains.

5

Bias Mitigation for Reinforcement Learning

5.1 Introduction

5.1.1 Overview

The existing literature on mitigating bias at the algorithmic level has focused on conventional supervised learning methodologies, as outlined in Chapter 4. These methodologies incorporate a range of techniques, including adversarial learning, the integration of regularization methods, and the imposition of constraints on a model's loss function. They have shown effectiveness in mitigating undesired biases by adapting the learning process to incorporate fairness considerations across various domains, including our investigations into COVID-19 screening.

Nevertheless, although these methods have proven effective, they might not comprehensively tackle the intricacies of bias mitigation, especially in situations requiring specialized approaches to fulfill specific tasks and attain desired fairness goals. Recognizing this gap, our aim is to create an innovative technique using a reinforcement learning framework to enhance fairness outcomes. Reinforcement learning was briefly introduced in the previous chapter; however, in this chapter, we will explore it in greater detail, as it serves as the primary focus.

Reinforcement learning constitutes a distinctive set of tasks that not only find relevance across numerous real-world scenarios but also diverge significantly from conventional machine learning domains. Typically, machine learning tasks are categorized as either supervised (where an input is mapped through a model to predict a class label) or unsupervised (aiming to discern patterns from unlabeled data). RL, however, stands apart by concentrating solely on maximizing rewards garnered from interactions with the environment [186, 225]. Although it has commonly been associated with successes in game playing and control, the core elements of RL have been shown to be successful on a wider range of tasks, including those which, on the surface, do not appear to have a particular "agent" interacting with an "environment". Such problems include classification tasks, which have commonly been addressed using standard supervised learning algorithms. RL instead, uses an agent to interact with the input to determine which class it belongs to, and then receives an immediate reward from its environment based on that prediction [226, 227]. This offers a unique framework for bias mitigation, as we can use a specialized reward function with the specific purpose of improving algorithmic fairness and mitigating unwanted biases. Thus, by formulating bias mitigation as a reinforcement learning problem, we can design agents that learn to make decisions that not only optimize predictive accuracy but also promote fairness and equity in outcomes.

This approach allows us to incorporate fairness considerations directly into the learning process, enabling the agent to adapt its behavior in response to feedback. By training agents to navigate decision-making tasks while explicitly mitigating biases associated with sensitive features, we can develop models that are inherently more robust and equitable across diverse populations.

This chapter aims to extend the current repertoire of bias mitigation techniques by introducing a reinforcement learning-based approach that prioritizes fairness outcomes. By using the strengths of reinforcement learning, we aim to develop more effective and adaptive methods for addressing bias in machine learning algorithms.

5.1.2 Reinforcement Learning

In this section, we offer a concise introduction to reinforcement learning, with a specific emphasis on Q-learning and deep Q-learning within the realm of classification. Variants of reinforcement learning beyond these are not covered within the scope of this thesis. Chapter 5.2 will delve deeper into the distinct iterations of Q-learning, methodologies, and implementation specifics used.

Recall that reinforcement learning is a machine learning paradigm where an agent learns to make decisions by interacting with an environment to achieve specific goals. The agent learns through trial and error, receiving feedback in the form of rewards or penalties based on its actions [179, 184–186]. We use discrete timesteps to consider each observation-action-reward cycle used within the RL framework. Thus, to train an agent for a task effectively, a reward value R is provided to the agent after it takes an action at each time step. A positive reward is assigned to the agent upon correct prediction of a label, and a negative one is allocated otherwise [227]. This feedback mechanism aids the agent in learning the optimal "behavior" to accurately classify samples, aiming to accumulate the highest possible rewards. To achieve this, the agent executes actions that establish memory cells, which are subsequently used by the agent, in conjunction with the original input, to determine actions and classify samples. The overall objective is to acquire a policy that maximizes the cumulative reward throughout an entire run of the training process, commonly referred to as an episode.

Formally, an RL environment can be defined as a sequential decision-making problem within a finite Markov Decision Process (MDP) framework [184, 225, 227], characterized by a tuple of five variables (s, a, r, p, γ). During the training phase, data batches are randomly shuffled and sequentially presented to the model, with each variable defined as follows:

- s: The current state, comprising the features of the current observation (sample) being processed.

- a: The subsequent action taken by an agent, used for selecting a classification label.
- r: The anticipated reward associated with a specific action, determined by the accuracy of the classification.
- p: The transition probability resulting from an action.
- γ: The discount rate applied to future rewards.

The agent engages with the environment in discrete time steps denoted as $t = 0, 1, ..., T$, where T represents the final time step of the episode. At each time step, the agent selects an action a_t, and the environment transitions to a new state according to the transition function $p(s_t, a_t) = s_{t+1}$. In our case, this transition is deterministic, as the agent progresses from one state to the next in accordance with the order of samples in the training data. Subsequently, the agent receives a reward $r_t = R(s_t, a_t)$.

The objective is to maximize the reward across all time steps, often referred to as the return. A simple approach involves defining this return as the summation of the sequence of subsequent rewards, $r_t + r_{t+1} + \cdots + r_T$.

Nevertheless, in certain tasks where the agent continually takes actions without reaching an end state, this method becomes inadequate, as the sum of all rewards will tend to infinity as $T \to \infty$. Additionally, this would hinder the agent's ability to differentiate actions that yield greater rewards more promptly. Hence, a discount factor denoted by γ, where $0 \leq \gamma \leq 1$, is employed to control the significance of future rewards in comparison to immediate rewards. Consequently, the total return [184] is recalculated as:

$$G_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots + \gamma^T r_{t+T} = \sum_{k=0}^{T} \gamma^k r_{t+k} \tag{5.1}$$

The rewards at each time step are contingent on both the environment's state and the action chosen by the agent. Therefore, an action-value function, $Q(s_t, a_t)$, is established to estimate the values associated with particular state-action pairs.

Commonly known as the Q-function, this function provides values termed as Q-values [184].

5.1.3 Q-Learning

Q-Learning is an iterative reinforcement learning algorithm aimed at constructing a look-up table of Q-values to represent the Q-function [184, 228]. The Bellman equation serves as the theoretical foundation for updating Q-values. The Q-value update equation in Q-learning is a form of the Bellman equation known as the Bellman optimality equation. In the context of Q-learning, it is used to update the Q-values based on observed experiences. The Bellman optimality equation for Q-learning is given by [229]:

$$Q(s_t, a_t) = \mathbb{E}\left[r_t + \gamma \max_{a'} Q(s', a')\right], \tag{5.2}$$

where the expectation is over possible state transitions. Here, $Q(s_t, a_t)$ denotes the Q-value for taking action a_t in state s_t. r_t is the immediate reward received after taking action a_t from state s_t, and to this immediate reward we add the product of the discount factor and $\max_{a'} Q(s', a')$, which represents the highest expected Q-value in the subsequent state for all possible actions a'. This function is formulated with the inherently greedy rationale that the optimal strategy for maximizing overall reward involves selecting actions that optimize the expected return at each time step. Consequently, numerous reinforcement learning algorithms aim to acquire a precise Q-function through their experience.

Q-learning is one such process which uses iterative updates to learn the true Q-value of each state-action pair. The Q-value update equation in Q-learning is:

$$Q(s_t, a_t) := Q(s_t, a_t) + \alpha\left(r_t + \gamma \max_{a'} Q(s', a') - Q(s_t, a_t)\right) \tag{5.3}$$

Here, $\alpha \in [0, 1]$ is the learning rate, which controls the extent to which new information overrides old information.

Subsequently, at each time step t, Q-learning executes the following steps:

1. Observe the current state s_t.

2. Select and execute action a_t.

3. Observe the resulting new state s'.

4. Receive the immediate reward r_t.

5. Update the Q-Table update equation 5.3.

These steps are repeated for each time step until the learning converges or reaches a predefined stopping criterion. The learning rate α in Q-learning modulates the influence of new experiences on updating the Q-values, striking a balance between incorporating new information and maintaining stability during the learning process. It plays a crucial role in controlling the convergence and performance of the Q-learning algorithm. Through iteratively updating Q-values based on observed rewards and transitions, Q-learning enables the agent to learn the optimal action-selection strategy for maximizing cumulative rewards over time.

In practice, adopting this basic approach is impractical, as it involves estimating the action-value function separately for each sample, lacking any form of generalization [230]. Moreover, it is important to highlight that implementing this method necessitates memory cells to store the Q-table, which may not be feasible for domains with extensive state spaces [231, 232]. To address this limitation, a function approximator for the Q-value is used instead of storing values in a table. This function approximator is often a linear function, although a nonlinear function approximator can also be used, such as a neural network. The use of a neural network as a function approximator forms the foundation of Deep Q-learning [230].

5.1.4 Deep Q-learning

Deep Q-learning (DQN) is a significant advancement in reinforcement learning, using deep neural networks to approximate Q-values [225, 230]. Unlike traditional tabular Q-learning, which struggles with high-dimensional state spaces, deep Q-learning efficiently handles complex inputs such as images or raw sensory data. This

capability is crucial for real-world applications like robotics, autonomous vehicles, and clinical data, where input data can be vast and intricate.

Moreover, deep Q-learning offers improved generalization across similar states [232]. By extracting relevant features from input data, neural networks can generalize their knowledge more effectively, enabling agents to learn optimal policies in unseen states. This scalability and flexibility make deep Q-learning highly adaptable to various reinforcement learning problems, including both discrete and continuous action spaces.

Deep Q-learning does not store a Q-table, but instead uses a neural network as a function approximator, where the parameterized Q-function, $Q(s, a; \theta) \approx Q(s, a)$, is used. Here, θ represents the parameters of the neural network. For training, we substitute the Q-Learning iterative update process with gradient descent, minimizing the mean-squared error loss function:

$$L(\theta) = (y_t - Q(s_t, a_t; \theta))^2 \tag{5.4}$$

$Q(s_t, a_t; \theta)$ is the network prediction and, as in standard supervised learning, y_t can be treated as the target to be predicted. We use the target of the Q-learning iterative update in Equation 5.3:

$$y_t = r_t + \gamma \max_{a'} Q(s', a'; \theta) \tag{5.5}$$

Overall, deep Q-learning represents a powerful combination of reinforcement learning and deep learning techniques, unlocking new possibilities for learning in high-dimensional and complex environments.

5.1.5 The Advantage Function

As well as the action-value function $Q(s_t, a_t)$, we can also introduce the state-value function $V(s_t) = \mathbb{E}[G_t|s_t]$ which represents the expected return G_t that can be achieved from the state s_t in the remaining timesteps. Here the expectation is over

the policy $\pi(a_t|s_t)$ of possible actions from this state. We can therefore write the state-value function in terms of the action-value function as:

$$V(s_t) = \sum_{a_t} \pi(a_t|s_t)(s_t, a_t)$$

We can then introduce the advantage function $A(s_t, a_t)$ to quantify the advantage of selecting each of the different actions:

$$A(s_t, a_t) = Q(s_t, a_t) - V(s_t)$$

$A(s_t, a_t)$ then indicates how much better taking a specific action a_t is compared to the average expected reward over all possible actions in that state s_t.

Finally, the Q-function can then be reconstructed using the value function and the advantage function:

$$Q(s_t, a_t) = V(s_t) + A(s_t, a_t)$$

The value function V can be viewed as a proxy for the "goodness" of a particular state, and the Q function evaluates the value of selecting a particular action in this state [233]. Thus, A can be interpreted as the relative importance of each action. By decomposing the Q-function into these two components, it becomes easier to analyze and understand the relative value of actions in different states.

The subsequent section will delve deeper into the specific techniques and implementations used in our proposed approach.

5.2 Method

5.2.1 Reinforcement Learning for Classification

To formulate classification as a reinforcement learning task, we start by modeling the COVID-19 diagnosis task in a sequential decision-making format using a finite MDP. We define the MDP using a tuple of five variables (s, a, p, r, γ), as defined in Section 5.1.2

With respect to the training dataset, we define it as size $N \times D$, where N is the total number of samples and D is the number of features in each sample.

During training, a batch of data is randomly shuffled and presented to the model in order. Here, p is deterministic, as the agent moves from one state to the next according to the order of samples in the training data, with the features of each sample making up the state, s.

The action, a, is the prediction the agent makes when presented with the state, s. Given a total number of classification labels, K, each action a is selected from one of K classes. With respect to COVID-19 classification, $a \in \{0, 1\}$, where 0 corresponds to COVID-19 negative cases and 1 corresponds to COVID-19 positive cases.

Because the selection of an action, a, does not determine the following sample, s, presented to the agent, an alternative dependency must be introduced between s and a. To achieve this, a training episode is terminated when an agent incorrectly classifies the minority class, preventing any further reward, r. This allows for a relationship between s and a to be learned, especially when there are severe data imbalances between majority (COVID-19 negative) and minority (COVID-19 positive) classes. We have specifically chosen to use this off-policy Monte Carlo (i.e. model-free) RL approach, as an off-policy algorithm allows for the samples presented to the network to be independent and uncorrelated; and the model-free element means we don't learn a transition function (and thereby don't learn a trajectory), but instead, learn a mapping of state to appropriate action for all considered states. The overall RL framework is shown in Figure 5.2.

Defining Reward for Bias Mitigation

In the realm of reinforcement learning, customized reward functions have demonstrated effectiveness in mitigating significant data imbalances concerning the predicted label, as evidenced in a previous study [227], and further confirmed in our own investigations with COVID-19 diagnosis [234]. Building upon this foundation, we started our evaluations by implementing this previous method as a benchmark against our proposed method, which specifically aims to address data imbalances related to a sensitive feature. In Chapter 5.2.3, we provide further details on this comparative method. Here, rather than focusing on addressing label imbalances, our

methodology focuses on mitigating imbalances associated with the sensitive feature, specifically the hospital attended by a patient. By targeting imbalances related to sensitive features through the reward function, our proposed method aims to tackle potential biases that may arise from disparities across different hospital sites.

Standard classification models which use gradient descent, estimate the marginal distribution with a differentiable error term. This can skew models towards the majority class present in a batch, due to aggregation of the errors. However, reinforcement learning provides a way of indicating error using a non-differentiable signal that can be uniquely designed for each situation at hand. For example, for the task presented here, we can detect minority classes by representing this in the reward function, which aggregation typically doesn't allow you to do. As a result, a reinforcement learning paradigm allows for the learning of minority classes without needing to compromise on learning of majority classes, implicitly. This is particularly important in the tasks presented here, where we aim to train models that can generalize well across different patient demographics, patient outcomes, and hospital centres, even if their distributions are unequal at the time of model development.

The reward, R, is the signal evaluating the success of the agent's selected action. We introduce a specialized function for reward, uniquely formulated for the purpose of mitigating biases of the chosen sensitive feature, z. To do this, we separate the reward function into two components – one to help train a strong classifier, and one to debias with respect to the sensitive attribute. Furthermore, given that most prior investigations have primarily focused on assessing bias mitigation solely for binary attributes, we designed the reward function to address debiasing in multi-class attributes. This is particularly crucial in clinical contexts, where a greater level of detail is often necessary, as condensing values into fewer (e.g., binary) classes may lack biological relevance—especially in cases where classes are categorical—and may exhibit significant bias based on the sample population. Hence, to address class imbalance in multi-class sensitive attributes, we adjust the reward to be inversely proportional to the relative frequency of a class within the dataset. This approach resembles the use of cost-sensitive weights in standard supervised learning. However,

it is important to note, as discussed in Chapter 4, that while adjusting weights based on costs can mitigate class imbalances to some extent, it still relies on cross-entropy loss, which provides the network with a learning signal regardless of the presented data. Consequently, models may exhibit bias towards the majority class within a batch due to the aggregation of errors. By contrast, employing a reinforcement learning setup (as opposed to a supervised learning framework reliant on gradient descent) allows for more precise control over when and how a learning signal is backpropagated (further explained in the following sections). This was demonstrated through our previous investigations on imbalanced learning with respect to the outcome class label, where RL (with a specialized reward function) was compared to other common imbalanced learning methods (SMOTE, cost-sensitive/-adjusted weights), and found to improve on balanced classification [227, 234]. Results for this study can be found in Appendix C.6.

In our implementation, a positive reward is given when the agent correctly classifies the sample (as either COVID-19 positive or negative), and a negative reward is given otherwise. If a negative reward is given (i.e. a sample was misclassified), the absolute reward value given is inversely proportional to the relative presence of the label of the sample in the training data. Thus, the absolute reward value of a sample from the minority class is higher than that in the majority class, making the model more sensitive to the minority class. This helps accommodate for label imbalance during training; and since the primary purpose of the model is to effectively classify COVID-19 status, this sensitivity-differential will help the agent learn the optimal behavior for COVID-19 prediction. To consider the sensitive class, z (which we aim to debias), we make the absolute reward for the positive case (i.e. when a sample is correctly classified) inversely proportional to the relative presence of each respective z label in the training data, accommodating for any class-imbalances present in a multi-class sensitive attribute. Here, the absolute reward value of a sample from a minority z class is therefore higher than that from the majority class, making the model more sensitive to minority z labels. By performing debiasing on the positively rewarded states, we already know that the sample was correctly

classified for the main task, as debiasing would be inconsequential if the model was not clinically effective for use to begin with. The formulation introduced allows for evaluation of both binary and multi-class tasks and sensitive features.

To formulate the problem, we first use $m \in \{0, ..., M\}$ to represent the possible values for the sensitive feature and $k \in \{0, ..., K\}$ to denote the possible class labels. Then, for a patient with features s_t and sensitive feature z_t, if the model predicts COVID-19 status a_t, and the true COVID-19 status is l_t, the reward function is calculated as follows:

$$R(s_t, a_t, l_t, z_t) = \begin{cases} -\lambda_k & \text{if } a_t \neq l_t \text{ and } l_t = k \\ \lambda_m & \text{if } a_t = l_t \text{ and } z_t = m \end{cases} \tag{5.6}$$

$$\lambda_{\mathrm{k}} = \frac{\frac{1}{N_k}}{\left\| (\frac{1}{N_0}, \frac{1}{N_1}, ..., \frac{1}{N_K}) \right\|^2} \tag{5.7}$$

$$\lambda_{\mathrm{m}} = \frac{\frac{1}{n_m}}{\left\| (\frac{1}{n_0}, \frac{1}{n_1}, ..., \frac{1}{n_M}) \right\|^2} \tag{5.8}$$

where N_k represents the number of instances in class k, and n_m represents the number of instances with sensitive feature m. As seen in Equation 5.6, the reward is assigned differently depending on whether or not the model correctly predicts the COVID-19 label k.

In the first scenario, where the COVID-19 label is predicted incorrectly, the reward $-\lambda_k$ is inversely proportional to the frequency, N_k, of the COVID-19 label k. As discussed above, the motivation for this is that by penalizing incorrect COVID-19 predictions, the model will learn to predict COVID-19 status correctly. The weighting by label frequency is to address imbalance in the dataset.

In the second scenario, where the COVID-19 label k is predicted correctly, the reward λ_m is inversely proportional to the frequency, n_m, of the sensitive label z. The reason for this is that if the model is already able to predict COVID-19 status correctly, we now want to mitigate any unintentional biases.

After experimenting with different configurations for λ_k and λ_m, we found that models achieved desirable performance when these were set to be the vector-

normalized reciprocal of the class frequencies, as shown in Equations 5.7 and 5.8. To balance immediate and future rewards, a discount factor, $\gamma = 0.1$, is used.

Dueling Q-Network Architecture

In a typical deep Q-network (DQN) setup, the output layer of the network corresponds to predicted Q-values for state-action pairs. Since only one state-action pair can be trained at a time, it can be difficult to provide update information about the state. To address this, we choose to implement a dueling Q-network, which is capable of training state representations and action representations independent of one another.

For a DQN, the Q-network is implemented as a standard, single-stream neural network, where fully-connected layers are connected in a continuous sequence. The dueling Q-network, shown in Figure 5.3, instead, implements a fully-connected neural network with two streams - one for estimating the value (which is scalar) and another to estimate the advantages of each action (which is a vector). These two streams are combined to produce a single output, which is the Q function [233]. We compared this method to a non-dueling network by analyzing the training curves, where the dueling DDQN appears to outperform the non-dueling DDQN. This analysis can be found in Figure 5.1.

Based on the definition of the advantage function, we represent Q as:

$$Q(s_t, a_t; \theta, \alpha, \beta) = V(s_t; \theta, \beta) + \left(A(s_t, a_t; \theta, \alpha) - \text{softmax}(A(s_t, a_{t+1}; \theta, \alpha)) \right), \quad (5.9)$$

where α and β represent the parameters of the A and V streams of the fully connected layers, respectively. The additional softmax module is used to address identifiability issues and improve performance [233]. Additionally, this extra term does not change the relative rank of A (and subsequently, Q-values), which preserves the ϵ-greedy policy (which we use in our training; explained in Section 5.2.2).

For the Q-network, we used a fully-connected neural network with one hidden layer, alongside the rectified linear unit (ReLU) activation function and dropout. For updating model weights, the Adam optimizer was used during training. We

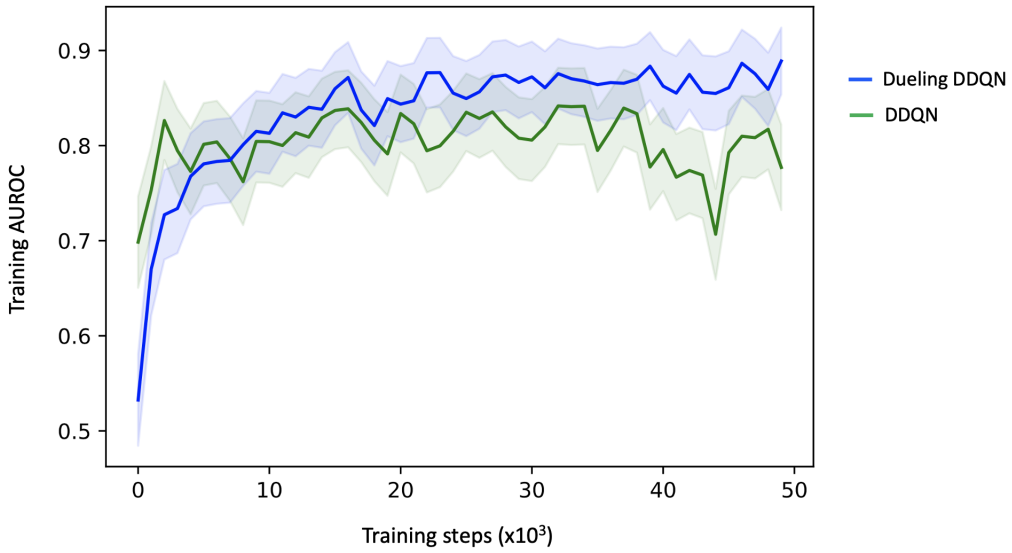

Figure 5.1: AUROC scores during training, comparing DDQN and Dueling DDQN models. Curves are shown for the COVID-19 prediction task.

set the exploration probability, ϵ, to be linearly attenuated from 1 to 0.01 over the entire training process. Each training period consists of 120,000 steps (i.e. iterations of updating parameters θ).

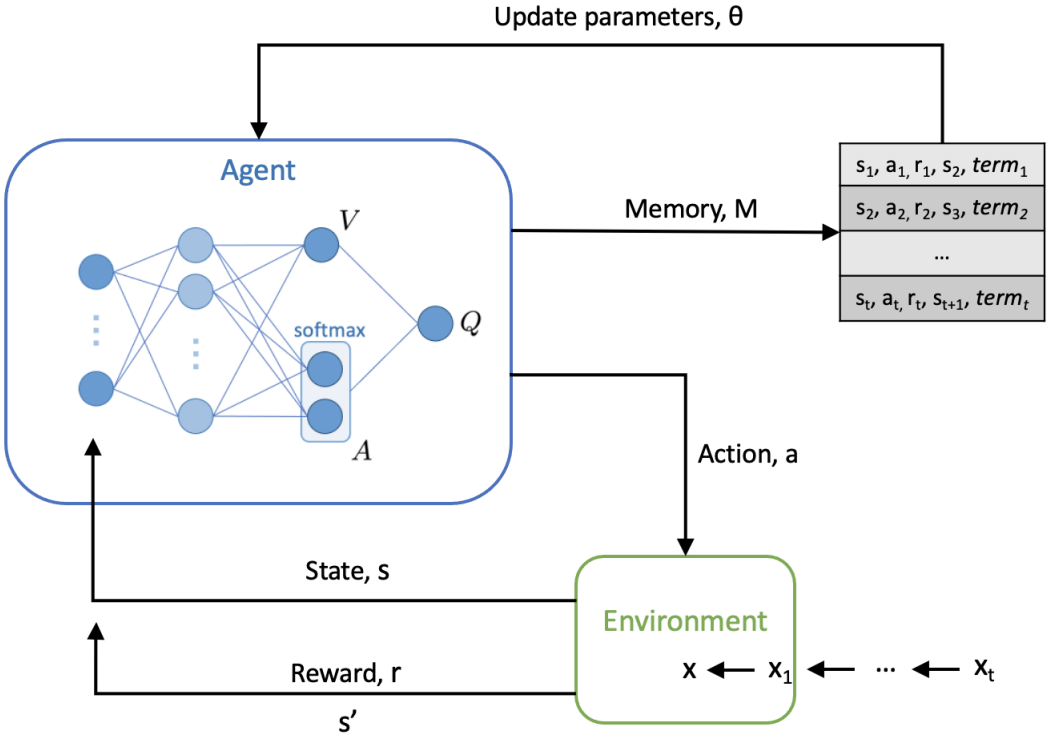

Figure 5.2: Overview of the reinforcement learning framework used.

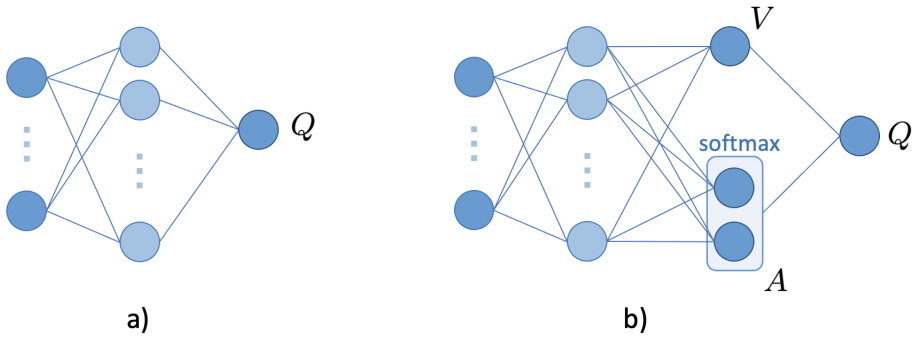

Figure 5.3: A typical single-stream Q-network is shown in a). A dueling architecture, with two streams to independently estimate the state-values (scalar) and advantages (vector) for each action is shown in b) (this implements Equation 5.9).

Double Deep Q-Learning

During each episode, combinations of states, actions, and rewards at each step, (s_t, a_t, r_t, s_{t+1}), are saved in the agent's working memory, M. To learn the parameters of the Q-network, θ, a randomly sampled mini batch of these transitions, B, are used in the gradient descent step. Recall the mean-squared error loss function, where y can be treated as the target to be predicted and $Q(s, a; \theta)$ as the prediction:

$$L(\theta) = \sum_{(s_t, a_t, r_t, s_{t+1}) \in B} (y - Q(s_t, a_t; \theta))^2 \tag{5.10}$$

We choose to define y using the format of a double deep Q-Network (DDQN). As a standard DQN uses the current Q-network to determine an action, as well as estimate its value, it has been shown to give overoptimistic value estimates [235]. This increases the likelihood of selecting overestimated values (which can occur even when action values are incorrect), making it harder to learn the optimal policy. Double deep Q-Learning (DDQN) was introduced as a method of reducing this overestimation [236]. Unlike a DQN, a DDQN uses the current Q-network to select actions, and the target Q-network to estimate its value [237]. Through decoupling the selection and evaluation steps, a separate set of weights, θ', can be used to provide an unbiased estimate of value.

The DDQN algorithm is implemented using the following target function:

$$y = r_t + (1 - term)\gamma Q(s_{t+1}, \underset{a_{t+1}}{\text{argmax}} Q(s_{t+1}, a_{t+1}; \theta); \theta') \tag{5.11}$$

As previously mentioned, a dependency between a state and action needs to be established for the agent to learn a relationship. As well as the regular feedback from rewards, the value of *term* is set to 1 once the agent reaches its terminal state, and 0 otherwise. A terminal state is reached after the agent has iterated through all samples in the training data (or a set number of samples, specified at the beginning of training), or when the agent misclassifies a sample from the minority class (preventing any further reward).

5.2.2 Reinforcement Learning Training Procedure

By framing our learning problem with a RL set-up, learning can be regulated through the design of the reward function. This enables control of how and when a learning signal is backpropagated, which aggregation (through standard supervised learning) typically doesn't allow for.

The environment's reward procedure is outlined in Algorithm 1. Here, D_1 represents the minority class, which in our case corresponds to the COVID-19 positive label. As previously discussed, when the COVID-19 label is predicted correctly ($a_t = l_t$), a reward λ_m is assigned, which is inversely proportional to the frequency of the sensitive label. Conversely, when the COVID-19 label is predicted incorrectly ($a_t \neq l_t$), a penalty of $-\lambda_k$ is applied, with λ_k being inversely proportional to the frequency of the COVID-19 label. The value of the *term* variable is initially set to *False* and is updated to *True* once the agent reaches its terminal state. As outlined earlier, a terminal state occurs when the agent has either processed all samples in the training data (or a predefined number of samples specified at the start of training) or misclassifies a sample from the minority class, thereby halting further reward accumulation.

The Q-network is trained following the DDQN procedure outlined in Algorithm 2. Training begins by shuffling the dataset D and initializing the initial state s_1 as x_1. During each training episode, the agent uses an ϵ-greedy behavior policy to select actions. With probability ϵ, the agent chooses a random action,

while with probability $1 - \epsilon$, it selects the action that maximizes the optimal Q-function, $\arg\max_a Q^*(s_t, a_t)$.

At each step of the episode, the transitions—combinations of states, actions, and rewards (s_t, a_t, r_t, s_{t+1})—are stored in the agent's working memory, M. A mini-batch of these transitions is then randomly sampled from M. For each sampled transition, y_j is set to r_j if the episode has terminated. Otherwise, y_j is computed using the target function defined in Equation 5.11. The network parameters θ are updated through gradient descent to minimize the loss. The resulting optimized Q-network serves as the trained classifier.

Algorithm 1: Environment Reward Procedure

State D_1 is the minority class set in the training data.
Function `Reward`(a_t, l_t)**:**
    Initialize $term_t = False$
    if $a_t = l_t$ **then**
        Set $r_t = \lambda_m$
    else
        Set $r_t = -\lambda_k$
        if $l_t \in D_1$ **then**
            Set $term = True$
    return r_t, $term$

5.2.3 Baseline Evaluation and Model Comparators

In Chapter 4, we've established robust baseline and state-of-the-art benchmarks to assess the proposed RL-based bias mitigation method.

Given that we introduce a deep Q-learning approach, a neural network serves as the foundation of our function approximator, facilitating the generalization of the relationship between state (represented by patient samples) and action (the COVID-19 prediction) pairs. This setup enables a more direct comparison with the consistency regularization and adversarial debiasing methods, which also rely on neural network frameworks for implementation.

Algorithm 2: DDQN Training Procedure

```
Initialize memory, M
Function Main(a_t, l_t):
    Initialize term_t = False
    for episode ∈ {1, 2, ..., E} do
        Shuffle training data, D
        Initialize state s_1 = x_1
        for t ∈ {1, 2, ..., T} do
            Choose action using ϵ-greedy policy:
            a_t = π_θ(s_t)
            r_t, term_t = Reward(a_t, l_t)
            Set s_{t+1} = x_{t+1}
            Store (s_t, a_t, r_t, s_{t+1}, term_t) to M
            Randomly sample (s_j, a_j, r_j, s_{j+1}, term_j) from M
            if term_j = True then
                Set y_j = r_j
            else
                Set y_j = r_j + γQ(s_{j+1}, argmax_{a_{j+1}} Q(s_{j+1}, a_{j+1}; θ); θ')
            Perform gradient descent on L(θ) w.r.t. θ
            Set θ' = θ
            if term_t = True then
                Break
```

Moreover, by incorporating cost-adjusted weighting within our proposed RL debiasing framework, we can directly compare it with the implementations of neural network and XGBoost models using the same weighting strategy.

Lastly, as we delve into a new machine learning paradigm, we also assess our method against an RL classification model lacking any debiasing component (as mentioned in Chapter 5.2.1). This comparative analysis ensures that our model initially trains as a robust classifier and validates the utility of incorporating a debiasing component. Specifically, we implement the same RL framework outlined, however the reward function used is:

$$R(s_t, a_t, l_t) = \begin{cases} \lambda_k & \text{if } a_t = l_t \text{ and } l_t = k \\ -\lambda_k & \text{if } a_t \neq l_t \text{ and } l_t = k \end{cases} \tag{5.12}$$

$$\lambda_{\mathrm{k}} = \frac{\frac{1}{N_K}}{\left\| \frac{1}{N_0}, \frac{1}{N_1}, ..., \frac{1}{N_k} \right\|^2} \tag{5.13}$$

Here, we let the reward for correctly/incorrectly labeling an instance of a particular class be inversely proportional to the relative presence of the class in the data. The absolute reward value of a sample from the minority class is thus higher than that in the majority class, making the model more sensitive to the minority class. Here, we do not provide any feedback regarding the label of the sensitive feature.

5.2.4 Evaluation Metrics

We employ identical evaluation metrics as specified in Section 4.2.5. Thus, to evaluate the performance of COVID-19 prediction, we report sensitivity, specificity, PPV, NPV, and AUROC, accompanied by 95% confidence intervals (CIs), computed using 1000 bootstrapped samples drawn from the test set. Tests of significance, indicated by p-values, involve evaluating how often one model outperforms another across 1000 pairs of bootstrapped iterations. We use 0.05 as the threshold value for determining statistical significance. Results are based on the evaluation of final, held-out test set. With respect to fairness, we use the standard deviation of Equalized Odds, as defined in Equations 4.16 and 4.17.

5.2.5 Hyperparameter Optimization

Consistent with Chapter 4.2.6, we use grid search alongside standard five-fold cross-validation on the training set to identify the optimal hyperparameters for the reinforcement learning methods deployed. This process involves optimizing parameters such as the number of layers in the Q-network, the number of nodes within each layer, and the learning rate.

Detailed information regarding the software, implementation, and finals hyperparameter values selected for each model can be located in Appendix C.

5.2.6 Threshold Optimization

In alignment with the findings detailed in Chapter 4.2.7, we conduct threshold optimization using the continuous validation set. This allowed for binary classification (COVID-19 positive or negative) to align with the "green–amber–blue" categorization

Table 5.1: Equalized Odds evaluation for hospital bias and COVID-19 status prediction test results across different models, optimized to sensitivities of 0.9. Bolded values denoting the the best (underlined) and second best Equalized Odds scores. Classification metrics are reported alongside 95% CIs, with bolded values denoting best scores achieved on the test set. The RL algorithm is a dueling DDQN.

Model	$M_{EO(TP)}$	$M_{EO(FP)}$	Sensitivity	Specificity	PPV	NPV	AUROC
RL (debiasing)	**0.010**	**0.040**	0.887 (± 0.019)	0.622 (± 0.008)	0.155 (± 0.009)	0.986 (± 0.003)	0.879 (± 0.014)
RL	0.035	0.063	**0.892 (± 0.019)**	0.553 (± 0.008)	0.135 (± 0.008)	0.985 (± 0.003)	0.855 (± 0.015)
XGBoost	0.024	0.057	0.875 (± 0.02)	**0.720 (± 0.008)**	**0.196 (± 0.011)**	0.987 (± 0.003)	0.900 (± 0.012)
Neural Net.	0.022	0.065	0.879 (± 0.019)	0.676 (± 0.008)	0.175 (± 0.01)	0.986 (± 0.002)	0.891 (± 0.013)
Neural Net. (weighted)	0.033	0.055	0.883 (± 0.019)	0.686 (± 0.008)	0.180 (± 0.011)	0.987 (± 0.002)	0.894 (± 0.013)
XGBoost (weighted)	**0.014**	0.057	**0.892 (± 0.019)**	0.681 (± 0.008)	0.179 (± 0.01)	**0.988 (± 0.002)**	**0.901 (± 0.012)**
Neural Net. (reg.)	0.022	0.046	0.865 (± 0.021)	0.714 (± 0.008)	0.191 (± 0.012)	0.985 (± 0.003)	0.892 (± 0.013)
Adversarial	0.022	**0.045**	0.882 (± 0.019)	0.642 (± 0.008)	0.161 (± 0.009)	0.986 (± 0.003)	0.882 (± 0.014)

system employed by NHS Trust policy. Here, the objective is to identify a suitable threshold, optimized to attain sensitivities of 0.9, ensuring clinically acceptable performance levels in detecting positive COVID-19 cases.

5.3 Results

For consistency with Chapter 4, following hyperparameter optimization and model fitting, we assessed all models using a held-out set. The outcomes showcased for XGBoost, the conventional neural network, weighted models, adversarial debiasing, and NCR debiasing are extracted from Table 4.1. Across all models, there was relative consistency in terms of the AUROC metric, with values ranging from 0.855 to 0.901. Notably, the weighted XGBoost model achieved the highest AUROC score of 0.901 (95% CI ±0.12), while the standard RL and RL debiasing models achieved the lowest scores at 0.855 (±0.15) and 0.879 (±0.14), respectively.

Furthermore, with a sensitivity configuration of 0.9, sensitivity scores remained consistent across all models, ranging from 0.865 to 0.892. Specifically, the neural network regularized using NCR achieved the lowest score of 0.865 (±0.021), while both the weighted XGBoost and standard RL methods attained the highest score of 0.882 (±0.019). The RL debiasing method followed closely with the second highest sensitivity of 0.887 (±0.019). It is important to note the trade-off between sensitivity and specificity metrics, as models with high sensitivity tended to exhibit slightly lower specificity.

All models demonstrated high prevalence-dependent NPV scores exceeding 0.985, suggesting their ability to confidently exclude COVID-19 cases.

Despite the apparent similarity in overall predictive performance among models, the difference in AUROC values was found to be statistically significant ($p < 0.0001$, based on 1,000 bootstrapped samples).

In the context of bias mitigation, the debiasing RL model (dueling DDQN) demonstrated the fairest performance, achieving the most favorable outcomes for Equalized Odds. It outperformed other models in both true positive and false positive metrics, with scores of 0.010 and 0.040, respectively. Across the hospital groups, it achieved true positive rates of 0.878, 0.886, 0.906, and 0.900 for OUH, UHB, BH, and PUH, respectively, and false positive rates of 0.419, 0.350, 0.308, and 0.364.

The weighted XGBoost and adversarial debiasing models followed closely in performance. The XGBoost model achieved the second-best Equalized Odds for true positive rates (0.014), with a range of 0.850–0.888 across hospital subgroups—slightly lower than the debiasing RL model. Meanwhile, the adversarial debiasing model achieved a false positive Equalized Odds score of 0.045, with false positive rates ranging from 0.273 to 0.396. Although the adversarial model demonstrated better false positive rates overall compared to the RL model, its wider range resulted in a slightly lower ranking.

Overall, the adversarial model ranked second in terms of Equalized Odds performance, followed by the weighted XGBoost model and the neural network regularized using NCR.

One interesting observation is that the weighted neural network exhibited consistently lower true positive and false positive rates across all hospital subgroups. Its true positive rates ranged from 0.750 to 0.875, compared to 0.849 to 0.969 observed in other models, while its false positive rates ranged from 0.121 to 0.200, significantly lower than the 0.207 to 0.502 range seen in other models.

A detailed analysis of true positive and false positive rates for each hospital subgroup is provided in Tables C.2 and C.3 in the Appendix.

When comparing the two RL methods, the model implementing a bias-mitigating reward function notably enhanced Equalized Odds, showing improvements of 0.025 and 0.023 for TP and FP Equalized Odds, respectively. Furthermore, comparing models with a debiasing component (particularly the regularized neural network, RL debiasing, and adversarial methods) to all other methods revealed a significant enhancement in fairness, with only a slight trade-off in predictive performance compared to the models achieving the highest AUROC scores.

We also found that the adjusted thresholds, listed in Appendix C.5, for the RL-based methods are also much closer to the default threshold (0.5) than the other methods (0.46-0.49, compared to 0.01-0.31 for non-RL methods), demonstrating how threshold adjustment may not be necessary even when the training data is heavily imbalanced (unlike the baseline models).

Moreover, due to its inherent computational complexity, reinforcement learning often demands more computational resources in comparison to supervised learning. To evaluate the computational efficiency of various models, including XGBoost, Neural Network, and Reinforcement Learning Models, we conducted an analysis of the average training times across 10 training iterations. Our findings indicated that RL consistently exhibited longer execution times compared to other supervised methods. Detailed results of this evaluation are presented in Appendix C.

5.4 Discussion

Comparable to the findings from the supervised learning-based models discussed in Chapter 4, we observed that the RL-based debiasing models yielded less biased outcomes compared to models lacking bias mitigation components. However, despite the reduction in bias, our models did not entirely meet the criteria for Equalized Odds. This discrepancy could be attributed to the imbalance in data concerning the sensitive attribute [102], with significantly more data available from OUH and PUH compared to UHB and BH. Additionally, given that we employed a neural network for Q-learning, skewed distributions within the data could potentially lead to inconsistent results.

The RL debiasing approach exhibited superior fairness performance compared to supervised-learning-based bias mitigation techniques, particularly in terms of both true positive and false positive rates. This superiority can be attributed to the ability of a RL setup to regulate the propagation of learning signals and refine error aggregation processes. In contrast to standard supervised learning frameworks, where cross-entropy loss provides a uniform learning signal irrespective of input data, RL frameworks offer more control over when and how learning signals are backpropagated to improve error aggregation. Consequently, the use of cost-sensitive weighting based on sensitive attribute frequency may have been more effective within the RL context. This could explain the relatively weaker fairness outcomes observed in the weighted neural network and XGBoost implementations. In these models, the indiscriminate learning signals generated by cross-entropy loss may have amplified biases. This may also explain the notable shifts in true positive and false positive rate ranges across hospital subgroups when using the weighted neural network model.

Although the prediction outcomes between RL methods and supervised learning methods exhibited similarity, a notable disparity lies in their training times, particularly the significantly longer duration required for training an RL model relative to supervised approaches, as observed in Appendix C. While this posed minimal problems with this dataset, where we dealt with relatively small feature sets, it could present considerable challenges when applying these methods to high-dimensional datasets like images. Thus, it is essential to consider the computational requirements associated with each modeling approach to facilitate the selection of the most suitable method based on computational constraints and performance prerequisites.

While the training time for reinforcement learning-based models may exceed that of other methods, such as supervised learning with XGBoost or neural networks, there are notable advantages to this approach. One significant advantage is the increased flexibility of applying reinforcement learning across various types of learning tasks. While supervised learning models are often tailored to specific tasks and require large amounts of labeled data, RL algorithms can be applied to a wide range of problems with minimal modifications. This inherent versatility

renders RL well-suited for tackling intricate, real-world challenges characterized by noisy, incomplete, or heterogeneous data, a common scenario encountered in clinical data settings.

Secondly, RL provides a framework for learning optimal decision-making policies in dynamic and uncertain environments. Unlike supervised learning, where the model's objective is to minimize prediction errors based on labeled data, RL agents aim to maximize cumulative rewards over time. This objective-driven methodology empowers RL models to dynamically adjust to evolving conditions and formulate decisions geared towards optimizing long-term outcomes, even without explicit training instances. Such adaptability proves particularly advantageous in clinical prediction tasks, where distributions, prevalence rates, and patient conditions may exhibit fluctuations over time. By embracing the inherent uncertainty of real-world scenarios, RL models have potential to navigate the complexities of clinical environments and provide decisions that align with evolving healthcare needs and circumstances.

Regarding threshold adjustment, we made modifications to the decision threshold to prioritize achieving high sensitivity in our models. However, as previously demonstrated, data (and subsequent downstream analyses like machine learning) can be influenced by site-specific factors, suggesting that optimal decision thresholds derived from one dataset/location may not necessarily generalize well to others. Nevertheless, we observed that the optimized thresholds for RL-based methods were already considerably closer to the default classification threshold (0.5) compared to non-RL methods, even when the training data exhibited significant imbalance with respect to the predicted label. Consequently, RL methods may offer an advantage in that the optimal threshold may require less adjustment to the training set, facilitating greater generalizability to new, unseen data. In conclusion, further investigation into the selection of an optimal decision threshold is warranted, given its direct impact on both model performance and fairness metrics, resulting from shifts in true positive and true negative rates.

Finally, RL enables agents to learn complex behaviors and strategies that may not be feasible with supervised learning alone. By exploring different actions and observing their consequences, RL agents can discover novel solutions and adapt their strategies over time based on feedback from the environment. Specifically, RL provides a mechanism for expressing errors through a non-differentiable signal that can be customized to suit the specific scenario. This means that if it is imperative to detect a minority or sensitive class, this requirement can be integrated into the reward function, a feat typically unattainable through simple aggregation methods. Consequently, embracing an RL framework facilitates the learning of the minority class without compromising the learning of the majority class, implicitly. This aspect is particularly critical in the context of the clinical tasks at hand, where the objective is to train models capable of effectively generalizing across diverse patient outcomes, even when the distributions of these outcomes are imbalanced during model development.

6

Mitigating Machine Learning Bias Across Low-Middle- & High-Income Countries

6.1 Introduction

6.1.1 Overview

Earlier sections have illustrated the presence of bias in machine learning models, stemming from the composition of training data. Such biases lead to varying performance across specific subgroups in predictive tasks and impede a model's capacity to accurately capture the relationship between features and the target outcome. Consequently, this results in suboptimal generalization and unfair decision-making [71, 72, 77]. Previous chapters have also demonstrated the efficacy of specific machine learning training frameworks in mitigating biases, particularly in successfully addressing site-specific biases across four distinct hospital groups in the UK, as applied to a COVID-19 screening task. Building on prior investigations, this chapter concentrates on evaluating the efficacy of machine learning debiasing techniques across hospitals situated in diverse socioeconomic contexts, particularly across high-income countries (HICs) and low-middle-income countries (LMICs). In our study, these correspond to the UK and Vietnam, respectively.

Hospitals in LMICs often contend with resource constraints, including inadequate funding, outdated infrastructure, a shortage of technical expertise, and

a limited availability of comprehensive and digitized healthcare data [238–241], all of which are essential requirements for the development and validation of AI algorithms. This often results in a significant disparity in resource availability between HIC and LMIC settings. Notably, this discrepancy in data availability creates a bias during the development and training of collaborative models, affecting their relative effectiveness when deployed across participating sites [239, 240, 242, 243]. Consequently, addressing and mitigating unintentional biases in data-driven algorithms is imperative to prevent the perpetuation or exacerbation of existing disparities in healthcare and society.

6.1.2 AI Challenges in Low-Middle-Income Countries

LMICs face several significant challenges in the development and implementation of artificial intelligence technologies. One of the foremost challenges is the lack of access to high-quality data [239–241]. LMICs often have limited resources and infrastructure for data collection, storage, and management. This scarcity of data can hinder the training and validation of AI algorithms, leading to models that may not generalize well or perform accurately in real-world settings. Additionally, existing data may be incomplete, biased, or of poor quality, further complicating AI development efforts in LMICs. For instance, a comprehensive systematic review focusing on LMICs assessed various aspects of health data quality management [244]. Among the studies analyzed, the review revealed instances of poor data completeness (ranging between 19-50%) [245, 246], inconsistencies in data reporting [247–249], and inaccuracies [246, 247].

Another obstacle lies in the scarcity of AI-related skills and expertise within LMICs. These regions often face challenges in both attracting and retaining qualified professionals proficient in critical areas such as data collection, analysis, and AI research, development, and implementation [250–252]. The limited pool of skilled workers constrains the capacity of LMICs to innovate and effectively utilize AI technologies [241]. Moreover, many LMICs lack educational and training programs specifically tailored to AI, further exacerbating the skills gap and hindering the

development of local AI talent [252]. This shortage not only hampers the adoption and deployment of AI solutions but also undermines efforts to address pressing societal and healthcare challenges within LMICs.

Infrastructure and technological limitations also present formidable obstacles to AI development in LMICs. Many of these countries grapple with inadequate internet connectivity, unreliable power supply, and limited computational resources, all of which are crucial for conducting AI research and deploying AI systems [250, 251, 253]. The absence of robust infrastructure impedes LMICs' ability to access cloud computing resources, deploy AI applications at scale, or effectively harness emerging AI technologies. Furthermore, regulatory and policy frameworks governing AI may be deficient or outdated in many LMICs, resulting in ambiguity and barriers to the adoption and implementation of AI solutions [241, 253, 254].

Ethical and societal concerns surrounding AI also present challenges in LMICs. Issues such as privacy, fairness, transparency, and accountability are particularly salient in contexts where marginalized populations may be disproportionately impacted by AI technologies [253, 254]. Moreover, LMICs may lack the resources and expertise to develop and enforce appropriate ethical guidelines and regulations for AI development and deployment [252]. Cultural and societal norms may also influence the acceptance and adoption of AI technologies in LMICs, underscoring the importance of contextual sensitivity and engagement with local communities in AI development endeavors [241]. Addressing these ethical and societal concerns is paramount to fostering trust, equity, and responsible use of AI technologies in LMICs, ensuring that these innovations contribute positively to societal well-being and development.

In summary, the predominant challenges in AI development within LMICs revolve around several key areas, including data scarcity, skills shortages, infrastructure limitations, regulatory gaps, and ethical considerations. These challenges collectively hinder the effective adoption and utilization of AI technologies in LMIC contexts. By addressing these challenges, LMICs can unlock the potential of AI to address pressing societal issues, improve healthcare delivery, enhance educational outcomes,

stimulate economic growth, and advance progress towards sustainable development goals [253]. However, concerted action and sustained commitment are needed to realize the transformative potential of AI in LMIC contexts.

6.1.3 Collaborative AI Development

Collaborative AI development between HICs and LMICs presents a significant opportunity to overcome some of the key barriers facing LMICs in the realm of AI [253]. This collaborative approach has the potential to significantly enhance technology access, build critical capacities, and ensure ethical, responsible AI development tailored to the unique needs and challenges of LMICs.

One of the primary benefits of such collaborations is the facilitation of technology transfer and the strengthening of local capacities in LMICs. Through partnerships with HICs, LMICs can gain access to advanced technologies, resources, and expertise [255–257]. HICs can offer specialized training, mentorship programs, and technical support, thereby enhancing the skills and knowledge base of researchers, tech entrepreneurs, and policymakers in LMICs. This transfer of knowledge and technology empowers LMICs to spearhead AI initiatives, fostering innovation and self-reliance in the local AI landscape. By building a robust AI ecosystem, these collaborations can drive sustainable development, economic growth, and technological self-sufficiency in LMICs.

Moreover, collaborative AI initiatives provide a platform for addressing complex ethical, regulatory, and societal challenges associated with AI. By bringing together a diverse array of stakeholders, including ethicists, policymakers, civil society organizations, and local communities, these collaborations can ensure that AI technologies are developed with a keen awareness of ethical considerations, regulatory compliance, and societal impact [258, 259]. Such a multifaceted, inclusive approach to AI development helps to build systems that are not only technologically advanced but also ethically sound, culturally sensitive, and socially inclusive. Engaging a wide range of perspectives in the development process promotes trust, transparency, and

accountability, key factors in achieving broad acceptance and successful integration of AI technologies within LMIC settings.

In addition to fostering AI innovation at a broad level, these collaborations can also significantly enhance AI model development through direct, practical engagements. One of the key avenues for this direct enhancement is through data sharing and the utilization of high-quality datasets [260], a critical component in the development and refinement of AI models. HICs often possess extensive, diverse, and meticulously curated datasets [261], which are essential for the training and validation of sophisticated AI algorithms. These datasets, encompassing a wide range of variables and scenarios, can significantly improve the performance of AI models, making them more robust, accurate, and applicable across different settings. For LMICs, access to such datasets can be a game-changer, enabling them to leapfrog some of the initial barriers to AI development related to data scarcity and quality.

Collaborative projects can facilitate access to AI resources in various ways. Firstly, by merging datasets from HICs with local data from LMICs, the resulting combined datasets become larger, more diverse, and more representative of various global contexts. This not only improves the training process but also ensures that the developed AI models are more generalizable and effective across different geographical and cultural settings. Secondly, LMICs can utilize the model weights from AI models initially trained on HIC data, and apply transfer learning techniques to fine-tune these models to specific LMIC contexts [262]. This approach provides a practical solution to the often stringent data sharing and privacy regulations that might restrict direct access to raw data across borders [263, 264]. By sharing model weights and utilizing transfer learning, collaborations can respect these legal and ethical boundaries while still facilitating the mutual exchange of AI knowledge and resources. Overall, these methods allow LMICs to build upon the advanced AI groundwork laid by HICs, tailoring it to their unique local needs and constraints without the necessity for extensive local datasets.

To summarize, partnerships in AI development between HICs and LMICs opens doors for LMICs to tackle obstacles, seize opportunities, and explore AI's capabilities

for sustainable progress. Such collaborations go beyond merely enhancing AI technologies; they aim to narrow the digital gap and promote worldwide participation in creating and applying AI solutions. However, the benefits of these collaborative efforts must be balanced with a commitment to fairness and equity, particularly given the distinct challenges and conditions faced by LMICs compared to their HIC counterparts.

6.2 AI Generalizability Across Low-Middle- & High-Income Countries

Machine learning generalization refers to a model's ability to accurately apply its learned knowledge from training data to new, unseen data. This capability is particularly valuable when models are deployed in real-world scenarios, where they must perform well on independent datasets encountered in real-time. In clinical contexts, two common types of generalizability are temporal generalizability (applying prospectively within the center where a model was developed) and external/geographic generalizability (applying a model at an independent center). Based on the previously discussed challenges surrounding location bias, we will focus on external/geographic generalizability.

This chapter considers the collaboration between the Oxford University Clinical Research Unit (OUCRU) in Ho Chi Minh City, Vietnam, and the University of Oxford Institute of Biomedical Engineering in Oxford, England, along with the Hospital for Tropical Diseases in Ho Chi Minh, Vietnam, and the National Hospital for Tropical Diseases in Hanoi, Vietnam. This partnership is dedicated to enhancing critical care in low-middle-income country contexts. Their primary objective is to accurately identify patients requiring critical care and enhance the quality of care they receive, addressing the unique challenges encountered within LMIC healthcare systems.

Previously, we developed an AI-driven rapid COVID-19 triaging tool using data across four UK NHS Trusts. As such, through our collaboration with Vietnam-based centres, we aimed to translate the UK-based models to LMIC settings, specifically

at the Hospital for Tropical Diseases (HTD) in Ho Chi Minh, Vietnam, and the National Hospital for Tropical Diseases (NHTD) in Hanoi, Vietnam. Applying a similar screening tool at the HTD and NHTD in Vietnam could offer a systematic approach to prioritize and manage patient care. It would allow for the efficient use of limited resources, including clinician expertise, ventilators, and beds, ultimately optimizing patient outcomes and ensuring timely access to appropriate interventions. These benefits are especially valuable in LMIC settings where resource constraints pose significant challenges to healthcare delivery.

With this COVID-19 case study, we start by assessing the viability of adapting models across hospitals serving diverse socioeconomic demographics, illustrating the significant impact of inherent biases on the generalizability of the models.

6.2.1 Methods

Datasets

We use clinical data comprising of linked and deidentified demographic information from patients across four hospital centers in the UK and two hospitals in Vietnam. As discussed in Chapter 3, from the UK, the datasets included electronic health records from hospital emergency departments in OUH, PUH, UHB, and BH. The four UK datasets used are identical to those introduced in Chapter 3. However, to emphasize the difference in digital data availability (particularly with respect to historical data), we additionally include historical ("pre-pandemic") data from OUH. As before, we received UK NHS approval via the national oversight/regulatory body, the HRA, for the development and validation of artificial intelligence models aimed at detecting COVID-19 (CURIAL; NHS HRA IRAS ID: 281832).

The data from Vietnam was sourced from the intensive care units (ICUs) in the Hospital for Tropical Diseases (HTD) and the National Hospital for Tropical Diseases (NHTD). The ethics committees of the HTD and the NHTD approved use of the HTD and NHTD datasets for COVID-19 diagnosis, respectively. For the Vietnam hospitals, we extracted data from the Critical Care Asia Registry

(we will refer to this as "Registry"), a dedicated prospectively acquired database facilitating quality improvement initiatives.

In previous chapters, to counteract bias stemming from hospital locations, it was imperative to include data from all sites in both the training and validation datasets. Consequently, we amalgamated all datasets to ensure representation from each site. Here, to address generalizability across sites, we perform prospective and external validation across multiple datasets.

From OUH, we have three data extracts corresponding to distinct periods: pre-pandemic presentations (before December 1, 2019), the first wave of the COVID-19 epidemic in the UK (December 1, 2019, to June 30, 2020), and the second wave (October 1, 2020, to March 6, 2021). The positive COVID-19 presentations from "wave one" and pre-pandemic controls were used for training and continuous validation, with an 80% to 20% random split, respectively.

As highlighted in Chapter 3, challenges such as incomplete testing and the imperfect sensitivity of the PCR swab test led to uncertainties in determining the viral status of patients who were either untested or tested negative [144]. To address this, following the methodology used in [144], each positive COVID-19 presentation from "wave one" was matched with a set of pre-pandemic negative controls based on age. Using patient presentations from OUH prior to the global COVID-19 outbreak guarantees that these cases are COVID-free. This careful selection of data ensures the accuracy of COVID-19 status labels used during the training phase of the model. For our purposes, we employed a ratio of 20 controls to 1 positive presentation for the training set. This matching approach aimed to simulate a disease prevalence of 5%, consistent with the actual COVID-19 prevalences observed at all four UK sites during the data extraction period (ranging from 4.27% to 12.2%). It is important to note that this matching process was exclusively applied to the training set, as it directly influences the model weights and biases. The continuous validation set, used solely to evaluate model training and determine an evaluation threshold (without altering the model itself), retains the full stratification without simulating a 5% prevalence. To account for the uncertainty in negative

PCR results, a sensitivity analysis was conducted, yielding improvements in the apparent accuracy of the models, as outlined in [144].

Thus, for the training dataset, we used 114,957 patient presentations from OUH prior to the global COVID-19 outbreak, guaranteeing that these cases are COVID-free. Additionally, we included 701 patient presentations that tested positive for COVID-19. This careful selection of data ensured the accuracy of COVID-19 status labels used during the training phase of the model.

We then validated the model on four UK cohorts (OUH "wave two", UHB, PUH, BH), totalling 72,223 admitted patients (4,600 COVID-19 positive), and two Vietnam cohorts (HTD and NHTD), totalling 3,431 admitted patients (2,413 COVID-19 positive). A summary of each respective cohort is in Table 6.1.

HTD considered all patients admitted between December 10, 2020 and December 30,2022. NHTD considered all patients admitted between November 1, 2020 and December 21,2022.

COVID-19 status at the UK sites and HTD was determined through confirmatory PCR testing, while at NHTD, both PCR and/or rapid antigen testing were used. Nonetheless, concerning NHTD, there were numerous instances where the specific test type was not recorded. Therefore, in order to maximize testing coverage, in cases where the test type was unspecified (Table 6.2), we examined how COVID-19 was documented during patient evaluation, including terms such as "COVID-19 lower respiratory infection," "COVID-19 pneumonia," "SARS-COV-2 Infection," "COVID-19 acute respiratory distress syndrome," "Acute COVID-19," and others, to determine the presence of COVID-19. A full count of all COVID-19 severity levels in the HTD and NHTD cohorts can be found in Figures D.1 and D.2 in Appendix D. For our analysis, alongside confirmatory testing, we considered any indication and severity of COVID-19 presence as COVID-19 positive. These diagnoses were confirmed by attending specialist infectious diseases clinicians, and thus, we consider these diagnostic labels to be robust. A breakdown of the labels available within the NHTD database is provided in Table 6.2. Furthermore, we conducted a sensitivity analysis for NHTD, comparing PCR-confirmed outcomes

Table 6.1: Total patients and positive COVID-19 cases in the OUH training cohorts (OUH pre-pandemic and "wave one"), prospective validation cohort (OUH), external validation cohorts of patients admitted to three independent NHS Trusts (UHB, PUH, BH), and external validation cohorts of patients admitted to two Vietnam-based hospitals (NTD, NHTD).

	Cohort	Total Patients	COVID-19 Positive Cases
OUH pre-pandemic	Before Dec 1/19	114,957	0
OUH "wave 1"	Dec 1/19-June 30/20	701	701
OUH "wave 2"	Oct 1/20-Mar 6/21	22,857	2,012 (8.80%)
UHB	Dec 1/19-Oct 29/20	10,293	439 (4.27%)
PUH	Mar 1/20-Feb 28/21	37,896	2,005 (5.29%)
BH	Jan 1/21-Mar 31/21	1,177	144 (12.2%)
HTD	Dec 10/20-Dec 30/22	1,820	1,360 (74.7%)
NHTD	Nov 1/20-Dec 21/22	1,611	1053 (65.4%)

with those incorporating rapid antigen tests and other written documentation of COVID-19, which is detailed in Section 6.2.2.

Table 6.2: Diagnostic testing method for COVID-19 detection.

COVID-19 Test-type	NHTD
PCR	880
Rapid Antigen Test	69
Other	58
Unspecified	604

Clinical Features

Consistent with previous chapters, the focus is on rapid patient triaging, with datasets encompassing a segment of regularly acquired clinical data, comprising initial blood tests (encompassing full blood counts, liver function tests, and electrolytes), vital signs, and the confirmation of COVID-19 diagnosis through a PCR swab test.

It is important to note that certain features, such as C-reactive protein, bilirubin, albumin, alkaline phosphatase, urea, and estimated glomerular filtration rate (found in the UK datasets), are not part of the standard admission protocol at HTD and NHTD. We were also unable to obtain many of the blood test features from NHTD. Therefore, in order to integrate data from the UK, HTD, and NHTD, we had to match the features available in the UK hospitals with those present in the NHTD database. Additionally, any features with missing values exceeding 30% were excluded. Table 6.3 summarizes the final features included.

Table 6.3: Clinical predictors considered for COVID-19 diagnosis.

Category	Matched UK and Vietnam	UK Features
Vital Signs	Heart rate, respiratory rate, systolic blood pressure, diastolic blood pressure, temperature	
Blood Test	Haemoglobin, haematocrit, white cell count, platelets	Mean cell volume, neutrophil count, lymphocyte count, monocyte count, eosinophil count, basophil count
Liver Function Tests & C-reactive protein	Bilirubin	Albumin, alkaline phosphatase, alanine aminotransferase, C-reactive protein
Urea & Electrolytes	Sodium, potassium, creatinine, urea	Estimated glomerular filtration rate

Data Pre-processing

Once we confirmed consistency in the units applied to identical features, we proceeded with implementing the pre-processing procedures outlined in Section 3.2.4. Consequently, all features were standardized to attain a mean of 0 and a standard deviation of 1. Additionally, we used the population median to impute any missing values.

Model Architectures

In order to evaluate the generalizability of developed models, we conducted investigations using XGBoost and a standard neural network, the baselines used in previous chapters. Also consistent with previous chapters, we trained a fully-connected neural network which used the rectified linear unit (ReLU) activation function in the hidden layers and the sigmoid activation function in the output layer. For updating model weights, the Adam optimizer was used during training.

Evaluation Metrics

We adopt the same evaluation metrics outlined in Section 4.2.5, reporting on sensitivity, specificity, PPV, NPV, and AUROC with 95% CIs derived from 1000 bootstrapped samples from the test set. Significance tests, marked by p-values, assess model performance differences over 1000 bootstrapped comparisons, using a significance threshold of 0.05. These analyses are performed on final test sets.

Hyperparameter Optimization

As in with Chapter 4.2.6, we employ a grid search together with standard five-fold cross-validation on the training set to determine the optimal hyperparameters for both XGBoost and the neural network.

Grid search was applied to the neural network model to determine the best hyperparameter configurations, encompassing factors like the number of hidden layers, nodes per layer, and learning rate. In the case of XGBoost, we explored a range of parameters such as learning rate, depth, and the number of trees.

Detailed information regarding the software, implementation, and final hyperparameter values selected for each model can be located in Appendix D.

Threshold Optimization

In alignment with the findings detailed in Chapter 4.2.7, we conduct threshold optimization, employing binary classification (COVID-19 positive or negative). Again, the objective is to identify a suitable threshold, optimized to attain sensitivities of 0.9, ensuring clinically acceptable performance levels in detecting positive COVID-19 cases.

For each task, we used a training set to develop, select hyperparameters, train, and optimize the models. A separate validation set was employed for ongoing validation and threshold adjustment. Following successful training, six independent test sets were used to evaluate the performance of the final models.

6.2.2 Results

To start, we used the OUH pre-pandemic controls and "wave one" positive cases to develop models, using the matched feature set (Table 6.3).

In the original study mentioned in Chapter 2.5.1, laboratory blood markers, such as eosinophils and basophils, were identified as having a significant impact on model predictions. This determination was made through the application of Shapley Additive Explanations (SHAP) analysis during the development and evaluation of models using patient cohorts from the UK (this can be found in Appendix A).

However, these particular features were not accessible in the Registry dataset, and consequently, were not incorporated into the initial models developed for compatible testing across UK and Vietnam cohorts. We hypothesize that without the inclusion of these features during training, the models' performance would be inferior compared to the previously reported scores.

COVID-19 prevalences observed at all four UK sites during the data extraction period ranged from 4.27% to 12.2%. COVID-19 prevalence was highest in the BH cohort, owing to the evaluation timeline spanning the second UK pandemic wave during January 1, 2021 to March 31, 2021 (12.2% vs 5.29% in PUH and 4.27% in UHB; Fisher's exact test $p < 0.0001$ for both). Prevalance at the Vietnam sites was significantly higher (74.7% and 65.4% at HTD and NHTD, respectively, $p < 0.0001$), as these were exclusively infectious disease hospitals, and handling the most severe cases of COVID-19.

Between all UK and Vietnam cohorts, all matched features had a significant difference in population median (Kruskal-Wallis, $p < 0.0001$). In the case of features exclusive to the UK cohorts, a significant distinction in population median across hospital sites was observed for all features, except for mean cell volume, where the population median appeared to be similar ($p = 0.210$). Full summary statistics (including median and interquartile ranges) of vital signs and blood tests for all patient cohorts are presented in Supplementary Tables D.4 and D.5, respectively.

It is important to highlight that, upon a preliminary examination of the summary statistics of the datasets, we observed the presence of extreme values in the Vietnam datasets. For instance, the minimum haemoglobin value was recorded as 11 g/L, which is notably rare, as values this low are typically considered highly unlikely [265–267]. Another instance is observed in the white blood cell count feature, where the dataset's maximum value was registered at 300, an exceptionally high value [267]. While such levels of deviation theoretically can occur in cases of haematological malignancy [268], they remain exceedingly rare occurrences. In the Vietnam datasets, there were some extreme values in patients with lymphoma. For our experiments, we made a deliberate choice to retain these extreme values in the

Table 6.4: COVID-19 diagnosis performance across XGBoost and neural network models trained on the UK data. Results are optimized to sensitivities of 0.9. Bolded values denote the best scores across models for each test set. Metrics are reported alongside 95% confidence intervals (CIs).

Test Set	Model	Sensitivity	Specificity	PPV	NPV	AUROC
OUH	XGBoost	0.711(±0.018)	0.716(±0.005)	0.195(±0.005)	0.962(±0.002)	0.784(±0.010)
	Neural Net.	**0.718(±0.017)**	**0.725(±0.005)**	**0.201(±0.005)**	**0.964(±0.002)**	**0.803(±0.010)**
PUH	XGBoost	0.766(±0.016)	**0.709(±0.005)**	**0.128(±0.003)**	0.982(±0.001)	**0.817(±0.009)**
	Neural Net.	**0.835(±0.014)**	0.592(±0.005)	0.103(±0.002)	**0.985(±0.002)**	0.816(±0.009)
UHB	XGBoost	0.617(±0.039)	**0.754(±0.008)**	0.101(±0.006)	0.978(±0.002)	0.76(±0.020)
	Neural Net.	**0.69(±0.036)**	0.743(±0.007)	**0.107(±0.006)**	**0.982(±0.002)**	**0.776(±0.021)**
BH	XGBoost	0.576(±0.070)	**0.811(±0.020)**	**0.299(±0.034)**	0.932(±0.011)	0.773(±0.038)
	Neural Net.	**0.688(±0.069)**	0.74(±0.056)	0.269(±0.027)	**0.944(±0.012)**	**0.804(±0.034)**
HTD	XGBoost	0.803(±0.017)	**0.202(±0.034)**	0.748(±0.009)	0.258(±0.035)	0.533(±0.026)
	Neural Net.	**0.908(±0.013)**	0.139(±0.028)	**0.757(±0.007)**	**0.339(±0.056)**	**0.577(±0.027)**
NHTD	XGBoost	0.727(±0.024)	**0.272(±0.032)**	**0.654(±0.013)**	**0.346(±0.033)**	0.478(±0.026)
	Neural Net.	**0.831(±0.019)**	0.159(±0.026)	0.651(±0.009)	0.333(±0.045)	**0.515(±0.026)**

dataset. This decision was motivated by our aim to evaluate the performance of models using real-world data, acknowledging the presence of extreme values and potential errors. This is further discussed in Chapter 6.2.3.

Following the training of models on the OUH pre-pandemic and "wave one" data, we conducted prospective and/or external validation on six datasets. As anticipated, when utilizing the matched dataset based on the available features in Registry, the performance of the models was approximately 5%-10% lower in terms of AUROC compared to our previous investigations using the same training and test cohorts. The AUROC ranges were as follows: OUH (0.784-0.803), PUH (0.816-0.817), UHB (0.76-0.776), BH (0.773-0.804), in contrast to the results reported from the validation study highlighted in Chapter 2.5.1 [133]: OUH (0.843-0.878), PUH (0.842-0.872), UHB (0.836-0.858), BH (0.854-0.881). The AUROC scores remained relatively consistent across all UK test sets, with a standard deviation (SD) of 0.017 for the neural network model. However, the AUROC was lower for the HTD and NHTD centers, with an neural network AUROC of 0.577 (CI ±0.027) and 0.515 (±0.026), respectively.

Sensitivity scores varied across all test sets, with an SD of 0.090 for the neural network model. The highest sensitivities were observed at HTD, PUH, and NHTD (0.908, 0.835, 0.831 for the neural network model, respectively), while the lowest sensitivities were observed at OUH, UHB, and BH (0.718, 0.690, 0.688 for the neural

network model, respectively). Even within the same country, there was a significant range in sensitivity, with ranges of 0.688-0.835 for UK centers and 0.831-0.908 for Vietnam centers in the neural network model. In the UK test sets, specificity exhibited a reasonable balance with sensitivity. However, for the Vietnam datasets, specificity was notably poor, with values of 0.139 ($\pm$0.028) and 0.159 ($\pm$0.026) for neural network models at HTD and NHTD, respectively.

Consistent with previous studies, our models achieved high prevalence-dependent NPV scores (>0.944) on the UK datasets, demonstrating their ability to confidently exclude COVID-19 cases.

We conducted an additional sensitivity analysis to address the uncertainty surrounding the viral status of patients who underwent rapid antigen testing or where the testing method was unspecified at NHTD. Utilizing the neural network model, which demonstrated superior performance, and evaluating solely on the subset of NHTD patients with confirmed PCR testing, we attained AUROC scores of 0.492 ($\pm$0.056) for the NHTD set. Generally, we observed comparable results, indicated by overlapping confidence intervals, when compared to datasets incorporating alternative testing methods.

6.2.3 Discussion

Using ready-made HIC models (UK models) in LMIC settings (Vietnam hospitals) without customization resulted in the lower predictive performance and the high variability in AUROC and sensitivity/specificity. This finding aligns with prior research indicating that model performance deteriorates when models trained in one context are then used in a different context, such as the shift from HIC to LMIC settings [239, 243]. Thus, these outcomes were anticipated, as diverse hospital settings can significantly differ in terms of unobserved factors, protocols, and cohort distributions, posing challenges to model generalization. Despite potential similarities in human pathophysiology for specific outcomes, neural networks heavily rely on the specific datasets and patient cohorts used during training [44]. Therefore, considering the unique attributes of each setting is crucial for achieving optimal

model performance. In particular, the datasets analyzed in this study exhibited variations in patient demographics, genotypic/phenotypic characteristics, and other determinants of health, such as environmental, social, and cultural factors. For example, the HTD and NHTD datasets were primarily composed of Southeast Asian (Vietnamese) patients, as opposed to the UK datasets, which had a majority of patients from a white demographic, and this may have influenced the models generalization capabilities.

Regarding data quality, we also detected the presence of outliers within the Vietnam datasets, such as the minimum recorded haemoglobin value of 11 g/L. This particular value would typically be considered highly improbable [265, 266]. The existence of such outliers could be attributed to a unit conversion error, where values were erroneously shifted by a factor of 10 (some locations utilize g/dL instead of g/L), or they may be the result of data entry errors. Since we aimed to work with real data, our model incorporates such instances of incorrect data entry and outliers. In the case of extreme values for white blood cell count, there were some extreme values found in patients with lymphoma in the HTD and NHTD datasets. In certain scenarios, outliers like these may contain unique information that can enhance a model's ability to generalize effectively, rendering the models more robust and less susceptible to noise. The decision of whether to retain extreme values in a dataset or not depends on the context and the problem under consideration. Extreme values can indeed offer valuable information, but it is important to handle them appropriately to prevent any adverse impact on model performance [269, 270]. Therefore, for future studies, it may be worthwhile to explore additional filtering and preprocessing steps to address these anomalies and enhance the dataset's quality before model development and testing.

It is essential to consider that HTD and NHTD are specialized hospitals for infectious diseases. They specifically designated as "COVID-19" hospitals during the pandemic, primarily receiving referrals for severe cases of COVID-19. While both the UK and Vietnam datasets included the first recorded blood tests and observations, it is important to acknowledge that in LMICs, there might be some

delay in recording these features after the initial presentation. Moreover, COVID-19 negative cases in these facilities typically involved other infectious diseases, and critical cases, including patients with various comorbidities, were treated at these hospitals. Given that the Vietnamese cohorts primarily consisted of severely ill patients, this might account for the more noticeable fluctuations in blood test results. Due to these differences, models may encounter challenges in accurately differentiating COVID-19 for patients at HTD and NHTD based on vital signs and blood test features, as other diseases (including infectious diseases) might also be present. Furthermore, in the case of UK hospitals, there was a broader spectrum of COVID-19 case severity. The UK datasets encompassed all individuals coming to the hospital, with only a small subset of patients progressing to ICUs. Consequently, diagnosing COVID-19 using AI is a significantly more challenging task at HTD and NHTD because we must distinguish the specific reason for ICU admission, particularly in cases of infectious diseases. For instance, distinguishing COVID-19 from bacterial pneumonia (which is frequently encountered at HTD and NHTD) is more challenging than distinguishing it from injuries such as a fractured leg.

This difficulty may also account for the lower level of specificity observed in the HTD and NHTD datasets compared to the UK sites. Thus, even if AUROC metrics are high at external hospitals sites, it may be necessary to tailor the classification threshold (i.e., the criterion for categorizing COVID-19 status as positive or negative) for each site independently, to maintain the desired levels of sensitivity and specificity [44].

While we analyzed patient cohorts admitted to ICUs at HTD and NHTD, the datasets and features we used were those readily available and documented upon hospital admission. These models can provide swift insights and facilitate efficient and precise triage during a patient's initial presentation at the hospital. It is important to note that in many cases, such as those observed in Vietnam, by the time patients are transferred from the hospital to the ICU, the diagnosis is typically already established. Therefore, even though similar features are recorded upon ICU admission, in these scenarios, the relevance of a machine learning-based classification

algorithm may appear redundant, and the benefits of diagnosing at ICU admission may be limited. Ultimately, the decision to employ machine learning algorithms should consider various factors, including the clinical context, the patient's condition, and the urgency of the situation. Additionally, similar approaches could be applied to other diseases or integrated into local hospital protocols, including guidelines for patient transfer, among other considerations.

It is also important to acknowledge that prediction models can never be fully validated due to inherent variability in their performance across different locations, settings, and time periods [271, 272]. A single external validation study conducted in a specific geographical area, during a particular time frame, and within a distinct patient population offers only a limited view and cannot assert universal applicability. In this study, our investigation spanned a significant time period, from December 1, 2019, to December 30, 2022. During this extended duration and particularly during peak pandemic periods, such as the COVID-19 outbreak, the relationship between patient and disease factors with clinical events, including hospital-acquired infections, may undergo changes [272]. Additionally, over time, there may be variations in practice patterns such as hardware and software updates and changes in protocols, which can impact data capture and outcomes. Although this retrospective study offered valuable insights into historical data, future research should ideally focus on prospective analysis. Models should be updated regularly to maintain their relevance. This approach enables a more dynamic assessment of model performance and provides timely feedback for refining and improving predictive models. Therefore, future validation efforts should aim to quantify and comprehend the heterogeneity in model performance, rather than solely focusing on point estimates [271]. This broader understanding of performance variability is crucial for refining and improving the models over time. For instance, in LMIC settings, real-time data preprocessing and curation can be achieved through cost-effective and accessible strategies. In the study highlighted here, an offline, in-house version of the algorithm can be used, where a doctor manually enters feature values in real-time (feasible with only 14 features). These values can then be

automatically processed through a script that imputes missing features and performs standardization, ultimately outputting a diagnosis for further triaging. Additionally, emphasizing the use of open-source tools and scalable, cost-effective infrastructure ensures applicability in resource-constrained settings.

Finally, the adoption of AI in LMICs encounters significant infrastructural and capacity-building challenges [238, 240, 241, 273]. These challenges encompass power outages, unreliable internet connectivity, cybersecurity concerns, inadequate digital infrastructure (such as data and storage), and a shortage of skilled AI professionals. As a result, prioritizing AI solutions may divert resources from more urgent foundational needs. These issues also impact the broader concern of AI governance, which remains a challenge even in HICs [274], and is likely even more challenging in LMICs. Therefore, while AI holds promise, its adoption in LMICs necessitates a careful, context-sensitive approach to address these underlying challenges.

6.3 AI Bias Mitigation Across Low-Middle- and High-Income Hospitals

The preceding section underscored the challenge of generalizability when attempting to directly implement a HIC model in a LMIC context. Consequently, this section delves into a collaborative training initiative that uses datasets from both UK and Vietnamese hospitals during the training process. In this endeavor, we exclusively utilize data from HTD, as the NHTD dataset exhibits a higher prevalence of missing features. This strategic decision enables us to develop more robust models, effectively showcasing the impacts of bias mitigation techniques. Furthermore, the contrast between data sourced from four hospitals in the UK and only one in Vietnam underscores the common resource disparity between HICs and LMICs. This inequity sheds light on potential biases inherent in the development and training of models, which could compromise the efficacy of these models in diverse settings, particularly those marked by significant socioeconomic disparities [239, 240, 242, 243]. It is imperative to address and mitigate these biases in data-driven technologies to prevent further reinforcing or exacerbating existing healthcare and societal inequalities.

Building on the work of previous chapters that examined the creation of fair machine learning systems in relation to various UK hospital sites, this part of the thesis extends the examination to include an assessment of the effectiveness of machine learning debiasing techniques across hospitals operating under differing socioeconomic conditions.

In line with insights from Chapters 4 and 5, our focus will be on reinforcement learning-based debiasing and adversarial debiasing, as these were the best and second best performing models in terms of fairness. As before, we employ the statistical definition of Equalized Odds. Using the same COVID-19 case study, our aim is to mitigate any site-specific biases and assess the effectiveness—considering both classification performance and fairness—of bias mitigation models trained collaboratively across both HIC and LMIC hospital settings.

6.3.1 Methods

Datasets

We use the same data cohorts as those mentioned previously in Chapter 6.2.1, except we do not include the NHTD dataset.

Thus, we use the pre-pandemic and "wave one" cohort as the training and continous validation set, with an 80% to 20% random split, respectively. The "wave two" dataset, comprising both negative and positive COVID-19 cases confirmed through PCR testing, was designated as the held-out test set.

Since UHB, PUH, BH, and HTD each provided a single extract, we divided each into training, continuous validation, and test sets through a random split (allocated at 60%, 20%, and 20%, respectively). This division was stratified based on the COVID-19 status.

Finally, we merged data from all sites, yielding final training ($\mathcal{D}_{train}$), continuous validation ($\mathcal{D}_{val}$), and held-out test ($\mathcal{D}_{test}$) sets comprising 42,385, 33,371, and 33,095 presentations, respectively (including 2,912, 944, and 2,805 COVID-19 positive cases, respectively). A summary of each dataset is provided in Table 6.5.

Table 6.5: Aggregate of patients, positive COVID-19 instances, and distribution across hospitals in the training, continuous validation, and test sets.

	Training ($\mathcal{D}_{train}$)	Continuous Validation ($\mathcal{D}_{val}$)	Test ($\mathcal{D}_{test}$)
Total Patients	42,385	33,371	33,095
COVID-19 positive (%)	2,912 (6.9%)	944 (2.8%)	2,805 (8.5%)
OUH (%)	11,676 (27.5%)	23,132 (69.3%)	22,857 (69.1%)
UHB (%)	6,175 (14.6%)	2,059 (6.2%)	2,059 (6.2%)
PUH (%)	22,737 (53.6%)	7,580 (22.7%)	7,579 (22.9%)
BH (%)	705 (1.7%)	236 (0.7%)	236 (0.7%)
HTD (%)	1,092 (2.6%)	364 (1.1%)	364 (1.1%)

Clinical Features

As previously seen, in order to integrate data from both the UK and HTD, we had to match the features available in the UK hospitals with those present in the HTD database. Additionally, any features with missing values exceeding 30% were excluded. Table 6.6 summarizes the final features included. Note that this task has more features than the generalizability experiments previously mentioned in Chapter 6.2, as those features were limited by the ones available at NHTD.

Table 6.6: Clinical predictors considered for COVID-19 diagnosis.

Category	Matched UK and Vietnam
Vital Signs	Heart rate, respiratory rate, systolic blood pressure, diastolic blood pressure, temperature
Blood Test	Haemoglobin, haematocrit, white cell count, platelets, mean cell volume, neutrophil count, lymphocyte count, monocyte count, eosinophil count, basophil count
Liver Function Tests	Alanine aminotransferase
Electrolytes	Sodium, potassium, creatinine

Data Pre-processing

Once we confirmed consistency in the units applied to identical features, we proceeded with implementing the identical pre-processing procedures outlined in Section 3.2.4. Consequently, all features were standardized to attain a mean of 0 and a standard deviation of 1. Additionally, we used the population median to imputate any missing values.

Evaluation Metrics

We adopt the same evaluation metrics outlined in Section 4.2.5, reporting on sensitivity, specificity, PPV, NPV, and AUROC with 95% CIs derived from 1000 bootstrapped samples from the test set. Significance tests, marked by p-values, assess model performance differences over 1000 bootstrapped comparisons, using a significance threshold of 0.05. These analyses are performed on final test sets. For fairness evaluation, we measure Equalized Odds variance as specified in Equations 4.16 and 4.17.

Hyperparameter Optimization

As in Chapter 4.2.6, we employ a grid search coupled with standard five-fold cross-validation on the training set to determine the optimal hyperparameters for all conventional supervised learning and reinforcement learning methods used.

Grid search was applied to all neural network models to ascertain the best configurations, encompassing factors like the number of hidden layers, nodes per layer, and learning rate. Specifically concerning adversarial debiasing, this exploration was carried out independently for both the predictor and adversary networks, as well as for the α hyperparameter. In the case of XGBoost, we explored a range of parameters such as learning rate, depth, and the number of trees. For reinforcement learning methods, optimization efforts were focused on parameters including the Q-network's layer count, nodes per layer, and learning rate.

Detailed information regarding the software, implementation, and final hyperparameter values selected for each model can be located in Appendix D.

Threshold Optimization

In alignment with the the process detailed in Chapter 4.2.7, we conduct threshold optimization, employing binary classification (COVID-19 positive or negative). As before, the objective is to identify a suitable threshold, optimized to attain sensitivities of 0.9, ensuring clinically acceptable performance levels in detecting positive COVID-19 cases.

Experimental Outline

Consistent with previous chapters, for each task, we use the training set to select hyperparameters and train the models, whilst the continuous validation set is used for ongoing validation and threshold adjustment (optimizing for a sensitivity of 0.9). After successful training and validation, the held-out test set is used to assess the performance of the final models.

As before, to establish a benchmark for comparing our neural network-based models, we start by training an XGBoost model utilizing the complete set of features outlined in Table 6.6. This model will additionally be used to assess the significance of variables for the classification task.

Following this, we proceed to train a neural network baseline using the entire feature set, facilitating comparison to various bias mitigation methods.

Concerning bias, our initial approach involves investigating potential bias sources across various hospital sites. As detailed in Chapter 3, we begin by utilizing t-Stochastic Neighbor Embedding (t-SNE) at the sample level to generate low-dimensional visualizations encompassing all positive COVID-19 cases collected from the four NHS sites and HTD. At a feature level, we then employ the Kolmogorov-Smirnov test to assess covariate shift. By identifying features exhibiting the most significant distribution shift between the UK sites and HTD, we subsequently remove these features and train models using a reduced feature set. This methodology effectively mitigates the impact of covariate shift in the training data.

For comparison, we then proceed to train new models using the complete feature set, incorporating bias mitigation techniques, namely adversarial debiasing (Chapter 4) and RL-based debiasing (Chapter 5).

This thorough comparison enables us to assess the models and techniques that contribute to enhancing classification fairness, thereby reducing undesired bias, in collaborative training across various hospital sites.

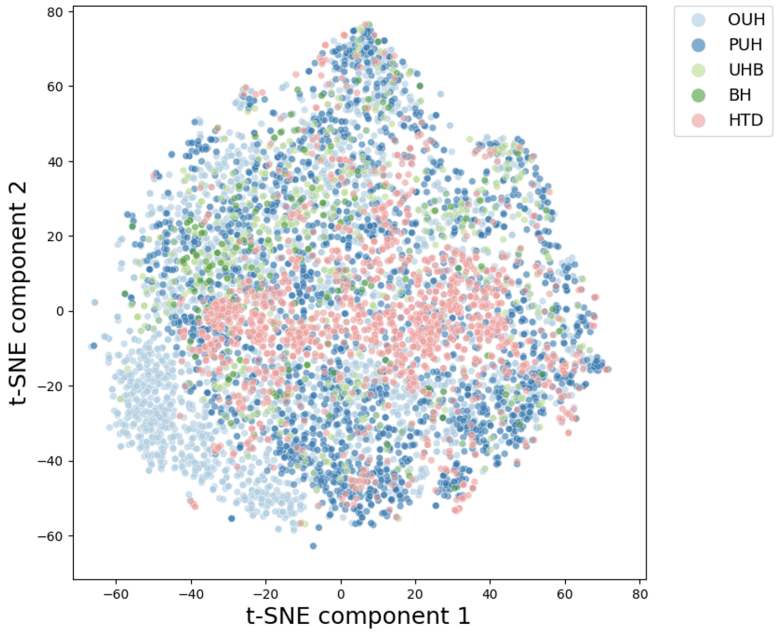

Figure 6.1: Visualization via t-SNE representation of datasets used in the study, including all positive COVID-19 cases across four NHS trusts and one Vietnamese hospital (OUH, PUH, UHB, BH, HTD).

6.3.2 Results

Starting with the t-SNE plot, in Figure 6.1, the presence of an isolated light blue cluster corresponding exclusively to a subset of presentations from one NHS site (OUH) suggests that the data from OUH has distinct features or characteristics that separates it from data collected at other sites. This is similar to results found in Chapter 3, when just comparing across the four NHS sites. Moreover, it underscores the importance of accommodating site-specific biases during model development, as these biases can be pronounced between hospitals from very different socioeconomic contexts.

During the data extraction period, COVID-19 prevalences at all four UK sites varied from 4.27% to 12.2%. The BH cohort exhibited the highest COVID-19 prevalence, attributed to the assessment timeframe covering the second wave of the UK pandemic from January 1, 2021, to March 31, 2021 (12.2% compared to 5.29% in PUH and 4.27% in UHB). As anticipated, the prevalence at HTD was

Table 6.7: Equalized Odds evaluation for hospital bias and COVID-19 status prediction test results across different models, optimized to sensitivities of 0.9. Classification metrics are reported alongside 95% confidence intervals based on 1,000 bootstrapped samples. Bolded values denoting the best (underlined) and second best Equalized Odds scores. Classification metrics are reported alongside 95% CIs, with bolded values denoting the best scores achieved on the test set. *RR: Respiratory Rate, T: Temperature.

Model	$M_{EO(TP)}$	$M_{EO(FP)}$	Sensitivity	Specificity	PPV	NPV	AUROC
XGBoost (All features)	0.086	0.264	0.801(± 0.012)	**0.788(± 0.004)**	**0.259(± 0.005)**	0.977(± 0.002)	**0.876(± 0.006)**
Neural Net. (All features)	0.075	0.246	0.811(± 0.012)	0.761(± 0.004)	0.239(± 0.005)	0.978(± 0.002)	0.866(± 0.007)
Neural Net. (Remove RR)	**0.052**	**0.196**	0.819(± 0.012)	0.674(± 0.004)	0.189(± 0.003)	0.976(± 0.002)	0.839(± 0.007)
Neural Net. (Remove T)	0.070	0.226	0.827(± 0.012)	0.732(± 0.005)	0.222(± 0.004)	**0.979(± 0.002)**	0.863(± 0.007)
Neural Net. (Remove RR and T)	**0.053**	**0.186**	**0.835(± 0.012)**	0.634(± 0.005)	0.174(± 0.003)	0.977(± 0.002)	0.837(± 0.008)
RL (debiasing) (All features)	0.056	0.204	0.829(± 0.015)	0.716(± 0.004)	0.213(± 0.003)	0.978(± 0.002)	0.858(± 0.007)
Adversarial (All features)	0.074	0.227	0.784(± 0.012)	0.773(± 0.004)	0.242(± 0.005)	0.975(± 0.002)	0.852(± 0.007)

notably higher (74.7%), given its exclusive focus as an infectious disease hospital, managing the most severe cases of COVID-19.

Among all cohorts from the UK and Vietnam, every matched feature exhibited a statistically significant difference in population median (Kruskal-Wallis, $p < 0.0001$). Comprehensive summary statistics, including medians and interquartile ranges, for vital signs and blood tests across all datasets are provided in Appendix D.

Previous chapters have demonstrated how XGBoost exhibited strong classification performance, serving as a reliable benchmark for assessing neural network-based models. When utilizing all features during training, the XGBoost model attained an AUROC of 0.876 (95% CI ±0.006) and a sensitivity of 0.801 (±0.012). These results align closely with findings from the work in previous chapters that used similar features and patient cohorts, reporting AUROC performances ranging from 0.855 to 0.901.

Regarding fairness, when assessed on the test set, the XGBoost model attained Equalized Odds values (represented as standard deviations) of 0.086 and 0.264 for true positive and false positive rates, respectively (Table 6.7).

To evaluate the significance of variables in the classification process, we conducted feature ranking for the XGBoost model by examining the importance scores assigned to each feature during the training phase. Figure 6.2 depicts the relative importance of the features used in the training. Notably, respiratory rate emerged as the most crucial variable, followed by granulocyte counts (specifically basophils and eosinophils) and temperature. This observation aligns with the feature rankings

identified in [133, 144] using SHAP, where both granulocyte counts and respiratory rate were found to be influential features in the classification of COVID-19 (as seen in Figure A.1 in Appendix A).

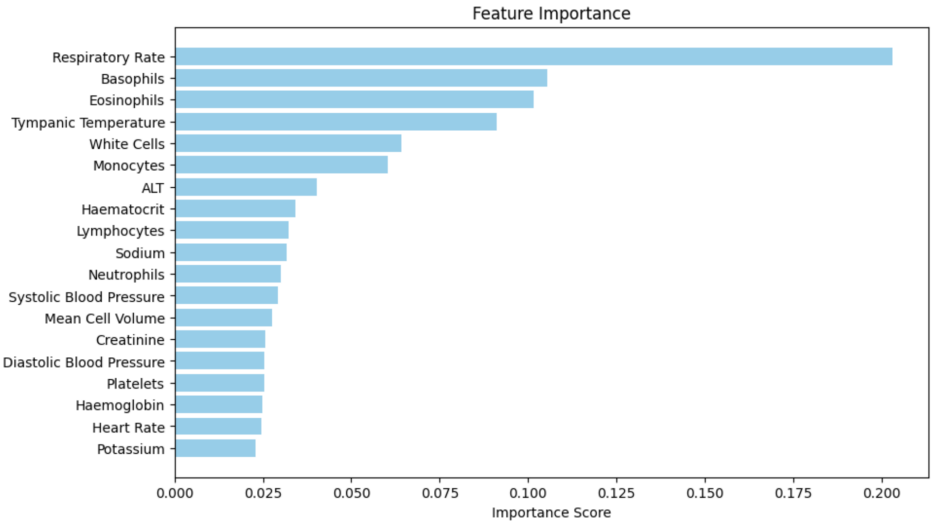

Figure 6.2: Feature ranking based on feature importance scores, obtained from the trained XGBoost model.

As in previous chapters, we then trained a conventional, fully-connected neural network utilizing the complete feature set as a reference point. This serves as a baseline for assessing the comparative impacts of different bias mitigation techniques. The neural network model demonstrated an AUROC of 0.866 (95% CI $\pm$0.006) and a sensitivity of 0.811 ($\pm$0.012) on the test set. This performance is within 1% of the XGBoost model, indicating the effectiveness of our initial training in establishing a robust neural network model. When comparing the neural network model to XGBoost, the difference in performance was found to be statisically significant ($p < 0.0001$ based on 1,000 bootstrapped iterations).

Regarding fairness, the baseline neural network model achieved Equalized Odds values of 0.075 and 0.246 for the TP and FP rates, respectively. These values, slightly lower than those obtained with XGBoost, suggest an enhancement in fairness compared to the XGBoost model.

To investigate potential sources of bias among hospital sites, we analyzed covariate shift using the Kolmogorov-Smirnov (KS) test for two-sample hypothesis

testing. Using each of the UK datasets as the reference, we compared the input features between each UK site and the HTD dataset. Our examination revealed respiratory rate and temperature as features displaying the most notable distribution shifts. Here, the KS statistics ranged from 0.636 to 0.804, with $p < 0.0001$ across all UK and HTD pairs (these p-values remained significant after Bonferroni Correction). In contrast, all other features showed KS statistics below 0.5 across all UK and HTD pairs. Notably, when scrutinizing feature distribution differences among solely the UK sites, the KS statistic remained below 0.5 across all features for all UK hospital pairs. This suggests a more pronounced bias between the UK sites and HTD, compared to biases exclusively between the UK sites.

Therefore, in addition to developing a comprehensive model incorporating all features, we proceeded to train separate models (without employing any inherent bias mitigation techniques) on feature subsets. These subsets excluded respiratory rate, excluded temperature, and excluded both features. By adopting this approach, we can evaluate the performance of models trained on feature sets with reduced covariate shift (achieved by excluding the features exhibiting the greatest shift), as compared to models that retain these features but implement bias mitigation techniques aimed at addressing shifts between different centers.

Training the same neural network with all features except respiratory rate resulted in a decreased AUROC of 0.839 (95% CI: ±0.007) when compared to the baseline). While sensitivity remained comparable at 0.819 (±0.012), there was a reduction in specificity from 0.761 (±0.004) to 0.674 (±0.004) compared to the baseline neural network. In contrast, excluding temperature during training led to a much smaller decline in performance, with the model achieving an AUROC of 0.863 (±0.007) ($p = 0.053$ when compared to the baseline using 1,000 bootstrapped iterations). In this scenario, there was a slight increase in sensitivity to 0.827 (±0.012), but at the cost of a decrease in specificity to 0.732 (±0.005).

Omitting both respiratory rate and temperature from training resulted in the model achieving an AUROC of 0.837 (±0.008), comparable to the situation where only respiratory rate was excluded. In this instance, sensitivity improved to 0.835

($\pm$0.012). Once again, this heightened sensitivity came at the cost of a reduction in specificity to 0.634 ($\pm$0.005). Although the deviation in AUROC from the standard model was statistically significant ($p < 0.0001$), it was found to be statistically similar to the model trained on all features except respiratory rate ($p = 0.168$).

Regarding fairness, models trained with the exclusion of respiratory rate and without both respiratory rate and temperature demonstrated a distinct enhancement, as evidenced by decreased Equalized Odds for TP and FP rates, ranging between 0.052-0.053 and 0.186-0.196, respectively. When only temperature was excluded from the training set, there was still an improvement in TP and FP Equalized Odds, albeit to a lesser extent than when respiratory rate was excluded, achieving values of 0.070 and 0.226, respectively.

Despite the apparent improvement in fairness achieved by removing features exhibiting the most bias (in terms of data drift), there was a notable decline in classification performance. As both respiratory rate and temperature were identified as important features during training (according to the XGBoost rankings in Figure 6.2), we proceeded to train two state-of-the-art bias mitigation models - a reinforcement learning debiasing model and an adversarial debiasing model. These models enable us to leverage all features while mitigating site-specific biases during the training process.

Both RL and adversarial debiasing techniques achieved AUROCs similar to the baseline neural network, with values of 0.858 ($\pm$0.007) and 0.852 ($\pm$0.007), respectively. Although slightly lower than the baseline neural network's performance ($p = 0.001$ for RL and $p < 0.0001$ for adversarial models), it outperformed all NNs trained on reduced feature sets, except for the neural network trained without temperature. When comparing RL and adversarial debiasing models to those trained on reduced feature sets, the differences in performance were found to be statistically significant ($0 < p < 0.021$).

The RL debiasing model exhibited a notable improvement in fairness, with TP and FP Equalized Odds decreasing to 0.056 and 0.204, respectively. These values are comparable to the Equalized Odds observed in models trained with the

Table 6.8: COVID-19 status prediction test results on the HTD subset, across different models. Model performance is optimized to sensitivities of 0.9, and PPV and NPV are reported using true prevalences. Metrics are reported alongside 95% confidence intervals based on 1,000 bootstrapped samples. Bolded values represent the best (underlined) and second best scores. *RR: Respiratory Rate, T: Temperature.

Model	Sensitivity	Specificity	PPV	NPV	AUROC
XGBoost (All features)	**0.989(±0.011)**	0.159(±0.067)	0.787(±0.013)	**0.824(±0.164)**	**0.836(±0.040)**
Neural Net. (All features)	**0.971(±0.018)**	0.170(±0.070)	0.786(±0.015)	**0.652(±0.171)**	0.723(±0.051)
Neural Net. (Remove RR)	0.931(±0.027)	0.205(±0.075)	0.786(±0.017)	0.486(±0.131)	0.694(±0.054)
Neural Net. (Remove T)	0.957(±0.021)	0.216(±0.072)	**0.793(±0.016)**	0.613(±0.145)	0.768(±0.047)
Neural Net. (Remove RR and T)	0.917(±0.029)	0.216(±0.074)	0.786(±0.016)	0.452(±0.119)	0.677(±0.055)
RL (All features)	0.942(±0.025)	**0.239(±0.074)**	**0.795(±0.017)**	0.568(±0.129)	**0.781(±0.047)**
Adversarial (All features)	0.938(±0.028)	**0.227(±0.080)**	0.792(±0.015)	0.541(±0.029)	0.757(±0.046)

exclusion of respiratory rate and without both respiratory rate and temperature, although slightly higher. Importantly, this represents a significant enhancement in fairness compared to the baseline neural network. Conversely, the adversarial debiasing model only marginally improved upon the baseline, achieving Equalized Odds of 0.074 and 0.227 for TP and FP standard deviations, respectively.

Full numerical results including AUROC, sensitivity, specificity, PPV, NPV, and Equalized Odds metrics can be found in Table 6.7.

When analyzing the subset of test data from HTD, the XGBoost model demonstrated the highest AUROC among all models, achieving a score of 0.836 (95% CI ±0.040). This figure is slightly lower when compared to the AUROCs achieved across each UK site, ranging from 0.859 to 0.909. In contrast, the standard neural network (trained using all features) attained a lower AUROC at HTD than XGBoost, with a score of 0.723 (±0.051) ($p < 0.0001$ when comparing the difference in performance between XGBoost and the baseline neural network). This is notably lower than the AUROCs attained on the UK datasets, which ranged from 0.853 to 0.901. Despite XGBoost's robust performance at HTD compared to the baseline neural network, performance across the UK sites was similar for both models.

Models trained with the exclusion of respiratory rate and without both respiratory rate and temperature exhibited lower AUROC performance at HTD, suggesting a potential decrease in model generalizability and, consequently, increased algorithmic bias between different hospital sites. In this context, the AUROC at HTD was 0.677 (±0.055) and 0.694 (±0.054) for each model, respectively ($p < 0.0001$

and $p = 0.032$ when comparing the difference in performance to the baseline neural network, respectively). This contrasts with significantly higher AUROCs ranging from 0.826-0.885 and 0.812-0.881 on the UK datasets, respectively. Models trained without temperature improved in terms of performance on the HTD dataset, relative to the baseline, improving to 0.723 ($\pm$0.051) ($p = 0.008$). And, similar to the baseline, AUROCs on the UK datasets ranged from 0.845 to 0.911. Again, despite varying scores at HTD compared to the XGBoost and baseline neural network models, performances across the UK sites remained similar across all models.

The RL debiasing model attained the second-highest AUROC on the HTD dataset, achieving a score of 0.781 ($\pm$0.047). This was found to be statistically significant with $p < 0.0001$ when comparing the difference in performance to the baseline neural network and the models trained on reduced feature sets. Once again, this figure is lower than the AUROCs achieved at the UK sites, which ranged from 0.842 to 0.888. Adversarial debiasing similarly reached a high AUROC score of 0.757 ($\pm$0.046) on the HTD dataset ($p = 0.072$ when comparing the difference in performance to the baseline neural network, and $p = 0.002$ and 0.004 when compared to the models trained on reduced feature sets), in contrast to AUROCs ranging from 0.830 to 0.849 on the UK datasets. When using algorithm-level bias mitigation methods, generalizability across the UK sites and HTD seems to improve compared to models trained using reduced feature sets. As observed previously, while there were varying performances at HTD, performances achieved across the UK sites remained similar across all models.

Full numerical results for HTD including AUROC, sensitivity, specificity, PPV, and NPV can be found in Table 6.8.

6.3.3 Discussion

In general, we observed that models incorporating some form of bias mitigation, whether through the removal of biased features or through the inclusion of bias mitigation training methods, exhibited greater fairness (with respect to Equalized Odds) compared to those without such considerations. It is important to note that

this reduction in bias came at a modest cost to performance, as both the removal of features from the training set and the implementation of bias mitigation at the training-level resulted in a decline in performance. The act of excluding features from training removes potentially valuable information that the model could learn from, emphasizing the need to strike a balance when constructing models with the dual objectives of mitigating undesirable biases and training a robust classifier.

We observed that excluding respiratory rate from model training resulted in a more pronounced decrease in AUROC (compared to the baseline model trained on all features) than a model trained without temperature. This aligns with respiratory rate being identified as the most influential feature, relative to all other features, in determining the presence of COVID-19, as illustrated in Figure 6.2. Therefore, it is to be expected that omitting this feature from model development would have the greatest impact on test performance. This is reinforced by the similarity in performance between the model trained without respiratory rate and the model trained without both respiratory rate and temperature, with further removal of temperature showing no significant impact.

Similarly, as respiratory rate emerged as one of the most biased features (exhibiting greatest data drift) between the UK hospital sites and HTD in Vietnam, excluding this feature from training resulted in improved fairness, as evidenced by noticeable enhancements in both TP and FP Equalized Odds. In contrast, removing temperature from training led to only a slight improvement in fairness. Despite temperature being identified as highly biased across different sites, its influence on classification performance was comparatively lower than that of respiratory rate. Consequently, its presence or absence had a lesser impact on Equalized Odds.

Generally, discovering respiratory rate and temperature as the most biased features between the UK hospitals and HTD is not unexpected. Given that HTD is a specialized hospital receiving referrals from other medical facilities, its high workload is an important factor to consider. Consequently, the meticulous measuring of respiratory rate may not have been consistently performed unless a patient appeared unwell. In busy LMIC wards, reduced staffing may also contribute to less

precise counting. Similarly, the disparity in temperature readings may be attributed to monitoring differences, as HTD utilizes auxillary non-digital thermometers, resulting in less detailed output. Nonetheless, temperature measurements can exhibit significant variability in any case, thus using a single temperature measurement may be an unreliable indicator [275]. Subsequent experiments may explore the option of excluding temperature as a variable if it proves to be inconsistent, exhibits substantial data shift across different sites, or is determined to have limited impact on the classification.

Regarding algorithmic-level bias mitigation methods, the decrease in performance, as measured by AUROC, was notably less compared to the direct removal of biased features; however, there was still a slight performance decrease. Similar findings have been reported in prior studies exploring bias mitigation methods, where improvements in fairness were achieved at the expense of a marginal reduction in performance [81, 86, 94, 276].

The RL debiasing method significantly enhanced Equalized Odds, achieving a level comparable to models trained on reduced feature sets, while maintaining robust classification performance (similar to standard models trained on all features). In contrast, although adversarial debiasing also upheld strong classification performance, its impact on Equalized Odds improvement was comparatively modest. This difference may be attributed to the standard supervised learning setting employed in adversarial debiasing, where cross-entropy loss provides a learning signal irrespective of the presented data, which affects error aggregation and can skew a model (further details of which were highlighted in Chapter 5). Thus, in our case, the RL approach was found to have more success in mitigating the risk of biasing the model towards the UK sites (i.e. the location with the most data points in training). Nevertheless, since there is no universal solution applicable to all datasets and machine learning scenarios, both options (among others) should be considered for different tasks.

We found that generalizability, measured by AUROC, was optimal when using the XGBoost and baseline neural network models. However, corresponding fairness metrics were found to be the poorest amongst all models. It is crucial to highlight

that the Equalized Odds metrics used in the evaluation are based on TP and FP rates, determined after thresholding. Consequently, despite a high AUROC, the model exhibited bias towards the distributions in the UK datasets, resulting in the classification threshold performing suboptimally on the HTD dataset.

Moreover, the removal of features with the most distribution drift also led to a significant decrease in generalizability. This could be attributed to the fact that these features were identified as highly influential in accurately classifying COVID-19. Therefore, their removal made it more challenging to correctly classify patients, especially those from an independent hospital site or distribution.

On the other hand, with the application of bias mitigation techniques, generalizability increased, as evidenced by the improved AUROC on the HTD dataset compared to both the baseline neural network and neural network models trained on a reduced number of features. However, the neural network trained on all features except temperature slightly outperformed the adversarial debiasing model, which again, may relate to the trade-off between fairness and accuracy [81, 276]. Although the performance of these models did not match those of XGBoost, fairness, as measured by Equalized Odds, significantly improved. The focus of bias mitigation methods on reducing bias between different hospital sites likely contributed to increased classification accuracy at HTD, making the algorithm less biased toward the UK sites.

Although achieving widespread generalizability is desirable for scalability, cost-effectiveness, and relevance to diverse cohorts and environments, it is often unattainable. This limitation became evident when comparing performance on the HTD dataset with that on the UK datasets. When conducting subset analysis for each site independently, we noted that AUROCs across all UK datasets were consistently within a similar range across all tested models. However, there was significant variability in performance on the HTD dataset. Despite improvements in performance at HTD through the application of bias mitigation techniques, the AUROC remained significantly lower than the scores achieved on the UK hospital datasets. This is likely due to the fact that the HTD dataset represents the sole

LMIC hospital, whereas the UK datasets are all part of the NHS, sharing more similarities with each other. Factors contributing to this divergence between settings may include concept drift in disease patterns (such as alterations in presentation, prevalence, and characteristics), population variability (such as patients at one center may not represent those in another location), evolving medical practices (such as changes in diagnostic, treatment, and management methods), and data drift (such as changes in patient behaviors, trends, or data collection methods) [238, 240–242]. Moreover, the majority of the data originates from the UK sites. Consequently, despite efforts to mitigate bias, the model remains more attuned to the characteristics of the UK datasets, resulting in better performance on these specific subsets.

7 Additional Case Studies

7.1 Introduction

7.1.1 Overview

In this chapter, we delve deeper into bias mitigation techniques in machine learning by introducing additional clinical case studies and use cases. Building upon the knowledge established in preceding chapters, we explore additional real-world scenarios where biases pose significant challenges, further emphasizing the critical need for effective mitigation strategies.

It is essential to test the proposed methods across diverse applications to assess their effectiveness and generalizability thoroughly. By exploring various applications, our aim is to provide a comprehensive evaluation of the performance of these methods. This approach enables us to showcase how these methods can be applied and refined to effectively address diverse challenges.

While our previous discussions primarily focused on mitigating biases originating from different hospital sites, our attention now shifts towards combating another potential source of bias, specifically patient ethnicity. This is particularly relevant in scenarios where datasets exhibit unequal representation across ethnic groups. As machine learning models are increasingly deployed across various domains, ensuring fairness and equity becomes crucial, especially in contexts where demographic

factors like ethnicity can significantly impact outcomes. Therefore, in addition to the previously discussed COVID-19 screening task, we also examine a new task and dataset related to predicting patient discharge status.

7.1.2 Ethnicity Bias in Healthcare

As previously highlighted, implicit bias, including those based on ethnicity, represents a significant source of bias within healthcare. These biases are unconscious, however, they can arise due to a variety of interconnected factors, including socioeconomic factors, cultural beliefs, and access to healthcare services, all of which can lead to disparities in healthcare utilization and outcomes. For example, ethnic minorities may experience barriers to accessing quality healthcare, leading to differences in disease detection, treatment initiation, and health outcomes compared to majority populations.

The landmark report *Unequal Treatment* highlighted significant inequities in healthcare, drawing attention to the racial and ethnic disparities in the occurrence, management, and complications associated with conditions like hypertension, heart disease, and diabetes [277]. It revealed that these disparities persist even when factors such as insurance coverage, income, age, and condition severity are similar across different racial and ethnic groups. The sources of these disparities are multifaceted, involving various stakeholders within the healthcare system, including the institutions themselves, healthcare providers, plan managers, and the patients [277, 278].

One of the critical findings of the report was that biases, stereotypes, prejudice, and clinical uncertainties harbored by healthcare providers might contribute significantly to these disparities. Further research also supports the claim that healthcare providers' perceptions of their patients can vary significantly based on the race or ethnicity of those patients, providing evidence of bias within the healthcare system. For instance, one notable study highlighted how cardiologists' perceptions of black patients differed markedly from their perceptions of white patients[279]. The study revealed that cardiologists generally viewed black patients as less intelligent,

less likeable, less friendly, and more likely to engage in risky behaviors and non-compliance with medical advice compared to white patients. Such perceptions can adversely affect the quality of care provided, including the thoroughness of examinations, the recommendations made by healthcare professionals, and the level of communication and trust established between patients and providers.

It has also been observed that racial and ethnic minority patients in the USA were more prone to decline treatments compared to their white peers [277]. This phenomenon can be attributed to a range of factors, such as previous encounters with discrimination, a general distrust towards the healthcare system, and cultural variances in how health is perceived and managed. Supporting this observation, another study revealed that the sense of discrimination felt by patients during medical interactions points to a potential bias on the part of healthcare providers [280]. They found that relative to white patients, individuals from minority groups are more inclined to think that their treatment, and the respect they receive from medical personnel, would improve if they were of a different racial group.

On a similar note, healthcare providers may also unknowingly exhibit biases in diagnostic processes or treatment decisions based on a patient's ethnicity. This can result in differential treatment recommendations, potentially leading to underdiagnosis, misdiagnosis, or inadequate treatment for certain ethnic groups. For example, as previously mentioned, one study found that black patients in the emergency room received pain medication at a 40% lower rate compared to white patients [66]. Such biases can further exacerbate disparities in health outcomes and contribute to the unequal representation of ethnicities in clinical datasets.

Despite ongoing efforts to mitigate these disparities, racial and ethnic minority groups, including Black, Hispanic, Asian, Pacific Islander, and American Indian/Alaska Native populations, continue to experience inferior healthcare services and outcomes. The Agency for Healthcare Research and Quality, in its annual disparities report, has consistently documented that these disparities are widespread within the United States [281]. Up to 2013, it was reported that Blacks, Hispanics, and American Indians/Alaska Natives received lower-quality care for 40% of the

assessed quality measures, while Asians received worse care for 20% of the measures. These findings underscore the persistent challenge of achieving equity in healthcare, highlighting the need for targeted interventions to address the root causes of racial and ethnic disparities in health care and outcomes.

With respect to machine learning, these implicit ethnicity biases pose a complex challenge. Since data can reflect these biases, models may inadvertently learn and reinforce these biases. For example, if certain ethnic groups are systematically disadvantaged or marginalized in healthcare, biased data may lead to discriminatory outcomes in machine learning-based medical decision-making.

In addition to implicit bias, bias can emerge from unequal representation of ethnicities in specific regions, thereby influencing the data. When certain ethnic groups are either overrepresented or underrepresented in the training data, the model might develop a tendency to produce predictions that benefit the predominant group while potentially discriminating against minority groups. This phenomenon restricts the model's ability to generalize effectively across diverse populations, consequently leading to disparities in its performance among various ethnicities. As a consequence, machine learning models may be less precise or less accurate when predicting on underrepresented groups.

In general, these combined biases have the potential to erode trust and faith in healthcare systems as well as in any machine learning models trained on existing data. When individuals perceive the predictions of a model as unjust or discriminatory, they may hesitate to rely on or trust the model, which can ultimately result in a reluctance to access healthcare services or use machine learning applications in critical decision-making scenarios. Hence, it is essential to understand the roots of bias and actively tackle them during subsequent analyses, such as those carried out in machine learning.

7.2 Case Study: COVID-19 Screening

7.2.1 Overview

In this section, we delve into the development and training of models aimed at performing COVID-19 screening while actively mitigating ethnicity bias.

7.2.2 Methods

In this chapter, we use the UK datasets introduced in Chapter 3, but adopt different stratifications for model training, continuous validation, and testing purposes. Previously, when trying to counteract bias stemming from hospital locations, it was imperative to include data from all sites in both the training and validation datasets. Consequently, we amalgamated all datasets to ensure representation from each site. However, to address ethnicity bias, it is essential for all ethnicities to be represented in the training and validation sets, with less emphasis on the necessity for all locations to be present. Hence, this chapter provides an opportunity to perform external validation—using independently derived datasets to assess the performance of models trained on the original data[282]—across multiple cohorts.

Training, Continuous Validation, and Test Dataset Splitting

For the training ($\mathcal{D}_{train}$) and continous validation ($\mathcal{D}_{val}$) sets used in the ethnicity debiasing models, we used patient presentations exclusively from OUH. As seen in Chapter 3, from OUH, we curated two data extracts - one from the first wave of the COVID-19 epidemic in the UK (December 1, 2019 to June 30, 2020), and one from the second wave (October 1, 2020 – March 6, 2021). As before, from the "wave one" dataset, we only included the positive cases (as determined through PCR tests) in training; and from the "wave two" dataset, we included both positive COVID-19 cases (by PCR) and negative controls. Recall that this was done to ensure that the label of COVID-19 status was correct during training, as there was uncertainty in the viral status of patients who were untested or tested negative during wave one.

To reasonably evaluate classification performance with respect to ethnicity, we removed any presentations where the label for ethnicity was ambiguous, including

those labeled as "unknown", "mixed", or "other". This resulted in 18,687 patients used in training and validation, including 2,083 of which were COVID-19 positive. A ratio of 80:20 was used to split the OUH cohort into training and continous validation sets, respectively. We then performed external validation on three independent patient cohorts from PUH, UHB, and BH ($\mathcal{D}_{test1}$, $\mathcal{D}_{test2}$, $\mathcal{D}_{test3}$, respectively), totalling 38,964 admitted patients, including 1,963 of which were COVID-19 positive. From Table 7.1, we can see that ethnicity is heavily skewed in our training dataset, making it a possible source of bias during training. The features included are the same as those listed in Chapter 3.3.

Table 7.1: Summary of number of patients, COVID-19 positive cases, and ethnicity distribution for training, validation, and external test set cohorts included in the ethnicity debiasing task.

	Training ($\mathcal{D}_{train}$)	Continuous Validation ($\mathcal{D}_{val}$)	External Tests ($\mathcal{D}_{test1}$, $\mathcal{D}_{test2}$, $\mathcal{D}_{test3}$)		
	OUH	OUH	PUH	UHB	BH
Total Patients	14,949	3,738	29,103	8,730	1,131
COVID-19 positive (PCR)	1,672	411	1,478	347	138
Ethnicity:					
White (%)	14,293 (95.6)	3,574 (95.6)	28,704 (98.6)	6,848 (78.4)	1,024 (90.5)
South Asian (%)	370 (2.5)	93 (2.5)	170 (0.6)	1357 (15.5)	71 (6.3)
Black (%)	243 (1.6)	61 (1.6)	187 (0.6)	484 (5.5)	36 (3.2)
Chinese (%)	43 (0.3)	10 (0.3)	42 (0.1)	41 (0.5)	0 (0.0)

Baseline Evaluation and Model Comparators

We showcase outcomes for baseline models, comprising a standard neural network, XGBoost, and a conventional RL classification model without any debiasing element. Additionally, we highlight the results for the two most effective bias mitigation techniques: an adversarial debiasing framework and an RL classification model integrated with a debiasing component—specifically, the methods proposed in Chapters 4 and 5, respectively.

Data Pre-processing

After ensuring consistency in the units applied to identical features, we proceeded to implement the same pre-processing procedures detailed in Section 3.2.4. As a result, all features were standardized to achieve a mean of 0 and a standard

deviation of 1. Additionally, we employed population median imputation to address any missing values.

Evaluation Metrics

We adopt the same evaluation metrics outlined in Section 4.2.5, reporting on sensitivity, specificity, PPV, NPV, and AUROC with 95% CIs derived from 1000 bootstrapped samples from the test set. Significance tests, marked by p-values, assess model performance differences over 1000 bootstrapped comparisons, using a significance threshold of 0.05. These analyses are performed on final test sets. For fairness evaluation, we measure Equalized Odds variance as specified in Equations 4.16 and 4.17.

Hyperparameter Optimization

As in Chapter 4.2.6, we use a grid search coupled with standard five-fold cross-validation on the training set to determine the optimal hyperparameters for all conventional supervised learning and reinforcement learning methods.

Grid search was applied to all neural network models to ascertain the best configurations, encompassing factors like the number of hidden layers, nodes per layer, and learning rate. Specifically concerning adversarial debiasing, this exploration was carried out independently for both the predictor and adversary networks, as well as for the α hyperparameter. In the case of XGBoost, we explored a range of parameters such as learning rate, depth, and the number of trees. For reinforcement learning methods, optimization efforts were focused on parameters including the Q-network's layer count, nodes per layer, and learning rate.

Detailed information regarding the software, implementation, and final hyperparameter values selected for each model can be located in Appendix E.

Threshold Optimization

In alignment with the findings detailed in Chapter 4.2.7, we conduct threshold optimization, employing binary classification (COVID-19 positive or negative) to align with the "green–amber–blue" categorization system employed by NHS Trust

policy. Here, we optimize outcomes to attain sensitivities of 0.9, ensuring clinically acceptable performance levels in detecting positive COVID-19 cases.

7.2.3 Results

Table 7.2: Equalized Odds evaluation for ethnicity bias and COVID-19 status prediction test results across different models, optimized to sensitivities of 0.9. Metrics are reported alongside 95% confidence intervals based on 1,000 bootstrapped samples. Bolded values denote the best (underlined) and second best Equalized Odds scores. Classification metrics are reported alongside 95% CIs, with bolded values denoting the best scores achieved on the test set.

Test Set	Model	$M_{EO(TP)}$	$M_{EO(FP)}$	Sensitivity	Specificity	PPV	NPV	AUROC
PUH								
	RL (debiasing)	**0.047**	0.037	0.876 (±0.017)	0.512 (±0.006)	0.088 (±0.005)	0.987 (±0.002)	0.834 (±0.013)
	RL	**0.048**	0.031	0.872 (±0.017)	0.518 (±0.006)	0.088 (±0.005)	0.987 ±0.002)	0.838 (±0.013)
	Adversarial	0.050	**0.014**	0.879 (±0.017)	0.595 (±0.005)	0.104 (±0.005)	0.989 (±0.001)	0.865 (±0.012)
	Neural Net.	0.066	**0.028**	0.890 (±0.016)	**0.631 (±0.005)**	**0.114 (±0.006)**	0.991 (±0.002)	0.875 (±0.012)
	XGBoost	0.133	0.053	**0.919 (±0.013)**	0.532 (±0.006)	0.095 (±0.005)	**0.992 (±0.001)**	**0.882 (±0.011)**
UHB								
	RL (debiasing)	**0.057**	**0.041**	**0.879 (±0.034)**	0.574 (±0.011)	0.079 (±0.009)	0.991 (±0.003)	0.849 (±0.025)
	RL	0.155	0.044	0.876 (±0.035)	0.538 (±0.011)	0.073 (±0.008)	0.991 (±0.003)	0.834 (±0.026)
	Adversarial	0.072	**0.039**	0.867 (±0.035)	0.637 (±0.010)	0.090 (±0.010)	0.991 (±0.003)	0.865 (±0.025)
	Neural Net.	**0.069**	0.052	0.873 (±0.035)	**0.667 (±0.010)**	**0.098 (±0.011)**	**0.992 (±0.002)**	**0.868 (±0.026)**
	XGBoost	0.106	0.071	0.867 (±0.036)	0.585 (±0.010)	0.080 (±0.009)	0.991 (±0.003)	0.854 (±0.025)
BH								
	RL (debiasing)	**0.030**	**0.010**	**0.935 (±0.041)**	0.691 (±0.029)	0.296 (±0.043)	**0.987 (±0.008)**	**0.923 (±0.031)**
	RL	0.055	0.057	0.906 (±0.049)	0.668 (±0.029)	0.275 (±0.031)	0.981 (±0.011)	0.898 (±0.035)
	Adversarial	**<0.001**	0.045	0.877 (±0.055)	0.779 (±0.026)	0.356 (±0.051)	0.979 (±0.011)	0.912 (±0.033)
	Neural Net.	0.059	0.039	0.870 (±0.057)	**0.803 (±0.024)**	**0.380 (±0.053)**	0.978 (±0.010)	0.912 (±0.033)
	XGBoost	0.107	**0.036**	0.928 (±0.044)	0.644 (±0.030)	0.266 (±0.040)	0.985 (±0.010)	0.908 (±0.033)

After training models on patient cohorts from OUH, we externally validated our models across three external patient cohorts from PUH, UHB, and BH. All models achieved reasonably high AUROC scores across all test sets (Table 7.2), comparable to those reported in the clinical validation study highlighted in Chapter 2.5, demonstrating that we trained strong classifiers to begin with. AUROC scores for predicting COVID-19 status stayed relatively consistent across all test sets, achieving the highest performances on the BH cohort (PUH: AUROC range 0.834-0.882; UHB: 0.834-0.868; BH: 0.897-0.923). With respect to the model used, all models achieved similar AUROCs; however, the highest AUROCs were generally achieved by the standard supervised learning-based models, including the adversarial model, and both weighted and unweighted XGBoost and neural network models (mean AUROCs of 0.869, 0.857, 0.881, 0.885, 0.881 for RL-based debiasing, RL, adversarial, neural network, XGBoost models, respectively). Using a sensitivity

configuration of 0.9, we obtained consistent scores for sensitivity across all models and cohorts (PUH: sensitivity range 0.872-0.919; UHB: 0.862-0.879 ; BH: 0.862-0.935), with RL-based debiasing achieving the highest sensitivities on the UHB and BH test sets (however, it should be noted that RL had either the lowest or second lowest specificities). And, as seen in previous chapters, our models achieved high prevalence-dependent NPV scores (>0.978), demonstrating the ability to exclude COVID-19 with high-confidence. Furthermore, these results demonstrate that an RL-based debiasing paradigm is more generalizable in diverse environments, with superior AUROC for the BH cohort, and superior sensitivity on the BH and UHB cohorts, which are the two most ethnically diverse cohorts, according to Table 7.1.

Although predictive performance of the RL-based debiasing model only varied slightly with respect to other models, the difference in accuracy of the RL-based debiasing model compared to that of other models was found to be statistically significant ($p<0.0001$, based on 1,000 bootstrapped samples).

In terms of fairness, the RL-based debiasing model achieved the best performance overall, achieving either the best or second best Equalized Odds performances (for both TP and FP rates) across all external test cohorts, except for the FP metric for PUH (Table 7.2). The adversarial model achieved the second best performance overall, usually achieving the best or second best scores for one of TP or FP Equalized Odds metrics. In general, models with an added dynamic debiasing functionality (i.e. RL-based or adversarial debiasing models) demonstrably improved Equalized Odds.

7.2.4 Discussion

Similar to the comparisons made in Chapters 4 and 5, our observations reveal that debiasing models produce less biased outcomes when compared to models without bias mitigation components. Specifically, the debiasing method using reinforcement learning demonstrated superior performance, typically achieving a combination of the best and second-best Equalized Odds for true positive and false positive rates. Following closely, the adversarial debiasing method achieved the second best performance, typically achieving the best Equalized Odds score for

either true positive rate or false positive rate, across all external test sets. Again, as mentioned in previous chapters, the superior performance of the RL debiasing approach compared to supervised-learning-based bias mitigation techniques can be attributed to the ability of a RL setup to regulate the propagation of learning signals and refine error aggregation processes.

Nevertheless, as previous chapters have shown, despite the mitigation of bias, these models did not fully satisfy the criteria for Equalized Odds. This persistent gap could be attributed to the data imbalance regarding the sensitive attribute [102], with a substantially larger volume of data available from white patients compared to other ethnicities. Additionally, the use of neural networks in both adversarial debiasing and Q-learning within the RL method might lead to inconsistent outcomes due to skewed distributions within the data.

In terms of meeting Equalized Odds criteria, the advantage of employing a RL framework was more evident and pronounced, particularly in the COVID-19 task aimed at mitigating inter-hospital biases. This was reflected in noticeable enhancements in the standard deviation of true positive and false positive rates for RL results compared to other models. The discernible difference in performance may stem from the larger volume of training data used in that task, compared to the COVID-19 task here, which focused on ethnicity debiasing. Specifically, 14,949 patients compared to 43,754 patients were used to train the ethnicity and hospital debiasing tasks, respectively. The larger amount of training data may have facilitated the models' ability to accurately distinguish between various classes, both for the primary task and the sensitive attribute. This is further exemplified by COVID-19 diagnosis performance, where superior predictive performance was achieved for the hospital-site mitigation task, which used a larger training set.

7.3 Case Study: Patient Discharge Status Prediction

7.3.1 Overview

To demonstrate the application of fairness-aware algorithms in a different predictive scenario (independent of the COVID-19 prediction task previously used), we introduce a task centered on predicting patient discharge status. Here, we also focus on debiasing ethnicity, demonstrating the generalizability of our method to a new, independent clinical task.

Accurately predicting the discharge status of patients carries substantial significance for hospitals, as it influences various aspects such as resource allocation, cost management, bed allocation, post-discharge care coordination, patient flow efficiency, re-admission rates, and financial outcomes. Consequently, machine learning decision support tools play a pivotal role in facilitating well-informed discharge decisions. This, in turn, fosters the advancement of patient-centered care initiatives and operational efficiency within healthcare facilities.

7.3.2 Critical Care Data

The expanding availability of clinical data alongside the advancements in machine learning technologies has played a pivotal role in tackling various healthcare challenges across different domains, encompassing risk assessment and prediction in acute, chronic, and critical care settings. However, within this spectrum of healthcare domains, critical care emerges as an especially data-rich and complex field. This distinction arises primarily from the nature of patient monitoring in ICUs, where individuals receive continuous and intensive medical attention due to severe illnesses or injuries.

In ICUs, patients are subjected to constant monitoring through a plethora of medical devices, sensors, and instruments, generating vast streams of data on their physiological parameters, vital signs, laboratory results, medication administration,

and other clinical metrics [283–285]. These data streams, often referred to as high-dimensional time-series data, offer a wealth of information regarding the patients' health status, disease progression, and response to treatment interventions. However, the sheer volume and complexity of these data present significant challenges in extracting actionable insights and making informed clinical decisions [284, 286, 287].

Machine learning, with its capacity to analyze large volumes of complex data and identify patterns and relationships not easily discernible by humans, has emerged as a powerful tool in leveraging ICU data for clinical decision-making. By applying machine learning algorithms to ICU data, healthcare providers can enhance their ability to predict adverse events, identify patients at risk of deterioration, optimize treatment strategies, and improve patient outcomes [288, 289]. Moreover, machine learning techniques enable the development of predictive models that can assist clinicians in early identification of conditions such as sepsis [290], acute respiratory distress syndrome [291], and organ failure [292, 293], allowing for timely interventions and improved patient care.

Machine learning algorithms can aid in optimizing resource utilization and operational efficiency within ICUs by predicting patient length of stay, facilitating bed management, and streamlining workflow processes [294, 295]. Additionally, these algorithms can support clinical research efforts by identifying novel biomarkers, elucidating disease mechanisms, and uncovering new avenues for therapeutic intervention in critical care settings. In the following sections, we focus on the task of patient discharge status prediction.

Overall, the combination of increasing availability of clinical data and advancements in machine learning holds tremendous promise for revolutionizing critical care delivery. By harnessing the wealth of data generated in ICUs and deploying sophisticated machine learning algorithms, healthcare providers can enhance patient monitoring, optimize clinical decision-making, and ultimately improve patient outcomes in this high-stakes healthcare environment.

7.3.3 Patient Discharge Status Prediction

Accurately predicting the discharge status or mortality of patients is of paramount importance for hospitals, as it has a profound impact on multiple aspects of healthcare delivery and operational management, such as predicting resource allocation [296–298]. Hospitals must efficiently allocate resources such as staff, medical supplies, and equipment based on patient needs. By predicting discharge status or mortality, hospitals can better align their resources to ensure that they are adequately prepared to provide the necessary care and support to patients, optimizing the allocation of resources and improving patient outcomes. One specific aspect is bed allocation within hospitals [299, 300]. Hospitals must manage bed availability efficiently to accommodate incoming patients while ensuring timely discharge for those who no longer require hospital care. By accurately predicting discharge status, hospitals can optimize bed utilization, reduce overcrowding, and minimize wait times for incoming patients.

Cost management is another crucial factor affected by accurate prediction of discharge status or mortality [301]. Predicting when patients will be discharged or their likelihood of mortality allows hospitals to optimize costs by reducing unnecessary hospital stays and associated expenses. This includes minimizing expenditures on medications, treatments, and supportive services during prolonged hospitalization periods, ultimately leading to improved financial outcomes for the hospital.

Post-discharge care coordination and likelihood of hospital re-admission is another critical aspect influenced by accurate anticipation of discharge status or mortality [302–304]. Healthcare providers must plan and coordinate post-discharge care services to ensure that patients receive the necessary support and follow-up care after leaving the hospital. By accurately predicting discharge status or mortality, hospitals can initiate post-discharge care planning in advance, arrange follow-up appointments, coordinate home health services, and provide patients with the resources they need to recover and manage their health effectively, ultimately reducing the risk of re-admission and improving patient outcomes.

Overall, accurate anticipation of discharge status or mortality contributes to improved patient flow efficiency within hospitals. By facilitating timely discharges or identifying patients at high risk of mortality, hospitals can optimize patient flow, reduce bottlenecks, and ensure that patients receive timely and appropriate care throughout their hospital stay. By leveraging predictive analytics and advanced decision support tools, hospitals can better anticipate and respond to the needs of their patients, ultimately enhancing operational efficiency, and improving the quality of care and the overall patient experience.

7.3.4 Methods

Data

For the task for patient discharge status prediction, we train models to predict the discharge status of patients staying in the ICU, using data from the eICU Collaborative Research Database (eICU-CRD) [305, 306]. The eICU Collaborative Research Database (eICU-CRD) is a publicly-available, anonymized database with pre-existing institutional review board (IRB) approval. The database is released under the Health Insurance Portability and Accountability Act (HIPAA) safe harbor provision. The re-identification risk was certified as meeting safe harbor standards by Privacert (Cambridge, MA) (HIPAA Certification no. 1031219-2).

The eICU Collaborative Research Database stands as a comprehensive and expansive resource within the realm of critical care, encompassing deidentified health data stemming from over 200,000 admissions to ICUs throughout the United States during the period of 2014 to 2015. This multi-center database provides insights into the management and treatment of critically ill patients across a diverse array of healthcare facilities.

At its core, the database comprises an assortment of clinical data, ranging from vital sign measurements, laboratory measurements, medications, APACHE components, patient history, care plan documentation, severity of illness measures, diagnosis information, and treatment details.

By leveraging the eICU Collaborative Research Database, researchers and healthcare professionals gain access to a wealth of anonymized patient data, enabling them to conduct in-depth analyses, develop predictive models, and uncover valuable insights into critical care practices and outcomes. Moreover, the multi-center nature of the database enhances its generalizability and applicability, offering a representative snapshot of ICU practices and patient populations across diverse healthcare settings throughout the United States.

Training, Continuous Validation, and Test Dataset Splitting

As discussed, the task is to predict the discharge status of a patient during the course of an ICU stay. Using similar inclusion and exclusion criteria to those used in a previous study [307], we selected adult patients (age > 18) with a minimum of 15 ICU records, and grouped these records into 1 hour windows. We removed any patients that did not have a clear discharge status (i.e., anything that was not "alive" or "expired"), and removed samples with any missing values.

We used a 60:20:20 training, continuous validation, and test ratio, resulting in 49,305 training, 16,436 continuous validation, and 16,436 test samples, respectively. As before, the training set, $\mathcal{D}_{train}$, was used for model development, hyperparameter selection, and training; the continous validation set, $\mathcal{D}_{val}$, is used for continuous validation and threshold adjustment; and after successful development and training, the held-out test set, $\mathcal{D}_{test}$, was used to evaluate the performance of the final model.

Table 7.3: Summary of number of patients, death cases, and hospital case distribution for training, validation, and held-out test set cohorts used in the ethnicity debiasing task.

	Training ($\mathcal{D}_{train}$)	Continuous Validation ($\mathcal{D}_{val}$)	Test ($\mathcal{D}_{test}$)
Total Patients	49305	16436	16436
Total Deaths	4501	1486	1486
Ethnicity:			
Caucasian (%)	40285 (81.7)	13514 (82.2)	13510 (82.2)
African American (%)	5704 (11.6)	1832 (11.1)	1869 (11.4)
Hispanic (%)	2073 (4.2)	665 (4.0)	667 (4.1)
Asian (%)	938 (1.9)	319 (1.9)	282 (1.7)
Native American (%)	305 (0.6)	106 (0.6)	108 (0.7)

Clinical Features Used for Prediction

This dataset contains a wide range of data-types, offering a comprehensive perspective on patients' ICU experiences by encompassing vital aspects such as health status, disease severity, and treatment interventions, we incorporate demographic characteristics, measurements recorded upon hospital admission, and measurements recorded upon ICU admission. The attributes used for training machine learning models are outlined in Table 7.4.

Table 7.4: Clinical predictors considered for predicting patient discharge status and patient diagnosis.

Category	Features
Demographic features	Gender, age, height, weight
Measurements at hospital admission	Non-invasive systolic blood pressure, non-invasive diastolic blood pressure, non-invasive mean arterial pressure, heart rate, supporting oxygen used at admission, blood oxygen saturation, Glasgow coma score, diagnosis at admission
Measurements at ICU admission	Glucose

Data Pre-processing

We employ one-hot encoding for categorical features and standardization for continuous features. One-hot encoding is a technique used to convert categorical variables into a binary representation, where each category is represented by a binary vector with a single 1 denoting the presence of that category and 0s elsewhere. This ensures that categorical variables are appropriately encoded for machine learning algorithms, which typically require numerical inputs.

For continuous features, we standardize all continuous features to have a mean of 0 and a standard deviation of 1. As mentioned in previous chapters, standardization ensures that all continuous features have comparable scales, which can help improve the performance of machine learning algorithms, particularly those sensitive to feature scales, such as linear models and neural networks.

Evaluation Metrics

We adopt the same evaluation metrics outlined in Section 4.2.5, reporting on sensitivity, specificity, PPV, NPV, and AUROC with 95% CIs derived from 1000 bootstrapped samples from the test set. Significance tests, marked by p-values, assess model performance differences over 1000 bootstrapped comparisons, using a significance threshold of 0.05. These analyses are performed on final test sets. For fairness evaluation, we measure Equalized Odds variance as specified in Equations 4.16 and 4.17.

Hyperparameter Optimization

In alignment with Chapter 4.2.6, we employ a grid search coupled with standard five-fold cross-validation on the training set to determine the optimal hyperparameters for all conventional supervised learning and reinforcement learning methods used.

As before, a grid search was applied to all neural network models to ascertain the best configurations, encompassing factors like the number of hidden layers, nodes per layer, and learning rate. Concerning adversarial debiasing, this exploration was carried out independently for both the predictor and adversary networks, as well as for the α hyperparameter. In the case of XGBoost, we explored a range of hyperparameters including learning rate, depth, and the number of trees. For reinforcement learning methods, optimization efforts were focused on parameters including the Q-network's layer count, nodes per layer, and learning rate.

Detailed information regarding the software, implementation, and final hyperparameter values selected for each model can be located in Appendix E.

Threshold Optimization

In alignment with Chapter 4.2.7, we perform threshold optimization for binary classification, to predict patient discharge status. Here, our objective is to find a suitable threshold, optimized to achieve sensitivities of 0.9.

Table 7.5: Equalized Odds evaluation for ethnicity bias and Patient ICU discharge prediction test results across different models, optimized to sensitivities of 0.9. Metrics are reported alongside 95% confidence intervals based on 1,000 bootstrapped samples. Bolded values denote the best (underlined) and second best Equalized Odds scores. Classification metrics are reported alongside 95% CIs, with bolded values denoting the best scores achieved on the test set.

Model	$M_{EO(TP)}$	$M_{EO(FP)}$	Sensitivity	Specificity	PPV	NPV	AUROC
RL (debiasing)	**<u>0.032</u>**	**<u>0.022</u>**	**0.897 (±0.015)**	0.539 (±0.008)	0.171 (±0.008)	0.980 (±0.003)	0.829 (±0.013)
RL	0.052	0.030	0.889 (±0.016)	0.502 (±0.008)	0.159 (±0.008)	0.977 (±0.003)	0.818 (±0.013)
Adversarial	0.040	**0.027**	0.885 (±0.016)	0.637 (±0.008)	0.205 (±0.010)	0.981 (±0.003)	0.861 (±0.012)
Neural Net.	**0.033**	0.037	0.884 (±0.016)	0.600 (±0.008)	0.189 (±0.009)	0.980 (±0.003)	0.847 (±0.012)
XGBoost	0.062	**<u>0.022</u>**	0.883 (±0.016)	**0.674 (±0.008)**	**0.223 (±0.011)**	**0.982 (±0.003)**	**0.875 (±0.012)**

7.3.5 Results

Following the development and training of the models, we assessed their performance on the held-out test set. As before, we found that all models achieved reasonably high AUROC scores on the test set (Table 7.5), comparable to previously reported benchmarks using the same dataset [307]. AUROC scores ranged from 0.818-0.875, with the XGBoost model achieving the highest score and the RL models (both with and without a debiasing component) achieving the lowest. However, when optimizing sensitivities to 0.9, the reinforcement learning-based debiasing method achieved the best results in terms of sensitivity, scoring 0.897 (±0.015), and Equalized Odds, scoring 0.032 and 0.022 for TP and FP rates, respectively. However, this did come at a small trade-off in AUROC. The difference in accuracy of the RL debiasing model compared to that of other models was found to be statistically significant ($p<0.0001$, based on 1,000 bootstrapped samples).

7.3.6 Discussion

As demonstrated in previous chapters, we found that the debiasing method using reinforcement learning demonstrated superior performance, achieving the best Equalized Odds for both true positive and false positive rates. Again, the enhanced effectiveness of the RL debiasing approach in contrast to supervised-learning-based bias mitigation methods can be credited to the RL setup's capability to control the propagation of learning signals, thereby diminishing the model's susceptibility to being biased toward the majority class present in each batch during training.

Nevertheless, despite attempts to alleviate bias, these models fell short of fully meeting the criteria for Equalized Odds. This enduring deficiency could be ascribed to the data imbalance related to the sensitive attribute [102], which was similarly evident in the previous case study, with a significantly larger dataset available from white patients compared to other ethnicities. Consequently, there tends to be a bias toward majority subgroups and this was especially noticeable when employing neural networks.

With respect to fulfilling Equalized Odds requirements, the advantage of using an RL framework was more observable and clear (i.e. noticeable improvements in TP and FP SDs for RL results over other models) for the patient discharge task compared to the previous COVID-19 task in Chapter 7.2 which also focused on mitigating ethnicity bias. Again, this may be due to the larger amount of training data used in this task compared to the previous one (14,949 patients compared to 49,305 patients for COVID-19 ethnicity and ICU patient discharge tasks, respectively). Thus, having a greater amount of training data may have made it easier for models to confidently differentiate between different classes.

It should be noted that mortality prediction may not always be an ideal prediction task in machine learning for several reasons, one of which involves the complexities surrounding the determination of patient discharge status. Patient discharge decisions are not solely based on clinical factors captured in datasets. Factors such as patient preferences, social support networks, financial considerations, and legal constraints can significantly influence discharge decisions. For instance, some patients may choose to self-discharge against medical advice, while others may have external circumstances affecting their discharge process [308–310]. These non-clinical factors are often not captured in clinical datasets used for machine learning tasks, leading to incomplete or inaccurate predictions of patient outcomes like mortality. Consequently, mortality prediction models may not fully account for the complexities of discharge decisions and may yield less reliable results compared to other prediction tasks that are more closely aligned with clinical variables. In essence, while mortality prediction models can provide valuable insights into patient prognosis, their utility

may be limited by the lack of comprehensive data on discharge determinants and non-clinical factors. Therefore, for more accurate predictions of patient outcomes, it may be beneficial for future research to focus on prediction tasks that incorporate a broader range of variables and considerations beyond clinical parameters alone.

Additionally, in both case studies, white patients form the predominant subgroup, a trend that's understandable given the datasets' curation in countries like the UK and US. However, this also underscores the potential for collaborative datasets and trained models to become disproportionately skewed toward one subgroup, thereby exacerbating existing imbalances. Hence, efforts beyond mere bias mitigation are imperative to bolster the diversity and representativeness of digital health data overall. This imperative extends not only to patients but also highlights the underrepresentation of racial and ethnic minorities among healthcare professionals, underscoring the urgency of addressing this disparity to promote equity in healthcare.

Finally, with respect to both case studies presented in this chapter, a limitation lies in the complexity of defining what constitutes an unwanted bias. While ethnicity should not influence outcomes in non-clinical scenarios like recidivism prediction, its role in clinical contexts is more nuanced. For example, ethnicity can be a crucial predictor for certain diagnoses, prognoses, and treatment recommendations. In our study, we addressed data imbalances to ensure equitable predictions for minority groups. However, ethnicity can also encompass factors such as place of residence and socioeconomic status, which can affect disease prevalence among specific ethnic cohorts. Thus, as more data becomes available, iterative adjustments are necessary to accurately assess bias and the true influence of these attributes.

8 Conclusion

Biased and unequally represented data can have detrimental effects on the performance, fairness, and trustworthiness of machine learning models, particularly in healthcare and other domains where equitable outcomes are crucial. Addressing these biases requires careful attention to data collection practices, model development methodologies, and evaluation metrics to ensure that machine learning models are fair, accurate, and inclusive across diverse populations.

While discussions on algorithmic fairness have been extensive, there is still no consensus on a universally applicable method, metric, or criterion. Consequently, determining and quantifying the significance of existing biases, as well as assessing the effectiveness of mitigation strategies, can be challenging. Moreover, the complexity is compounded by the diverse range of applications where fairness issues are pertinent. Therefore, it is essential to select and tailor methods according to the specific characteristics and requirements of each task.

Furthermore, as more data becomes available and as algorithms and empirical studies continue to evolve, alternative methods will emerge to further advance the field of algorithmic fairness and bias mitigation in machine learning. This ongoing research and development process will contribute to the refinement and enhancement of approaches to address fairness concerns across various domains and applications. By prioritizing fairness and accountability in machine learning

development, stakeholders can work together to build machine learning systems that uphold principles of fairness, transparency, and social responsibility.

8.1 Summary of Major Findings

In evaluating fairness-aware algorithms, our main focus was on investigating location-based biases across separate hospital sites. This approach enabled us to use existing insights into the diversity among healthcare facilities, showcasing how such disparities are manifested in data and consequently impact model training and effectiveness.

Our primary emphasis was on an extensive case study involving COVID-19 diagnosis across four distinct UK hospital trusts. This allowed us to underscore site-specific variations in population characteristics and feature distributions. Furthermore, employing a data-driven technique like t-SNE enabled us to visualize sample similarities based solely on inherent data patterns and structures. Consequently, the main insights from this analysis highlight the multifaceted nature of bias between hospitals, and underscore the importance of understanding unintended biases among hospital sites for both accurate interpretation of and the application of various downstream analyses or modeling methods.

We then highlighted the principles involved in the development, utilization, and assessment of fairness-aware algorithms. We started by showcasing various supervised learning techniques designed to alleviate unintentional biases inherent in training data, with the purpose of improving fairness in machine learning outcomes. Our study revealed that the fairness-aware algorithms used, such as the NCR regularization approach and the adversarial framework, generally led to enhanced fairness outcomes compared to models lacking built-in bias mitigation techniques, all while preserving high predictive accuracy.

Moreover, recognizing that the predominant focus of existing literature has been on addressing algorithm-level bias through traditional supervised learning methods, we aimed to pioneer a novel bias mitigation strategy within the realm of reinforcement learning. While supervised learning models typically demand large labeled datasets

tailored to specific tasks, RL algorithms offer versatility in addressing a broad spectrum of problems. This adaptability makes RL particularly adept at handling real-world challenges characterized by noisy, incomplete, or heterogeneous data, as commonly encountered in clinical settings. By expanding the repertoire of bias mitigation techniques through the introduction of a reinforcement learning-based approach prioritizing fairness outcomes, we presented a unique contribution to bias reduction within an alternative machine learning paradigm. Our findings indicated that the RL-based debiasing approach demonstrated superior fairness performance compared to supervised-learning-based bias mitigation techniques. This highlighted the importance of adopting fairness-aware methods across a wider range of tasks, including those that cannot be effectively addressed by supervised learning.

Finally, our proposed methods were put into practice across novel applications to conduct a comprehensive assessment of their effectiveness and applicability. Our exploration extended to encompass the utilization of fairness-aware algorithms within UK hospital sites as well as two hospitals in Vietnam, showcasing the potential of these algorithms to foster fairness across varying socioeconomic contexts. Our primary focus was on facilitating impactful collaborative AI advancement that benefits both high-income country and low-middle income country hospitals, even in instances where LMIC data is underrepresented in the training process. We illustrated that the integration of bias mitigation techniques not only improved algorithmic fairness but also enhanced the model's adaptability when deployed in LMIC settings. Consequently, the adoption of fairness-aware algorithms plays a crucial role in fostering trust among clinicians and patients in the reliability and efficacy of machine learning-based technologies. This, in turn, fosters and strengthens international cooperation and AI development initiatives.

Furthermore, we evaluated the effectiveness of fairness-aware techniques through two additional case studies: one addressing ethnicity bias in COVID-19 screening and another targeting ethnicity bias in patient discharge status prediction. Through these studies, we showcased the versatility and effectiveness of these methods in

promoting fairness across diverse applications, and further underscore the superior performance of our innovative reinforcement learning-based debiasing approach.

In summary, this thesis addressed the pressing issue of biases originating from data in machine learning, particularly within the domain of clinical machine learning. It not only highlighted the significance of recognizing and mitigating these biases but also introduced and demonstrated effective methods to promote fairness in this rapidly evolving field. Through a thorough exploration of various bias mitigation techniques and their application in clinical settings, this thesis contributes to the advancement of fair and equitable machine learning practices, ultimately aiming to improve the reliability and effectiveness of models used in healthcare decision-making.

8.2 Limitations

In our approach to COVID-19 prediction, we made adjustments to the decision threshold to prioritize achieving high sensitivity in our models. This strategy is particularly valuable when dealing with datasets characterized by significant imbalances, as was the case with our training sets. By focusing on sensitivity, we aimed to maximize the model's ability to correctly identify positive cases, which is crucial in a context like COVID-19 where early detection is paramount for effective intervention and containment efforts. However, as previously highlighted, the data used for model training can be influenced by site-specific factors, as indicated by the t-SNE representation depicted in Figure 3.2. Consequently, the optimal decision thresholds derived from one dataset may not necessarily generalize well to others. This variability in distributions across different sites can lead to disparities in model performance, particularly in terms of sensitivity/specificity. Overall, consistency in sensitivity and specificity scores across diverse hospital settings is crucial for ensuring that clinicians can confidently rely on the predictive capabilities of the model. Achieving this consistency is especially pertinent in clinical contexts where uniform performance characteristics are essential for informed decision-making and patient care. Additionally, the selection of an optimal decision threshold also directly

impacts statistical fairness, as the threshold shift controls the true positive and true negative rates. Thus, moving forward, future experiments should explore the possibility of implementing site-specific thresholds calibrated during deployment at different healthcare facilities. This approach would aim to standardize predictive performance across various sites, thereby promoting consistency and reliability in model predictions. By tailoring thresholds to specific contexts, we can enhance the applicability and effectiveness of predictive models in real-world clinical settings, ultimately improving patient care and outcomes.

Similarly, the balance between sensitivity and specificity holds significant importance and warrants careful consideration, contingent upon the specific objectives of the task at hand. In the COVID-19 task, we optimized thresholds for high sensitivity to aid in triaging. However, this trade-off negatively affected specificity, as evidenced by RL debiasing exhibiting high sensitivities but lower specificities. It is crucial to acknowledge the importance of specificity optimization as well, as a low specificity can lead to heightened resource utilization and costs, potentially burdening hospitals and adversely impacting patient well-being by inducing increased anxiety or discomfort.

This consideration of sensitivity and specificity in model evaluation also extends to the choice of fairness criteria used. Models should be customized to align with fairness definitions that best serve the objectives of each task. In scenarios where minimizing harm from false negatives is paramount, such as disease diagnosis, prioritizing high sensitivity may be preferable. For such cases, fairness metrics like equal opportunity ensure that the classifier's probability of predicting a sample as the positive class is uniform across all classes of the sensitive attribute [78, 80]. Similarly, when ensuring equality in predicting the negative class across sensitive attribute classes is more crucial, predictive equality can be employed [204, 207]. Furthermore, imbalances in real-world datasets concerning both outcome and sensitive feature labels must be considered. Therefore, selecting and devising evaluation metrics capable of adapting to and accurately representing these imbalances is essential. Such metrics would enable a more comprehensive assessment of model performance

and fairness across diverse contexts and scenarios, thereby enhancing the robustness and practical utility of machine learning solutions in real-world settings.

While the models achieved a high NPV, this metric should be interpreted with caution in the context of a class-imbalanced problem. A high NPV is expected a priori given the low prevalence of COVID-19 in the cohorts, and therefore does not, on its own, define clinical success. Moreover, without prospective outcome studies or evidence of impact on patient care, it remains unclear what threshold of NPV would be sufficient to claim clinical utility. For the purposes of this thesis, however, the dataset and application are not intended to serve as a definitive clinical deployment study. Instead, they provide a valuable case study and substrate for addressing the central methodological question: can fairness-aware algorithms mitigate bias while maintaining accuracy in clinical machine learning?

We also recognize the significance of considering the probability of disease occurrence as a valuable metric, contrasting with the approach of solely thresholding to a binary classification. In our investigations, we opted for binary classification over probability estimation to align with the categorization system mandated by NHS Trust policy. However, it is important to highlight that while binary classification was our chosen method to adhere to policy guidelines, probability estimation remains a viable option, particularly in scenarios where assessing disease severity is a critical aspect of the task. Therefore, although we implemented binary classification, the utilization of probability as a final output is also feasible and valuable, especially in contexts where a nuanced understanding of disease likelihood and severity is necessary.

Another significant limitation arises from the complexity of individual patients, and the collective influences of social, behavioral, and genetic factors on outcomes. While genetics play a role in determining certain outcomes, such as susceptibility to diseases, the interplay between genetic predispositions and environmental factors can significantly alter the outcome [311]. Additionally, factors like population admixture, which refers to the genetic mixing of different populations over time, further complicate the understanding of how genetic variations contribute to outcomes

across diverse populations. This complexity highlights the challenge researchers face in disentangling the intricate web of interactions between social, behavioral, and genetic factors and their collective impact on health outcomes [53, 312]. It underscores the necessity for comprehensive and interdisciplinary approaches to better understand and address these multifaceted influences on health and well-being. Thus, additional investigation into the main prediction task (and related variables) will be necessary to determine what biases exist and how to best mitigate them. For example, as highlighted in Chapter 7, while it is clear that ethnicity should not dictate outcomes in certain non-clinical scenarios like recidivism prediction, its significance in clinical contexts is not always straightforward [313]. Ethnicity can serve as a vital predictor for particular diagnoses, prognoses, and treatment recommendations. In our COVID-19 screening endeavor, our focus centered on rectifying data imbalances to ensure equitable predictions for minority groups, drawing from available data across UK hospital trusts. However, we recognize that ethnicity encompasses crucial facets such as place of residence and socioeconomic status, collectively influencing disease prevalence among specific ethnic cohorts. Pinpointing the precise impact of ethnicity (and related factors) on COVID-19 diagnosis during the initial stages of the pandemic presents challenges. Nevertheless, with the accumulation of more data over time, iterative adjustments are imperative to accurately gauge the true influence of these attributes.

Another critical consideration lies in striking a balance between fairness and predictive accuracy. Our investigation revealed instances where enhancements in fairness came at the expense of overall performance accuracy. Studies have previously shown that while efforts to improve fairness are commendable, they often entail a trade-off with performance metrics [108, 314–316]. Thus, it becomes imperative to carefully weigh whether certain fairness enhancements justify the sacrifice in achieving highly accurate predictions for specific subgroups.

This challenge also underscores the importance of evaluating the suitability of demographic-specific or site-specific models. While our models have demonstrated efficacy in handling multi-class sensitive features, it is essential to assess whether a

demographic-specific or site-specific model might be more suitable than a generalized multi-class model for a given task. For instance, if the model aims to provide support within a specific hospital care structure or predict the risk of a disease known to manifest significant variability between ethnicities, personalized models trained individually on each class may offer the optimal solution. However, it should be noted that the adoption of multiple models can pose computational challenges, presenting practical hurdles for hospitals to overcome efficiently. Therefore, careful consideration is warranted when deciding between the implementation of specialized models tailored to specific subgroups and the utilization of more generalized models to address fairness concerns without compromising overall performance accuracy. In scenarios where computational resources are limited or where the complexity of managing multiple models is prohibitive, adopting a more generalized model, such as the debiasing frameworks demonstrated in our study, can offer a viable solution. Despite its broader scope, a generalized model can still effectively address biases while providing a feasible alternative to individualized approaches. By using such frameworks, healthcare institutions can navigate the trade-offs between computational efficiency and model effectiveness, ensuring that biases are mitigated without imposing undue burdens on resource-strapped systems.

We acknowledge that the COVID-19 and eICU datasets used in our study only represent a fraction of the extensive data available within hospital record systems. For example, with respect to electronic health record data, crucial elements including intricate treatment records, lifestyle variables, and environmental factors, among others, are not entirely represented in the datasets used in this research. Consequently, it is imperative to pursue additional research endeavors aimed at attaining a comprehensive understanding of the implications of various types of bias and assessing how bias mitigation techniques affect model performance. Similarly, we also acknowledge that our examination of variations across specific hospitals and ethnicities only addresses a portion of healthcare disparities. However, from our investigations, we hope to promote the utilization of fairness-aware algorithms

in a wider range of prediction and debiasing tasks through the framework and concepts introduced in this thesis.

Another constraint observed in our study pertains to the inherent characteristics of clinical data, which often feature noise, inconsistency, and missing values. Clinical data is prone to noise, which can obscure meaningful patterns and insights. This noise originates from diverse sources, including measurement inaccuracies, errors in data entry, or variations in recording practices across different healthcare providers. Consequently, machine learning algorithms trained on noisy clinical data may yield unreliable predictions, thereby compromising the precision of healthcare decision-making processes. Additionally, inconsistency and missing values also represent common challenges in clinical data, affecting both data quality and representation. These factors collectively exert a substantial impact on our capacity to discern existing biases within the data and subsequently develop machine learning algorithms that are both equitable and precise. One method of overcoming this can include using methods similar to the NCR method introduced in Chapter 4, as this method relies on using feature representations, rather than labels (which can be noisy) to train models.

Moreover, for the COVID-19 screening task, our investigation spanned a significant time period, from December 1, 2019, to December 30, 2022. During this extended duration and particularly during peak pandemic periods, such as the COVID-19 outbreak, the relationship between patient and disease factors with clinical events, including hospital-acquired infections, may undergo changes [272]. Additionally, over time, there may be variations in practice patterns such as hardware and software updates and changes in protocols, which can impact data capture as well as outcomes.

Finally, another constraint lies in our reliance on statistical fairness metrics to evaluate fairness. Hence, a significant concern arises regarding the potential inconsistency between a patient's clinical trajectory and the fair predictions generated by a model [53]. Given the inherent heterogeneity of clinical data and human behavior, along with the myriad of other factors influencing patient outcomes, what

occurs if the patient's response deviates from the predicted one (compared to the reference training set on which the model is developed)? This variance between the idealized model and real-world behavior can profoundly affect metrics of model performance, such as specificity and sensitivity, as well as its practical clinical utility. If clinicians and patients perceive the model as impartial, any disparities between its predictions and the patient's actual clinical condition could be challenging to decipher. Consequently, this could obscure interventions that could have been more relevant and beneficial for the patient [53]. Hence, relying solely on output metrics to operationalize fairness proves inadequate.

8.3 Future Research Directions

8.3.1 Multi-Modal Analyses

An exciting area for future research is the implementation of multi-modal analyses. As diverse data types, such as imaging and genetic sequencing, continue to proliferate, there exists a promising opportunity to harness multi-modal analyses and algorithms to delve deeper into the complexities of human biology and health [317–319]. Through the integration of multiple data modalities, researchers gain access to a broader spectrum of information for model training, thereby enhancing the robustness of classifiers. For example, incorporating data from MRI scans, X-rays, and other imaging technologies can provide spatial and structural insights into biological tissues and organs. Using genetic information, such as DNA and RNA sequencing, can offer detailed views of the genetic underpinnings of diseases and biological processes. Including patient records, lab test results, and other clinical information adds another layer of context that is critical for understanding health outcomes.

Even prior to model development, these heterogeneous datasets can be subjected to thorough analysis to identify viable tasks for modeling and ascertain which features are most relevant for constructing accurate and resilient models. Such preliminary analyses serve as a foundational step in understanding the landscape of available data and formulating effective modeling strategies. For instance, feature

selection can help identify which variables from imaging and genetic data are most predictive of certain health outcomes. Additionally, correlation studies can analyze how different types of data correlate with each other to uncover hidden relationships and potential causal links.

Moreover, these multi-modal analyses are instrumental in uncovering existing biases within the data. By scrutinizing diverse sources of information, researchers can identify patterns of bias and devise strategies for their mitigation. For example, exploratory data analysis can reveal if there are disparities that exist across different patient populations. By addressing these issues early on, researchers can ensure that the data used for training models is more representative and balanced. Moreover, preliminary analyses can help in feature selection and engineering, ensuring that the most relevant and unbiased features are used in the model. For instance, if a feature like "socioeconomic status" introduces bias in health predictions, researchers might adjust its representation or find alternative features that capture the necessary information without the associated bias. This proactive approach ensures that biases are addressed early in the modeling process, leading to more equitable and reliable outcomes.

Case studies and applications of multi-modal analyses further highlight their potential. In cancer research, combining imaging, genetic, and clinical data has been used to improve the accuracy of cancer diagnosis and treatment personalization [317–322]. In the field of neurological disorders, integrating MRI scans with genetic data has helped in understanding complex conditions like Alzheimer's disease [323–325]. Similarly, in cardiovascular health, combining electrocardiogram data, genetic markers, and clinical records has advanced the prediction and management of cardiovascular diseases [326–330].

The integration of multi-modal data in developing fairness-aware algorithms holds significant potential for advancing personalized and equitable healthcare. By comprehensively understanding patient data, healthcare providers can deliver more customized treatments that account for each patient's unique genetic, clinical, and lifestyle factors. For example, personalized treatment plans for cancer patients can

be created by combining genomic data with clinical trial outcomes and imaging studies, resulting in more effective and targeted therapies. Furthermore, fairness-aware algorithms can ensure that these personalized treatments are distributed equitably across all patient populations. For example, in the development of predictive models for disease outbreaks and diagnosis, as highlighted in this thesis, these algorithms can guarantee that predictions are both accurate and fair across various geographic and demographic groups, leading to more effective and inclusive public health interventions.

Finally, an intriguing and valuable direction for future research lies in the development of more advanced bias mitigation methods capable of addressing and reducing bias across various data modalities. This approach would involve tailoring unique mitigation strategies to suit each specific type of data, whether it be imaging, genetic, clinical, or demographic. For instance, specialized techniques could be developed to correct biases inherent in medical imaging data, while different strategies might be applied to address biases found in genetic or clinical datasets. By customizing bias mitigation approaches to the characteristics of each data modality, researchers can more effectively ensure fairness and accuracy in multi-modal machine learning models, leading to more equitable and reliable outcomes in fields such as healthcare, finance, and beyond.

8.3.2 Explainable Methods

In addition to developing sophisticated algorithms capable of handling complex and high-dimensional data, future research in machine learning bias detection will also focus on incorporating techniques from explainable AI (XAI) to identify and understand subtle biases. By improving the transparency and explainability of AI systems, researchers can enhance their ability to detect and mitigate biases, ultimately fostering greater trust and ethical use of AI across various domains [331].

Explainable AI provides tools and methods to make the decision-making processes of machine learning models more transparent and understandable [331–335].

One approach to integrating XAI into fairness-aware algorithms is the use of model-agnostic explainability tools [331, 333]. These tools, such as Local Interpretable Model-agnostic Explanations [336], Shapley Additive Explanations (SHAP) [337], and others [338–340], can be applied to any machine learning model to explain its predictions. For example, SHAP values can quantify the contribution of each feature to a model's output [337], helping to identify if certain features are contributing to biased outcomes. By using these tools, researchers can pinpoint specific aspects of the data or model that may be introducing bias and adjust them accordingly.

Another important aspect is the development of inherently interpretable models. Unlike complex "black-box" models, interpretable models are designed to be easily understood by humans. Examples include decision trees [331, 341, 342], rule-based systems[331, 333, 334, 343], and Generalized Additive Models [333, 344–346]. These models provide clear insights into how decisions are made, which is crucial for identifying and correcting biases. For instance, a decision tree used in predicting patient treatment outcomes can be examined to ensure that it does not unfairly discriminate against patients based on race or gender. By using interpretable models, researchers can more easily ensure fairness in their algorithms.

Incorporating XAI techniques into fairness-aware algorithms not only aids in bias detection but also enhances the overall transparency and accountability of AI systems. Regularization techniques can be designed to promote fairness by encouraging the model to rely on features in a more balanced manner. For example, using explainability-driven regularization can use insights from XAI tools to guide the regularization process [347], ensuring that the model does not disproportionately depend on any biased features. For example, in a predictive policing model, regularization can be guided by SHAP values to ensure that the model's reliance on socio-economic data does not lead to biased predictions. By continuously monitoring feature importance during training, researchers can adjust regularization parameters to achieve a fairer model.

Moreover, the use of XAI fosters trust among users and stakeholders [331–335]. When people can understand and trust the decision-making processes of

AI systems, they are more likely to accept and adopt these technologies. This is particularly important in high-stakes domains like criminal justice, where AI-based risk assessments and sentencing recommendations must be transparent and justifiable to ensure public trust and ethical use. In healthcare, where AI directly impacts critical health outcomes and life-or-death decisions, XAI can help doctors understand why a model predicts a high risk of a certain disease for a patient. This allows them to validate the model's reasoning and ensure it is based on relevant medical factors rather than biased data. In finance, explainable models can provide clear justifications for credit decisions, helping to ensure that lending practices are fair and non-discriminatory.

In conclusion, incorporating XAI techniques into fairness-aware algorithms is essential for the future of machine learning bias detection. By utilizing model-agnostic tools, developing interpretable models, and directly applying these methods in bias mitigation, researchers can significantly enhance the transparency and explainability of AI systems. These advancements will improve our ability to detect and mitigate biases, leading to more equitable and trustworthy AI applications across various domains.

8.3.3 Foundation Models

Foundation models, with their large-scale and general-purpose capabilities, have transformed the landscape of artificial intelligence, enabling breakthroughs across a wide range of applications. They are now considered the basis for a wide range of downstream tasks across various data modalities [348, 349]. However, their immense scale and reliance on vast datasets also introduce unique challenges for bias mitigation and algorithmic fairness. These models often reflect the biases present in their training data, which can propagate and amplify in downstream tasks [350–354]. At the same time, their flexibility and adaptability provide opportunities to address fairness more holistically, offering new tools and methods to identify, mitigate, and monitor bias. Exploring the intersection of foundation models and fairness is critical to harnessing their potential responsibly and equitably.

Both large language models (e.g., GPT-4 [355]) and vision models (e.g., CLIP [356]), are pre-trained on massive datasets often sourced from the internet [348, 351, 352, 354, 357]. These datasets inherently reflect societal biases, which can be amplified and propagated into downstream tasks, potentially leading to unfair or harmful outcomes [352, 353, 357]. Addressing this issue requires identifying and quantifying the biases embedded within foundation models, as well as developing fairness metrics specifically tailored to assess their behavior across large-scale, multi-domain contexts [348, 354]. Additionally, understanding how biases present in pre-training datasets translate into specific downstream applications is crucial for designing effective mitigation strategies. This area of research holds significant promise for improving fairness and accountability in AI systems powered by foundation models [354]. For example, when fine-tuning foundation models for specific tasks, it will be essential to incorporate effective bias mitigation strategies to ensure fair and equitable outcomes [357, 358]. This could involve exploring techniques that preserve the valuable features and knowledge embedded in foundation models while simultaneously reducing inherited biases. Moreover, debiasing techniques inspired by adversarial training, reweighting, or targeted loss functions (as demonstrated in this thesis) can be implemented during the fine-tuning process to address and mitigate unfair tendencies. It is also important to investigate how task-specific adaptations interact with pre-existing biases in the model, as these interactions may introduce or exacerbate unintended disparities. Tackling these challenges during fine-tuning not only enhances fairness but also ensures that foundation models align more closely with ethical and societal expectations in real-world applications. Additionally, given the immense size of foundation models, it is far more computationally efficient to focus on debiasing for specific fine-tuned tasks rather than attempting to address biases throughout the entire pre-training process.

Being general-purpose, foundation models often combine multiple modalities—such as text, images, and videos—to support complex, cross-domain tasks [351, 352]. Ensuring fairness in these multi-modal settings is particularly challenging due to the interactions between modalities, which can introduce or amplify biases

[358]. Addressing these challenges requires the development of cross-modal fairness metrics and mitigation strategies that account for the interplay between different types of data [359]. Additionally, it is important to understand how biases in one modality, such as textual descriptions, may influence predictions in another, such as image classification [359, 360]. Thus, future studies should prioritize developing alignment techniques to ensure fairness across diverse input modalities, which is essential for achieving equitable outcomes and preserving the integrity of multi-modal foundation models in real-world applications.

Foundation models are also rapidly being adopted in critical domains such as healthcare, education, and legal systems, where fairness is essential for achieving equitable outcomes [351, 353]. These fields present unique challenges that necessitate tailored strategies to effectively address biases and disparities. Furthermore, as foundation models are continuously updated or adapted to new applications, it becomes crucial to develop robust frameworks for dynamically auditing and monitoring fairness [351, 361]. Such frameworks can help detect emerging biases and ensure that these powerful models remain aligned with ethical standards and societal expectations over time [362, 363].

Another intriguing direction for future research lies in using foundation models as tools for detecting and addressing biases in other systems. Their advanced generalization capabilities make them well-suited for identifying bias patterns within datasets or smaller models, revealing disparities that might otherwise remain hidden. Furthermore, foundation models can serve as benchmarks for evaluating the fairness of other algorithms, offering a standard for assessing and improving equity in machine learning systems. By using these capabilities, foundation models have the potential to play a pivotal role in fostering more transparent and equitable AI ecosystems.

8.3.4 Ethical AI Frameworks and Standards

The establishment of ethical AI frameworks and industry standards for bias mitigation is a crucial direction for future research and practice. These frameworks can provide comprehensive guidelines and best practices for developing and deploying fair

and unbiased AI systems, ensuring that the ethical implications of AI are adequately addressed throughout the model lifecycle [364–367]. By setting clear standards, these frameworks help ensure that AI technologies are developed responsibly, reducing the risk of harm and discrimination.

Regulatory bodies and professional organizations play a pivotal role in setting these standards and ensuring compliance. For example, the European Union's General Data Protection Regulation includes provisions that impact the development and deployment of AI, particularly regarding transparency and the right to explanation [368]. Similarly, the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems has developed guidelines to foster ethical AI practices [366, 367]. Such professional organizations can develop and promote ethical guidelines and standards to standardize bias mitigation practices across the field.

Several ethical AI frameworks have also already been proposed. For example, the European Commission issued the Ethical Guidelines for Trustworthy AI, outlining several key requirements for trustworthy AI, including diversity, non-discrimination, and fairness [369]. These guidelines serve as a comprehensive framework for developing and deploying AI systems that are ethical and fair.

Additionally, these organizations can also provide training and resources to help practitioners understand and implement these standards. For example, IBM's AI Fairness 360 is an open-source toolkit that includes metrics for datasets and machine learning models to test for biases, as well as algorithms to mitigate these biases, providing practical tools for ensuring that AI systems are fair and transparent [370].

Furthermore, as these standards and frameworks are implemented, they can be directly integrated into AI methods during the model training phase through techniques such as regularization and constraints. Constraint-based methods can enforce fairness criteria during the optimization process, while regularization techniques can incorporate fairness-aware terms in the model's objective function. This ensures that the model optimizes for both accuracy and fairness, as defined by various standards and frameworks. For example, demographic parity constraints can be applied to ensure that the model's predictions are equally accurate across

different demographic groups. This might involve setting constraints so that the rate of accurate disease diagnosis is consistent for patients of all racial and ethnic backgrounds. By doing so, the model helps prevent disparities in healthcare outcomes and ensures that all patients receive equally reliable diagnoses, regardless of their demographic characteristics. Additionally, these standards and frameworks will guide the evaluation of AI methods and their outcomes during *post-hoc* analyses, ensuring continuous adherence to ethical principles.

Overall, establishing ethical AI frameworks and industry standards for bias mitigation is crucial for the responsible development and deployment of AI systems. Regulatory bodies and professional organizations are vital in setting these standards and ensuring compliance. By integrating these standards into AI methods through techniques such as regularization and constraints, practitioners can develop models that are not only accurate, fair, and transparent but also aligned with ethical principles and guidelines. This alignment enhances the acceptance and trustworthiness of AI systems. These efforts will help build public trust in AI and ensure that its benefits are equitably distributed across all segments of society.

8.3.5 Continuous Prospective Model Evaluation

While the retrospective study presented in this thesis provides valuable insights into historical data, future research should prioritize prospective analysis. This approach allows for a more dynamic evaluation of model performance and facilitates timely feedback, which is essential for refining and enhancing predictive models. Factors such as variations in prevalence, data drift, and evolving practice patterns—including hardware and software updates and protocol changes—can influence data capture and outcomes over time. Therefore, incorporating prospective analysis in forthcoming studies is crucial for bolstering the robustness and applicability of the findings.

Moreover, external validation studies conducted in a specific geographical area, during a particular time frame, and within a distinct patient population offer only a limited view and cannot assert universal applicability. Thus, future validation efforts should aim to quantify and comprehend the heterogeneity in model

performance, rather than solely focusing on point estimates [271]. By dynamically assessing model performance and incorporating timely feedback, researchers can refine predictive models, ensuring their relevance and effectiveness in diverse and evolving clinical environments.

In conclusion, while fairness metrics are essential for evaluating model performance, they should be complemented by an understanding of how model predictions translate into tangible outcomes for patients in clinical practice at various points in time. This approach ensures that machine learning algorithms not only generate equitable predictions but also make a meaningful contribution to improving healthcare outcomes for all individuals.

8.4 Looking Forward: Ensuring Patient Equity in the Era of AI

I hope this thesis has highlighted the significance and intricacies involved in crafting and implementing fair and equitable AI systems in healthcare. Moving forward, there are still many ethical questions to consider, such as what constitutes fairness and who decides what's fair. There are fundamental biological questions to consider, such as how genetic variability truly affects different populations and the interplay between biology and social determinants of health. There are also technical considerations to address, including the computational resources necessary for training and executing models on increasingly large, high-dimensional, and complex datasets. Crucially, there's the continuous cycle of hypothesis generation, discovery, and the investigation of practical clinical outcomes, underscoring the dynamic nature of this field.

Overall, it is crucial to recognize that fostering fairness and equity in clinical machine learning goes beyond simply deploying fairness-aware algorithms. It demands a holistic approach that encompasses the entire lifecycle of machine learning development and deployment. This includes the initial design phase, where the selection of variables and the conceptual framework must consider equity, through to the data collection processes, which should aim to minimize biases by ensuring diversity and representation. Critical to this effort is the engagement

with stakeholders, including patients from diverse backgrounds, healthcare professionals, and policymakers, to understand and integrate their perspectives and needs. Furthermore, transparency in algorithm development and decision-making processes, along with continuous monitoring and updating of models to adapt to changing populations and healthcare practices, are essential. Such comprehensive measures ensure that the benefits of clinical machine learning are accessible and equitable, truly advancing healthcare outcomes for all.

I've been fortunate to explore just a few of the myriad of possibilities presented by fairness-aware machine learning. As we move forward into the age of artificial intelligence, I hope that these initiatives continue to broaden and provide valuable insights. I also hope that they are supported by the necessary clinical infrastructure to effectively translate these insights into actionable solutions.

Appendices

A

COVID-19 Clinical Data

A.1 Patient Inclusion and Exclusion Criteria

A.1.1 Oxford University Hospitals NHS Foundation Trust

We included all patients attending acute and emergency care settings at Oxford University Hospitals NHS foundation trust who received routine blood tests on arrival. From the "first-wave" of the COVID-19 pandemic in the UK, we included all presentations between December 1, 2019 and June 30, 2020, and from the "second-wave", the second wave we included all presentations between October 1, 2020, to March 6, 2021. Patients presenting with PCR confirmed SARS-CoV-2 infection formed the COVID-19-positive (cases) cohort. We excluded patients who opted out of electronic health record (EHR) research and those who did not receive laboratory blood tests or were younger than 18 years of age. Clinical features extracted for each presentation included first-performed blood tests, blood gases, vital signs measurements and PCR testing for SARS-CoV-2 (Abbott Architect [Abbott, Maidenhead, UK], TaqPath [Thermo Fisher Scientific, Massachusetts, USA] and Public Health England-designed RNA-dependent RNA polymerase assays).

Due to incomplete penetrance of testing during the first wave of the pandemic, and imperfect sensitivity of the PCR test, there is uncertainty in the viral status of patients presenting during the pandemic who were untested or tested negative.

We therefore selected a pre-pandemic control cohort during training to ensure absence of disease in patients labelled as COVID-19-negative. Thus, we additionally considered presentations before December 1, 2019, and thus before the pandemic, as the COVID-19-negative (control) cohort.

A.1.2 Portsmouth Hospitals University NHS Foundation Trust

PUH considered all patients admitted to the Queen Alexandria Hospital, serving a population of 675,000 and offering tertiary referral services to the surrounding region, between March 1, 2020 and February 28, 2021. Confirmatory COVID-19 testing was by laboratory SARS-CoV2 RT-PCR assay, considering any positive PCR result within 48hrs of admission as a true positive.

A.1.3 University Hospitals Birmingham NHS Foundation Trust

UHB considered all patients admitted to The Queen Elizabeth Hospital, Birmingham, between December 01, 2019 and October 29, 2020. The Queen Elizabeth Hospital is a large tertiary referral unit within the UHB group which provides healthcare services for a population of 2.2 million across the West Midlands. Confirmatory COVID-19 testing was performed by laboratory SARS-CoV-2 RT-PCR assay.

A.1.4 Bedfordshire NHS Foundation Trust

BH considered all patients admitted to Bedford Hospital between January 1, 2021 and March 31, 2021. BH provides healthcare services for a population of around 620,000 in Bedfordshire. Confirmatory COVID-19 testing was performed on the day of admission by point-of-care PCR based nucleic acid testing [SAMBA-II & Panther Fusion System, Diagnostics in the Real World, UK, and Hologic, USA].

A.2 Missing Data Imputation

Table A.1: Numbers of participants with data-completeness for each clinical feature, across each validation dataset.

	Prospective Test	External Validation (Admissions)		
	Oxford University Hospitals	Bedfordshire Hospitals NHS Foundation Trust	University Hospitals Birmingham NHS Foundation Trust	Portsmouth Hospitals University NHS Trust
	Oct 1/20 – Mar 6/21	Jan 1/21 - Mar 31/21	Dec 1/19 - Oct 29/20	Mar 1/20 - Feb 28/21
Haemoglobin (g/L)	22532/22857 (98.6%)	10243/10293 (99.5%)	37761/37896 (99.6%)	1177/1177 (100.0%)
White Cells ($10^9 l^{-1}$)	22532/22857 (98.6%)	10244/10293(99.5%)	37756/37896 (99.6%)	1177/1177 (100.0%)
Platelets ($10^9 l^{-1}$)	22511/22857 (98.5%)	10230/10293 (99.4%)	37719/37896 (99.5%)	1172/1177 (99.6%)
Mean Cell Vol. (fl)	22532/22857 (98.6%)	10288/10293 (100.0%)	37750/37896 (99.6%)	1177/1177 (100.0%)
Neutrophils ($10^9 l^{-1}$)	22417/22857 (98.1%)	10277/10293 (99.8%)	37734/37896 (99.6%)	1177/1177 (100.0%)
Haematocrit	22532/22857 (98.6%)	10288/10293 (100.0%)	37755/37896 (99.6%)	1177/1177 (100.0%)
Lymphocytes ($10^9 l^{-1}$)	22430/22857 (98.1%)	10274/10293 (99.8%)	37736/37896 (99.6%)	1177/1177 (100.0%)
Monocytes ($10^9 l^{-1}$)	22452/22857 (98.2%)	10273/10293 (99.8%)	37744/37896 (99.6%)	1177/1177 (100.0%)
Eosinophils ($10^9 l^{-1}$)	22452/22857 (98.2%)	10272/10293 (99.8%)	37736/37896 (99.6%)	1177/1177 (100.0%)
Basophils (109 l-1)	22448/22857 (98.2%)	10270/10293 (99.8%)	37745/37896 (99.6%)	1177/1177 (100.0%)
Sodium (mM)	22442/22857 (98.2%)	9664/10293 (93.9%)	36409/37896 (96.1%)	1173/1177 (99.7%)
Albumin (g/L)	20010/22857 (87.5%)	8783/10293 (85.3%)	35625/37896 (94.0%)	1160/1177 (98.6%)
Alkaline Phosphatase (IU/L)	19885/22857 (87.0%)	8799/10293 (85.5%)	35604/37896 (94.0%)	1111/1177 (94.4%)
ALT (IU/L)	19692/22857 (86.2%)	8689/10293 (84.4%)	35547/37896 (93.8%)	1037/1177 (88.1%)
Urea (mM)	22400/22857 (98.0%)	9667/10293 (93.9%)	36398/37896 (96.0%)	1141/1177 (96.9%)
Bilirubin (umol/L)	19705/22857 (86.2%)	8716/10293 (84.7%)	35550/37896 (93.8%)	940/1177 (79.9%)
Creatinine (umol/L)	22457/22857 (98.2%)	9655/10293 (93.8%)	36415/37896 (96.1%)	1172/1177 (99.6%)
eGFR (ml/min)	22405/22857 (98.0%)	9649/10293 (93.7%)	36415/37896 (96.1%)	1172/1177 (99.6%)
Potassium (mM)	22043/22857 (96.4%)	9306/10293 (90.4%)	34910/37896 (92.1%)	1057/1177 (89.8%)
CRP (mg/L)	19068/22857 (83.4%)	8204/10293 (79.7%)	35245/37896 (93.0%)	1136/1177 (96.5%)
Respiratory Rate (breath/min)	22794/22857 (99.7%)	1177/1177 (100.0%)	10091/10293 (98.0%)	33459/37896 (88.3%)
Heart Rate (beats/min)	22845/22857 (99.9%)	1176/1177 (99.9%)	10117/10293 (98.3%)	33461/37896 (88.3%)
Systolic Blood Pressure (mmHg)	22843/22857 (99.9%)	1171/1177 (99.5%)	10083/10293 (98.0%)	33459/37896 (88.3%)
Diastolic Blood Pressure (mmHg)	22841/22857 (99.9%)	1171/1177 (99.5%)	10082/10293 (98.0%)	33459/37896 (88.3%)
Oxygen Saturation (%)	22837/22857 (99.9%)	1177/1177 (100.0%)	10118/10293 (98.3%)	33459/37896 (88.3%)
Tympanic Temperature (C)	22767/22857 (99.6%)	1177/1177 (100.0%)	10115/10293 (98.3%)	33456/37896 (88.3%)

Table A.2: Evaluation of the performance of (a) CURIAL-Rapide and (b) CURIAL-Lab, optimized during training to achieve a sensitivity of 90%, on prospective set of all admissions to OUH during the second-wave of COVID-19, between October 1, 2020 and March 6, 2021. Mean values are reported alongside SD across population median, population mean, and age-based imputation methods. Values taken from [133].

Model	CURIAL-Rapide	CURIAL-Lab
Sensitivity	85.6 (0.6)	85.7 (0.9)
Specificity	59.1 (0.3)	68.6 (2.2)
PPV	9.46 (0.0)	12.0 (0.6)
NPV	98.8 (0.0)	99.0 (0.0)
AUROC	0.843 (0.002)	0.878 (0.001)

A.3 Distribution of Clinical Features

Table A.3: Distribution of vital sign and blood test features, reported as median and interquartile ranges, for each patient cohort in the training dataset. p-value is obtained from the Kruskal-Wallis test.

Feature	OUH	UHB	PUH	BH	p-value
Respiratory Rate	18 (16.6-19)	17 (16-19)	18 (17-20)	18 (16-20)	<0.0001
Heart Rate	84 (72-97)	82 (71-95)	86 (73-101)	84 (73-98)	<0.0001
Systolic Blood Pressure	134 (119-152)	128 (114-146)	136 (119-155)	131 (116-151)	<0.0001
Diastolic Blood Pressure	75 (65-85)	76 (67-84)	77 (68-87)	78 (68-89)	<0.0001
Temperature Tympanic	36.3 (36-36.8)	36.3 (36-36.8)	36.7 (36.4-37.2)	36.5 (36.4-36.9)	<0.0001
Albumin	35 (31-39)	36 (31-40)	36 (32-40)	35 (31-39)	<0.0001
Creatinine	74 (60-97)	74 (60-96)	78 (62-105)	80 (64-102)	<0.0001
Platelets	250 (197-313)	252 (198-313)	247 (196-311)	249 (199-317)	0.127
Bilirubin	9 (6-14)	10 (7-16)	10 (7-15)	10 (8-14)	<0.0001
Urea	5.7 (4.2-8.3)	5.2 (3.8-7.6)	6.2 (4.5-9)	5.8 (4.2-8.55)	<0.0001
Neutrophils	6.36 (4.34-9.47)	5.9 (4.2-8.5)	6.9 (4.8-10)	6.9 (4.8-9.8)	<0.0001
CRP	16.6 (3.6-69.3)	12 (3-69)	12 (3-61)	11.8 (2.8-46.6)	<0.0001
Lymphocytes	1.29 (0.84-1.87)	1.5 (0.98-2.1)	1.3 (0.9-1.9)	1.28 (0.86-1.8)	<0.0001
Haemoglobin	129 (114-142)	130 (114-143)	127 (113-140)	134 (119-146)	<0.0001
White Cells	8.85 (6.63-11.98)	8.5 (6.6-11.2)	9.4 (7.1-12.6)	9.3 (7-12.6)	<0.0001
Mean Cell Volume	90.2 (86.6-94.2)	89 (84.9-93)	89.9 (86.2-93.6)	88 (84-92)	<0.0001
Haematocrit	0.39 (0.35-0.43)	0.39 (0.345-0.425)	0.38 (0.34-0.42)	0.4 (0.35-0.43)	<0.0001
Basophils	0.04 (0.02-0.06)	0.04 (0.02-0.06)	0.1 (0-0.1)	0.05 (0.03-0.07)	<0.0001
Eosinophils	0.07 (0.02-0.16)	0.1 (0.02-0.2)	0.1 (1-0.2)	0.07 (0.02-0.16)	<0.0001
Sodium	138 (135-140)	138 (136-140)	137 (134-139)	138 (136-140)	<0.0001
ALT	20 (12-33)	19 (13-29)	19 (13-30)	20 (13-31)	<0.0001
Alkaline phosphatase	84 (66-111)	84 (67-109)	90 (71-119)	95 (76-125)	<0.0001
eGFR	84 (58-150)	83 (59-90)	76 (52-90)	77 (55-90)	<0.0001
Potassium	4 (3.8-4.4)	4.2 (3.9-4.5)	4.1 (3.8-4.4)	4.3 (4-4.6)	<0.0001
Monocytes	0.65 (0.47-0.89)	0.62 (0.48-0.85)	0.7 (0.5-0.9)	0.66 (0.48-0.9)	<0.0001

A.4 Feature Ranking for CURIAL Models

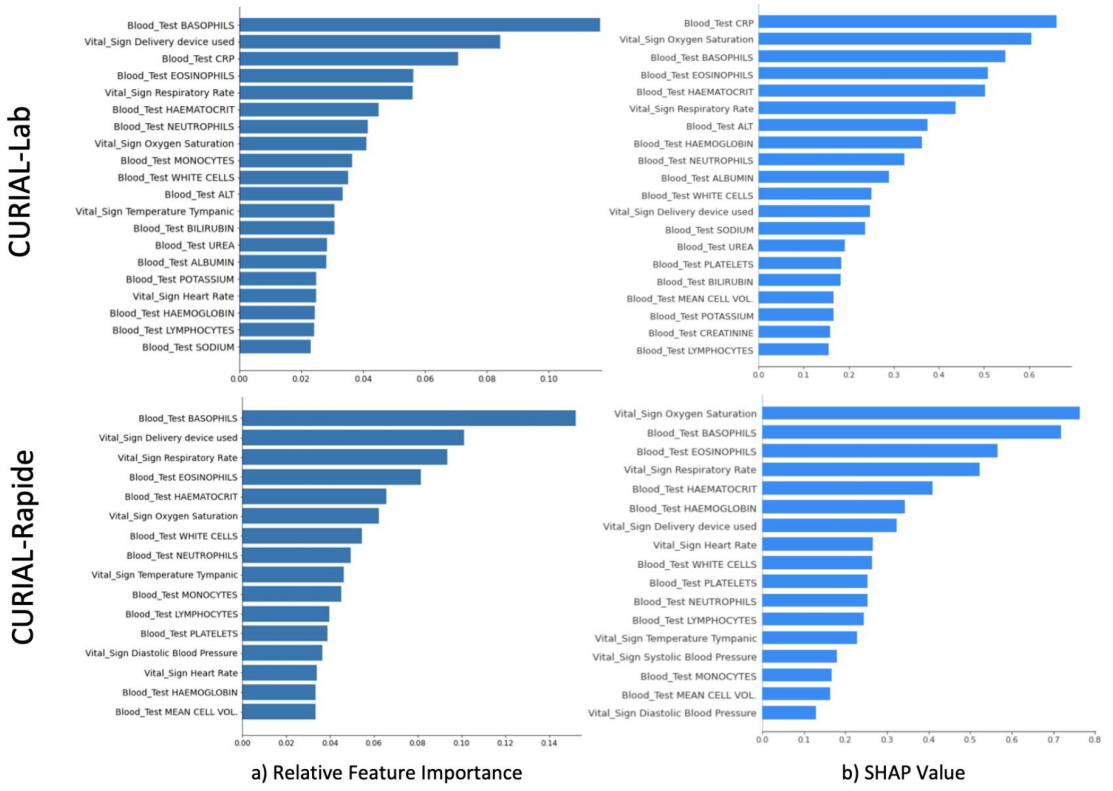

Figure A.1: Explainability analyses for CURIAL-Lab & CURIAL-Rapide. a) Relative feature importance of individual predictors within the trained models, b) SHAP (Shapley Additive Explanations) score analysis on the OUH second wave prospective set. Figure taken from [133]

B

Bias Mitigation for Supervised Learning

B.1 Software and Implementation

Models were implemented using Python (v3.6.9). Scikit Learn (v0.24.1) was used for standardization, median imputation, and dataset splitting. Performance metrics were calculated using Scikit Learn and manually programmed. XGBoost baseline models were implemented using the XGBoost library (v1.3.3). Neural network baseline models were implemented using Keras (v2.6.0). The NCR model and adversarial debiasing models were implemented using PyTorch (v1.13.1). All models were run using an Intel Xeon E-2146G Processor (CPU: 6 cores, 4.50 GHz max frequency).

B.2 Additional NCR Loss Functions

B.2.1 Jensen-Shannon Divergence

$$L_{NCR} := \frac{1}{m}\sum_{i=1}^{m} D_{JS}\left(\sigma(\mathbf{z_i}) \middle\| \sum_{j \epsilon NN_k} \frac{s_{i,j}}{\sum\limits_{k} s_{i,k}} \sigma(\mathbf{z_j})\right) \tag{B.1}$$

B.2.2 Mean Absolute Error

$$L_{NCR} := \frac{1}{m}\sum_{i=1}^{m} abs\left(\sigma(\mathbf{z_i}) - \sum_{j \epsilon NN_k} \frac{s_{i,j}}{\sum\limits_{k} s_{i,k}} \sigma(\mathbf{z_j})\right) \tag{B.2}$$

B.3 Final Hyperparameter Values

Table B.1: Final hyperparameter values used in COVID-19 status prediction for supervised learning methods. Models trained for 100 epochs.

Model	Hyperparameters
XGB	
	Learning rate = 0.1
	N estimators = 100
	Depth = 3
NN	
	Number hidden layers = 1
	Hidden nodes = 10
	Learning rate = 0.1
NN (weighted)	
	Number hidden layers = 1
	Hidden nodes = 10
	Learning rate = 0.1
XGB (weighted)	
	Learning rate = 0.1
	N estimators = 100
	Depth = 3
NN (reg.)	
	Number hidden layers = 1
	NCR starting epoch = 20
	Hidden Layer to calculate NCR = 1
	NCR weight = 1
	k = 10
	Loss = KL
ADV	
	Number hidden layers (predictor) = 1
	Number hidden layers (adversary) = 1
	Hidden nodes (predictor) = 100
	Hidden nodes (adversary) = 10
	Alpha = 1
	Learning Rate = 0.0001

C

Bias Mitigation for Reinforcement Learning

C.1 Software and Implementation

Models were implemented using Python (v3.6.9). Scikit Learn (v0.24.1) was used for standardization, median imputation, and dataset splitting. Performance metrics were calculated using Scikit Learn and manually programmed. Reinforcement learning was set up using Tensorflow (v2.6.2).

The code for the imbalanced learning and bias mitigation reinforcement learning methods are available online at `https://github.com/yangjenny/ImbalancedLearningRL` and `https://github.com/yangjenny/BiasMitigationRL`, respectively. All models were run using an Intel Xeon E-2146G Processor (CPU: 6 cores, 4.50 GHz max frequency).

C.2 Final Hyperparameter Values

Table C.1: Final hyperparameter values used in COVID-19 status prediction for reinforcement learning methods.

Model	Hyperparameters
RL	
	Number hidden layers = 1
	Gamma = 0.1
	Learning Rate = 0.00009
	Epsilon range = [0.01,1]
	Hidden nodes = 500
RL (debiasing)	
	Number hidden layers = 1
	Gamma = 0.1
	Learning Rate = 0.00009
	Epsilon range = [0.01,1]
	Hidden nodes = 500

C.3 Hospital Subgroup True Positive and False Positive Rates

Table C.2: True positive rates for hospital bias and COVID-19 status prediction test results across different models, optimized to sensitivities of 0.9.

Model	OUH	UHB	BH	PUH
RL (debiasing)	0.878	0.886	0.906	0.900
RL	0.871	0.920	0.969	0.911
XGBoost	0.871	0.920	0.906	0.858
Neural Net.	0.863	0.920	0.906	0.889
Neural Net. (weighted)	0.825	0.875	0.750	0.820
XGBoost (weighted)	0.888	0.886	0.875	0.850
Neural Net. (reg.)	0.849	0.898	0.906	0.876
Adversarial	0.860	0.920	0.906	0.903

Table C.3: False positive rates for hospital bias and COVID-19 status prediction test results across different models, optimized to sensitivities of 0.9.

Model	OUH	UHB	BH	PUH
RL (debiasing)	0.419	0.350	0.308	0.364
RL	0.502	0.393	0.328	0.432
XGBoost	0.362	0.282	0.263	0.215
Neural Net.	0.372	0.296	0.192	0.306
Neural Net. (weighted)	0.121	0.200	0.179	0.156
XGBoost (weighted)	0.349	0.264	0.273	0.215
Neural Net. (reg.)	0.338	0.262	0.207	0.274
Adversarial	0.396	0.317	0.273	0.350

C.4 Training Times

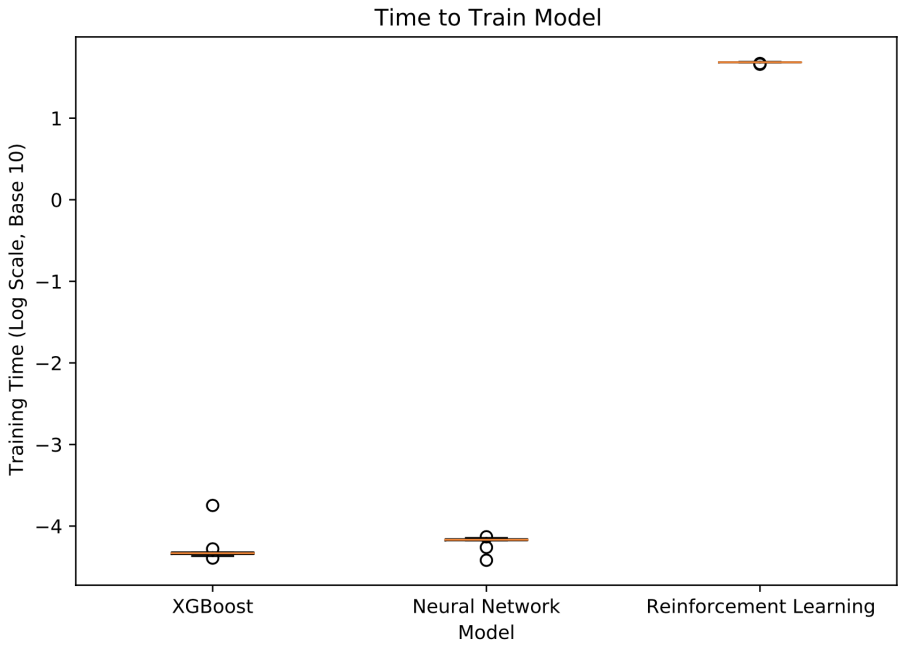

Figure C.1: Average training times across 10 training runs, for XGBoost, Neural Network, and Reinforcement Learning Models. Values displayed on logarithmic (base 10) scale.

C.5 Adjusted Thresholds Used for Binary Classification

Table C.4: Adjusted Threshold Values Used for COVID-19 status prediction (hospital-site debiasing).

Model	Threshold
	0.9
RL (debiasing)	0.46747
RL	0.49700
ADV	0.03206
NN	0.03203
XGB	0.01804
NN (weighted)	0.31231
XGB (weighted)	0.06306
NN (reg)	0.03403

C.6 Reinforcement Learning for Imbalanced Training

C.6.1 Defining Reward for Multi-class Imbalance

The reward, r_t, is the evaluation signal measuring the success of the agent's selected action. A positive reward is given when the agent correctly classifies the sample, and a negative reward is given otherwise, thus allowing the agent to learn the optimal behavior for prediction. We let the reward for accurately/inaccurately labelling an instance of a particular class be inversely proportional to the relative presence of the class in the data. The absolute reward value of a sample from the minority class is thus higher than that in the majority class, making the model more sensitive to the minority class. With l_k as the label of the sample, the reward function used is:

$$R(s_t, a_t, l_t) = \begin{cases} \lambda_k & \text{if } a_t = l_t \text{ and } l_t = k \\ -\lambda_k & \text{if } a_t \neq l_t \end{cases} \tag{C.1}$$

$$\lambda_{\mathrm{k}} = \frac{\frac{1}{N_K}}{\left\| \frac{1}{N_0}, \frac{1}{N_1}, ..., \frac{1}{N_k} \right\|^2} \tag{C.2}$$

N_k represents the number of class instances in class k and λ_k is a trade-off parameter used for adjusting the influence of the minority and majority classes.

We found that our model achieved desirable performance when λ_k is the sum of inverse squares of class frequencies, as shown in Equation C.2.

C.6.2 Model Comparators and Evaluation Metrics

We compare this imbalanced learning method against three baseline models - a fully-connected neural network and XGBoost (each with no added imbalanced learning strategy applied), as well as a DDQN and DQN (as introduced in [227]) with no dueling component. We also present results for the neural network and XGBoost models with the addition of two commonly used, state-of-the-art imbalanced data-learning methods:

SMOTE: SMOTE was applied to the training set using a minority oversampling strategy of 0.2 (i.e. the minority class was oversampled to have 20% of the number of samples in the majority class).

Cost-Sensitive Learning: Different weighted costs were assigned to each class during training. The value of class weights chosen were inversely proportional to class frequencies in the training data.

We trained both a neural network and XGBoost model as-is, and additionally trained implementations that utilized SMOTE and cost-sensitive learning. Appropriate hyperparameter values for all models were determined through standard 5-fold cross-validation (CV), using the training set. For the DDQN without a dueling component, we use the same hyperparameter settings as our method, to directly compare balanced classification performance of both methods.

To evaluate the classification performance, we calculate the sensitivity, specificity, and the area under receiver operator characteristic curve (AUROC) across all test sets, alongside 95% confidence intervals (CIs).

Since our model's objective is to train models effectively on imbalanced data, we assess the balanced classification performance using the Fowlkes–Mallows Index (FM score) and G-mean metrics. The G-mean metric is calculated as the geometric mean of recall and specificity, while the FM-score represents the geometric mean of recall and precision [227]. By utilizing geometric means, these metrics evaluate

the sensitivity and specificity of the model, ensuring that both the true positive and true negative rates are adequately considered.

As used in [227], we calculate FM-score and G-mean as follows:

$$\begin{aligned} F =& \sqrt{Sensitivity * Precision} \\ =& \sqrt{\frac{TP}{TP+FN} * \frac{TP}{TP+FP}} \end{aligned} \tag{C.3}$$

$$\begin{aligned} G =& \sqrt{Sensitivity * Specificity} \\ =& \sqrt{\frac{TP}{TP+FN} * \frac{TN}{TN+FP}} \end{aligned} \tag{C.4}$$

where TP is the number of true positives; FP is the number of false positives; TN is the number of true negatives; and FN is the number of false negatives.

C.6.3 Prediction Task and Datasets

We train models to predict the COVID-19 status for patients presenting to hospital emergency departments across four United Kingdom (UK) National Health Service (NHS) Trusts (Oxford University Hospitals NHS Foundation Trust [OUH], Portsmouth Hospitals University NHS Trust [PUH], University Hospitals Birmingham NHS Trust [UHB], Bedfordshire Hospitals NHS Foundations Trust [BH]), using anonymized EHR data (specifically, blood tests and vital sign features). We trained and optimized our model using 114,957 COVID-free patient presentations from OUH prior to the global COVID-19 outbreak, and 701 patient presentations during the first wave of the COVID-19 epidemic in the UK that had a positive PCR test for COVID-19 (ensuring that the label of COVID-19 status was correct during training). We then performed validation on a prospective OUH cohort, as well as external validation on three additional patient cohorts from PUH, UHB, and BH (totalling 72,223 admitted patients, including 4,600 of which were COVID-19 positive). During training, we used a simulated disease prevalence of 5% (i.e. a data imbalance ratio of 1 positive COVID-19 case: 20 negative controls). This aligns with real COVID-19 prevalences at all four sites (during the dates of data extraction), which ranged between 4.27%-12.2%.

C.6.4 Results

Table 1 shows results for COVID-19 prediction, where we performed prospective validation and external validation across four NHS Trusts. Scores for FM-score and G-mean are presented alongside sensitivity, specificity, and AUROC (with 95% CIs). The results presented use an adjusted decision threshold, optimized to a sensitivity of 0.9. As we are focused on evaluating balanced classification, we use red and blue to depict the best and second best scores, respectively, for F and G.

Results without any threshold adjustment (Supplementary Table 10), show that both baseline models performed poorly at predicting COVID-19 status (the minority class), achieving sensitivities below 0.5 on all test sets (mean sensitivities of 0.236 [CI range 0.071-0.388] and 0.340 [0.179-0.436] for neural network and XGBoost baseline models, respectively). The XGBoost model achieved slightly higher sensitivities than the neural network baseline, on all test sets. When SMOTE was applied to the training set, sensitivities slightly improved for both models (mean sensitivities of 0.463 [CI range 0.230-0.596] and 0.399 [0.205-0.528] for neural network and XGBoost models, respectively). When cost-sensitive learning is applied, the neural network model achieved much higher sensitivities than the baseline (mean sensitivity of 0.703 [CI range 0.539-0.785]); however, the XGBoost model only improved slightly with respect to its baseline comparator (mean sensitivity of 0.457 [CI range 0.303-0.578]). Compared to all baseline models and those additionally utilizing SMOTE and cost-sensitive weights, our method achieved the highest sensitivities, without threshold adjustment, on all test sets (mean sensitivity of 0.806 [CI range 0.733-0.864]), while maintaining high specificity as well (mean specificity of 0.756 [0.669-0.871]). The RL models without a dueling component achieved high sensitivity (mean sensitivities of 0.838 [CI range 0.765-0.888] and 0.930 [0.902-0.991] for the DDQN and DQN, respectively); however, had much lower specificity (mean specificities of 0.357 [CI range 0.280-0.440] and 0.111 [0.068-0.138]), with the DDQN architecture slightly outperforming the DQN. Comparison of the output from our method model to all other methods was found to be statistically significant ($p<0.0001$).

Although the models prior to threshold adjustment achieved poor performance on the minority class (except for our proposed RL method and the neural network with cost-sensitive weights), they still achieved reasonably high AUROC scores (>0.831, other than the RL methods without a dueling component, which achieved a slightly lower AUROC range of 0.659-0.762), suggesting that the models are able to distinguish between COVID-19 positive and negative classes. Thus, once threshold adjustment was applied, there was both higher and more balanced classification between COVID-19 positive and negative cases.

As our algorithm's primary objective is to accurately screen for COVID-19, we assess the balanced classification performance of models that can reliably predict the COVID-19 status of individuals. Specifically, we focus on models that have been optimized to achieve a sensitivity of 0.9. As shown in Table 1, all models using this optimization achieved high sensitivities (>0.792).

In terms of balanced classification, our proposed method achieved the highest F and G scores for three test sets - OUH, UHB, and BH. The XGBoost models using SMOTE and cost-sensitive weights achieved the best F and G scores on the PUH dataset. Similar results were found for models optimized to sensitivities of 0.85, with our method generally achieving the highest (or second highest) F and G scores, demonstrating model consistency. When no threshold adjustment was applied, our method also achieved the highest (or second highest) G scores on all test sets; however, F scores were not as high compared to other models that had much lower sensitivity (<0.61) but very high specificity (>0.93), due to the nature of how F is calculated. The non-dueling DDQN and DQN models consistently achieved the lowest F and G scores, across all test sets (recall that it also achieved the lowest classification performance). Comparison of the output from our method to all other methods was found to be statistically significant ($p<0.0001$).

All models trained achieved reasonably high AUROC scores across all test sets, demonstrating that we have trained strong models to begin with. Thus, the results show that the proposed method is both a strong classifier, in addition to being able to account for large data imbalances.

Table C.5: Performance metrics for COVID-19 prediction. Results reported as FM-score, G-mean, AUROC, sensitivity, and specificity for OUH, PUH, UHB, and BH test sets; 95% confidence intervals (CIs) also shown. Red and blue values denote best and second best scores, respectively, for FM-score and G-mean. Threshold adjustment applied to optimize models to sensitivities of 0.9.

Model	F	G	AUROC	Sensitivity	Specificity
OUH					
Reinforcement Learning (Ours)	**0.426**	**0.770**	0.861 (0.850-0.871)	0.838 (0.822-0.854)	0.707 (0.701-0.713)
Reinforcement Learning (DDQN)	0.306	0.534	0.758 (0.745-0.771)	0.852 (0.836-0.867)	0.334 (0.328-0.341)
Reinforcement Learning (DQN) [13]	0.293	0.350	0.751 (0.739-0.764)	0.921 (0.909-0.933)	0.133 (0.128-0.138)
Neural Network	0.388	0.715	0.877 (0.867-0.886)	0.899 (0.885-0.912)	0.568 (0.562-0.575)
Neural Network + SMOTE	0.398	0.736	0.871 (0.861-0.881)	0.871 (0.856-0.885)	0.622 (0.615-0.628)
Neural Network + Cost-Sensitive	0.400	0.737	0.872 (0.862-0.882)	0.881 (0.867-0.895)	0.616 (0.609-0.623)
XGBoost	0.399	0.734	0.877 (0.867-0.887)	0.889 (0.875-0.902)	0.607 (0.600-0.614)
XGBoost + SMOTE	**0.422**	**0.766**	0.876 (0.866-0.886)	0.846 (0.830-0.862)	0.694 (0.687-0.700)
XGBoost + Cost-Sensitive	0.399	0.739	0.869 (0.859-0.879)	0.857 (0.842-0.872)	0.637 (0.630-0.643)
PUH					
Reinforcement Learning (Ours)	0.306	0.727	0.831 (0.819-0.842)	0.828 (0.812-0.845)	0.638 (0.633-0.643)
Reinforcement Learning (DDQN)	0.248	0.606	0.762 (0.750-0.774)	0.804 (0.787-0.821)	0.457 (0.451-0.462)
Reinforcement Learning (DQN) [13]	0.223	0.324	0.732 (0.719-0.745)	0.915 (0.902-0.927)	0.115 (0.112-0.118)
Neural Network	0.289	0.676	0.857 (0.847-0.868)	0.903 (0.890-0.916)	0.506 (0.501-0.511)
Neural Network + SMOTE	0.309	0.728	0.856 (0.845-0.866)	0.859 (0.844-0.875)	0.617 (0.612-0.622)
Neural Network + Cost-Sensitive	0.288	0.681	0.850 (0.839-0.861)	0.883 (0.869-0.897)	0.526 (0.521-0.531)
XGBoost	0.321	0.741	0.881 (0.871-0.891)	0.898 (0.884-0.911)	0.612 (0.607-0.617)
XGBoost + SMOTE	**0.325**	**0.750**	0.881 (0.871-0.890)	0.877 (0.863-0.892)	0.641 (0.636-0.646)
XGBoost + Cost-Sensitive	**0.336**	**0.766**	0.881 (0.871-0.891)	0.862 (0.847-0.877)	0.680 (0.675-0.684)
UHB					
Reinforcement Learning (Ours)	**0.304**	**0.764**	0.837 (0.814-0.861)	0.815 (0.779-0.852)	0.717 (0.708-0.726)
Reinforcement Learning (DDQN)	0.209	0.516	0.721 (0.694-0.749)	0.841 (0.806-0.875)	0.317 (0.308-0.326)
Reinforcement Learning (DQN) [13]	0.203	0.322	0.723 (0.695-0.750)	0.927 (0.903-0.951)	0.112 (0.106-0.118)
Neural Network	0.279	0.718	0.866 (0.844-0.888)	0.913 (0.887-0.940)	0.565 (0.555-0.574)
Neural Network + SMOTE	0.290	0.746	0.850 (0.828-0.873)	0.845 (0.811-0.879)	0.658 (0.649-0.668)
Neural Network + Cost-Sensitive	0.284	0.733	0.861 (0.839-0.883)	0.879 (0.849-0.910)	0.611 (0.601-0.621)
XGBoost	0.287	0.740	0.861 (0.839-0.883)	0.872 (0.841-0.904)	0.627 (0.618-0.637)
XGBoost + SMOTE	**0.292**	**0.750**	0.853 (0.830-0.876)	0.827 (0.791-0.862)	0.680 (0.671-0.690)
XGBoost + Cost-Sensitive	0.289	0.746	0.851 (0.829-0.874)	0.838 (0.804-0.873)	0.663 (0.654-0.673)
BH					
Reinforcement Learning (Ours)	**0.561**	**0.815**	0.867 (0.829-0.906)	0.806 (0.741-0.870)	0.825 (0.802-0.848)
Reinforcement Learning (DDQN)	0.362	0.589	0.706 (0.656-0.756)	0.799 (0.733-0.864)	0.434 (0.403-0.464)
Reinforcement Learning (DQN) [13]	0.349	0.286	0.659 (0.608-0.710)	0.958 (0.926-0.991)	0.085 (0.068-0.102)
Neural Network	0.525	0.802	0.885 (0.849-0.921)	0.868 (0.813-0.923)	0.741 (0.714-0.767)
Neural Network + SMOTE	**0.540**	0.801	0.882 (0.845-0.919)	0.792 (0.725-0.858)	0.810 (0.786-0.834)
Neural Network + Cost-Sensitive	0.529	**0.804**	0.883 (0.847-0.920)	0.854 (0.797-0.912)	0.756 (0.730-0.782)
XGBoost	0.501	0.780	0.894 (0.859-0.929)	0.896 (0.846-0.946)	0.679 (0.650-0.707)
XGBoost + SMOTE	0.535	0.803	0.885 (0.849-0.921)	0.819 (0.757-0.882)	0.787 (0.762-0.812)
XGBoost + Cost-Sensitive	0.511	0.790	0.889 (0.854-0.925)	0.861 (0.805-0.918)	0.724 (0.697-0.751)

D

Mitigating Machine Learning Bias Between High-Income and Low-Middle Income Countries

D.1 Software and Implementation

Experiments were performed using Python (v3.8.3). All models were run using an Intel Xeon E-2146G Processor (CPU: 6 cores, 4.50 GHz max frequency). Statistical tests were performed using the SciPy (Statistical Functions) package (v1.10.1). Scikit Learn (v1.3.2) was used for standardization, median imputation, and calculating performance metrics. Performance metrics were calculated using Scikit Learn and manually programmed.

D.2 Patient Inclusion and Exclusion Criteria

The ethics committees of the Hospital for Tropical Diseases (HTD) and the National Hospital for Tropical Diseases (NHTD) approved use of the HTD and NHTD datasets for COVID-19 diagnosis, respectively.

Hospital for Tropical Diseases (HTD): HTD considered all patients admitted between December 10, 2020 and December 30,2022. Confirmatory COVID-19 testing was performed using PCR.

National Hospital for Tropical Diseases (NHTD): NHTD considered all patients admitted between November 1, 2020 and December 21,2022. Confirmatory COVID-19 testing was performed using PCR and/or rapid antigen testing.

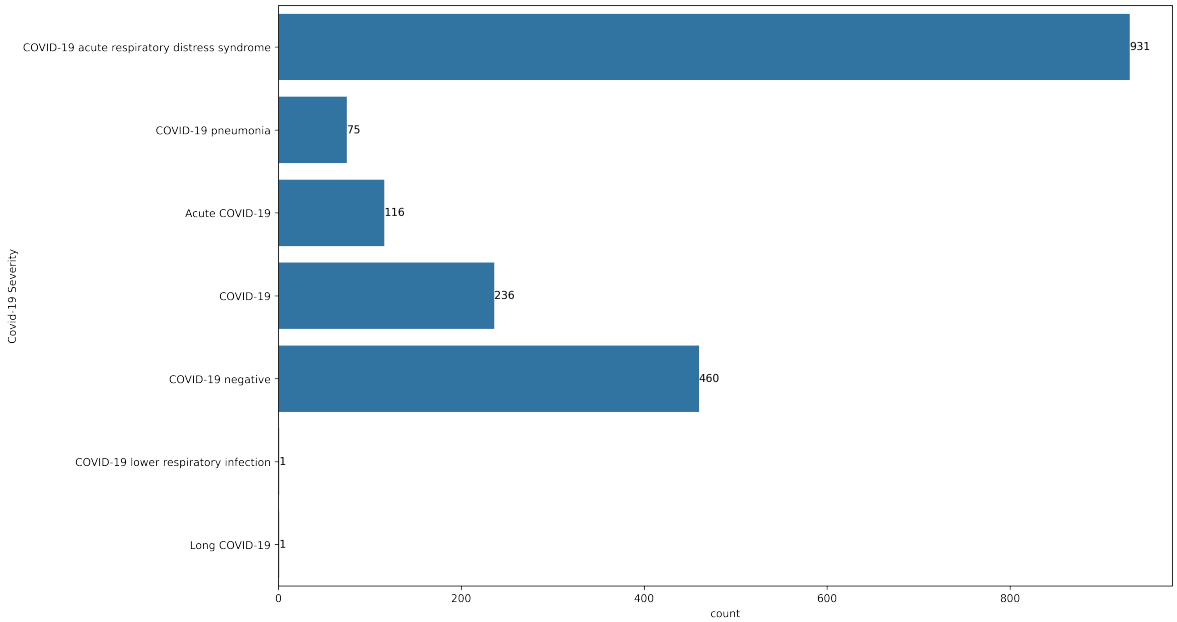

Figure D.1: Levels of COVID-19 severity recorded for patients admitted to the Hospital for Tropical Diseases.

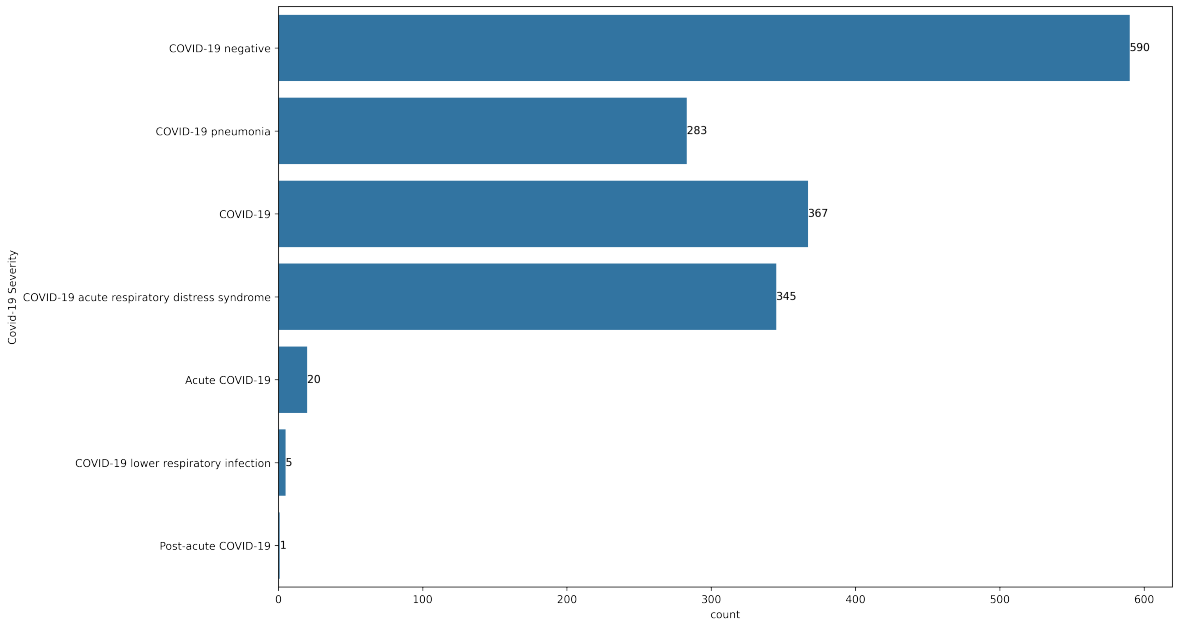

Figure D.2: Levels of COVID-19 severity recorded for patients admitted to the National Hospital for Tropical Diseases.

D.3 Final Hyperparameter Values

Table D.1: Final hyperparameter values used in COVID-19 status prediction during collaborative training between UK and Vietnam hospital sites.

Model	Hyperparameters
XGB	
	Learning rate = 0.1
	N estimators = 100
	Depth = 3
NN	
	Number hidden layers = 1
	Hidden nodes = 10
	Learning rate = 0.05
RL (debiasing)	
	Number hidden layers = 1
	Gamma = 0.1
	Learning Rate = 0.0001
	Epsilon range = [0.01,1]
	Hidden nodes = 500
ADV	
	Number hidden layers (predictor) = 1
	Number hidden layers (adversary) = 1
	Hidden nodes (predictor) = 100
	Hidden nodes (adversary) = 100
	Alpha = 1
	Learning Rate = 0.0001

D.4 Distribution of Clinical Features for Generalizability Task

Table D.2: Distribution of vital signs, reported as median and interquartile ranges, for each patient cohort.

	Oxford University Hospitals (pre-pandemic & wave 1 cases, to June 30/20)		Oxford University Hospitals	Portsmouth Hospitals University NHS Trust	University Hospitals Birmingham NHS Foundation Trust	Bedfordshire Hospitals NHS Foundation Trust	Hospital for Tropical Diseases	National Hospital for Tropical Diseases	Kruskal-Wallis, p-value
	Prepandemic	COVID-19-cases	Oct 1/20-Mar 6/21	Mar 1/20-Feb 28/21	Dec 1/19 - Oct 29/20	Jan 1/21-Mar 31/21	Jan 1/21-Dec 31/22	Jan 1/21-Dec 31/22	
Respiratory Rate (breath/min)	18.0 (16.0-19.0)	20.0 (18.0-24.0)	18.0 (16.6-19.0)	17.0 (16.0-19.0)	18.0 (17.0-20.0)	18.0 (16.0-20.0)	24.0(20.0-26.5)	25.0(22.0-30.0)	<0.0001
Heart Rate (beats/min)	82.0 (71.0-96.0)	88.0 (75.0-101.0)	84.0 (72.0-97.0)	82.0 (71.0-95.0)	86.0 (73.0-101.0)	84.0 (73.0-97.0)	94.0(84.0-108.0)	99.0(86.0-114.0)	<0.0001
Systolic Blood Pressure (mmHg)	132.0 (118.0-150.0)	131.0 (115.0-146.0)	134.0 (119.0-152.0)	128.0 (114.0-146.0)	136.0 (119.0-155.0)	131.0 (116.0-149.0)	130.0(111.5-140.0)	120.0(110.0-136.0)	<0.0001
Diastolic Blood Pressure (mmHg)	74.0 (65.0-84.0)	74.0 (64.0-84.0)	75.0 (65.0-85.0)	76.0 (67.0-84.0)	77.0 (68.0-87.0)	78.0 (68.0-88.0)	80.0(70.0-80.0)	70.0(60.0-80.0)	<0.0001
Tympanic Temperature (C)	36.5 (36.1-36.9)	36.9 (36.3-37.6)	36.3 (36.0-36.7)	36.3 (36.0-36.8)	36.7 (36.4-37.2)	36.5 (36.4-36.9)	37.0(37.0-37.3)	37.0(36.8-37.5)	<0.0001

Table D.3: Distribution of blood test features, reported as median and interquartile ranges, for each patient cohort.

	Oxford University Hospitals (pre-pandemic & wave 1 cases, to June 30/20)		Oxford University Hospitals	Portsmouth Hospitals University NHS Trust	University Hospitals Birmingham NHS Foundation Trust	Bedfordshire Hospitals NHS Foundation Trust	Hospital for Tropical Diseases	National Hospital for Tropical Diseases	Kruskal-Wallis, p-value
	Prepandemic cohort	COVID-19-cases cohort	Oct 1/20-Mar 6/21	Mar 1/20-Feb 28/21	Dec 1/19 - Oct 29/20	Jan 1/21-Mar 31/21	Jan 1/21-Dec 31/22	Jan 1/21-Dec 31/22	
HAEMOGLOBIN (g/L)	130.0 (116.0-142.0)	130.0 (114.0-144.0)	129.0 (114.0-142.0)	129.0 (114.0-143.0)	127.0 (113.0-140.0)	134.0 (119.0-146.0)	128.0(113.0-141.0)	112.0(95.0-127.0)	<0.0001
WHITE CELLS ($10^9 1^{-1}$)	8.45 (6.46-11.18)	6.98 (5.14-9.72)	8.94 (6.7-12.06)	8.6 (6.7-11.3)	9.4 (7.1-12.6)	9.2 (6.9-12.5)	9.55(6.69-13.565)	10.9(7.525-15.0)	<0.0001
PLATELETS ($10^9 1^{-1}$)	249.0 (199.0-307.0)	215.0 (163.0-283.5)	251.0 (198.0-314.0)	251.0 (199.0-312.0)	247.0 (196.0-311.0)	246.0 (196.0-310.0)	216.0(152.0-281.0)	189.5(116.75-260.25)	<0.0001
HAEMATOCRIT	0.39 (0.35-0.42)	0.4 (0.35-0.44)	0.39 (0.35-0.43)	0.39 (0.34-0.42)	0.38 (0.34-0.42)	0.39 (0.35-0.43)	0.392(0.35-0.428)	0.342(0.289-0.387)	<0.0001
SODIUM (mM)	138.0 (136.0-140.0)	136.0 (134.0-139.0)	138.0 (135.0-140.0)	138.0 (136.0-140.0)	137.0 (134.0-139.0)	138.0 (136.0-140.0)	134.0(130.0-137.0)	136.0(132.65-139.65)	<0.0001
UREA (mM)	5.3 (4.0-7.4)	5.9 (4.2-9.07)	5.7 (4.2-8.3)	5.2 (3.8-7.6)	6.2 (4.5-9.0)	5.8 (4.2-8.3)	7.5(5.0-12.75)	7.215(4.8-11.7)	<0.0001
BILIRUBIN (umol/L)	9.0 (6.0-13.0)	9.0 (7.0-13.25)	9.0 (6.0-14.0)	10.0 (7.0-16.0)	10.0 (7.0-15.0)	10.0 (7.0-14.0)	13.7(7.8-36.7)	9.8(6.4-16.375)	<0.0001
CREATININE (umol/L)	73.0 (60.0-93.0)	79.0 (65.0-106.0)	74.0 (60.0-97.0)	74.0 (60.0-96.0)	78.0 (62.0-105.0)	80.5 (65.75-104.0)	76.0(61.0-99.0)	77.65(56.0-121.0)	<0.0001
POTASSIUM (mM)	4.0 (3.7-4.3)	4.0 (3.7-4.3)	4.0 (3.8-4.4)	4.2 (3.9-4.4)	4.1 (3.8-4.4)	4.3 (4.0-4.6)	3.72(3.35-4.09)	3.8(3.4-4.3)	<0.0001
MEAN CELL VOL (fl)	89.6 (86.0-93.4)	90.2 (86.6-94.2)	90.2 (86.6-94.2)	89.0 (84.9-93.0)	89.9 (86.2-93.6)	88.0 (85.0-92.0)			0.210
NEUTROPHILS ($10^9 1^{-1}$)	5.72 (3.99-8.36)	5.11 (3.48-7.49)	6.44 (4.4-9.55)	5.9 (4.2-8.6)	6.9 (4.7-10.0)	6.8 (4.7-9.73)			<0.0001
LYMPHOCYTES ($10^9 1^{-1}$)	1.51 (1.0-2.13)	0.96 (0.65-1.38)	1.31 (0.85-1.89)	1.5 (0.97-2.2)	1.3 (0.9-1.9)	1.27 (0.86-1.83)			<0.0001
MONOCYTES ($10^9 1^{-1}$)	0.64 (0.48-0.85)	0.49 (0.35-0.74)	0.66 (0.48-0.89)	0.63 (0.48-0.85)	0.7 (0.5-0.9)	0.66 (0.48-0.92)			<0.0001
EOSINOPHILS ($10^9 1^{-1}$)	0.1 (0.04-0.2)	0.01 (0.0-0.06)	0.07 (0.02-0.16)	0.1 (0.02-0.2)	0.1 (0.0-0.2)	0.06 (0.02-0.16)			<0.0001
BASOPHILS ($10^9 1^{-1}$)	0.04 (0.03-0.06)	0.02 (0.01-0.03)	0.04 (0.02-0.06)	0.04 (0.02-0.06)	0.1 (0.0-0.1)	0.05 (0.03-0.07)			<0.0001
ALBUMIN (g/L)	36.0 (32.0-39.0)	32.0 (28.0-35.0)	36.0 (31.0-39.0)	36.0 (31.0-40.0)	36.0 (32.0-40.0)	35.0 (31.0-39.0)			0.006
ALKALINE PHOSPHATASE (IU/L)	80.0 (64.0-105.0)	82.0 (64.0-108.0)	84.0 (66.0-112.0)	84.0 (67.0-109.0)	90.0 (71.0-119.0)	94.0 (74.5-122.0)			<0.0001
ALT (IU/L)	18.0 (13.0-28.0)	25.0 (17.0-41.0)	20.0 (13.0-33.0)	19.0 (13.0-30.0)	19.0 (13.0-30.0)	20.0 (13.0-31.0)			<0.0001
eGFR (ml/min)	85.0 (63.0-150.0)	78.0 (53.0-150.0)	84.0 (58.0-150.0)	83.0 (60.0-90.0)	76.0 (52.0-90.0)	76.0 (54.0-90.0)			0.011
CRP (mg/L)	8.6 (2.3-39.0)	72.5 (23.8-143.6)	15.8 (3.5-67.4)	13.0 (3.0-71.0)	12.0 (3.0-61.0)	10.7 (2.8-48.78)			0.003

D.5 Distribution of Clinical Features for Bias Mitigation Task

Table D.4: Distribution of vital signs, reported as median and interquartile ranges, for each patient cohort.

	Oxford University Hospitals (pre-pandemic & wave 1 cases, to June 30/20)		Oxford University Hospitals	Portsmouth Hospitals University NHS Trust	University Hospitals Birmingham NHS Foundation Trust	Bedfordshire Hospitals NHS Foundation Trust	Hospital for Tropical Diseases	Kruskal-Wallis, p-value
	Prepandemic cohort	COVID-19-cases cohort	Oct 1/20-Mar 6/21	Mar 1/20-Feb 28/21	Dec 1/19 - Oct 29/20	Jan 1/21-Mar 31/21	Jan 1/21-Dec 31/22	
Respiratory Rate (breath/min)	18.0 (16.0-19.0)	20.0 (18.0-24.0)	18.0 (16.6-19.0)	17.0 (16.0-19.0)	18.0 (17.0-20.0)	18.0 (16.0-20.0)	24.0(20.0-26.5)	<0.0001
Heart Rate (beats/min)	82.0 (71.0-96.0)	88.0 (75.0-101.0)	84.0 (72.0-97.0)	82.0 (71.0-95.0)	86.0 (73.0-101.0)	84.0 (73.0-97.0)	94.0(84.0-108.0)	<0.0001
Systolic Blood Pressure (mmHg)	132.0 (118.0-150.0)	131.0 (115.0-146.0)	134.0 (119.0-152.0)	128.0 (114.0-146.0)	136.0 (119.0-155.0)	131.0 (116.0-149.0)	130.0(111.5-140.0)	<0.0001
Diastolic Blood Pressure (mmHg)	74.0 (65.0-84.0)	74.0 (64.0-84.0)	75.0 (65.0-85.0)	76.0 (67.0-84.0)	77.0 (68.0-87.0)	78.0 (68.0-88.0)	80.0(70.0-80.0)	<0.0001
Temperature (C)	36.5 (36.1-36.9)	36.9 (36.3-37.6)	36.3 (36.0-36.7)	36.3 (36.0-36.8)	36.7 (36.4-37.2)	36.5 (36.4-36.9)	37.0(37.0-37.3)	<0.0001

Table D.5: Distribution of blood test features, reported as median and interquartile ranges, for each patient cohort.

	Oxford University Hospitals (pre-pandemic & wave 1 cases, to June 30/20)		Oxford University Hospitals	Portsmouth Hospitals University NHS Trust	University Hospitals Birmingham NHS Foundation Trust	Bedfordshire Hospitals NHS Foundation Trust	Hospital for Tropical Diseases	Kruskal-Wallis, p-value
	Prepandemic cohort	COVID-19-cases cohort	Oct 1/20-Mar 6/21	Mar 1/20-Feb 28/21	Dec 1/19 - Oct 29/20	Jan 1/21-Mar 31/21	Jan 1/21-Dec 31/22	
HAEMOGLOBIN (g/L)	130.0 (116.0-142.0)	130.0 (114.0-144.0)	129.0 (114.0-142.0)	129.0 (114.0-143.0)	127.0 (113.0-140.0)	134.0 (119.0-146.0)	128.0(113.0-141.0)	<0.0001
WHITE CELLS (10^9l^{-1})	8.45 (6.46-11.18)	6.98 (5.14-9.72)	8.94 (6.7-12.06)	8.6 (6.7-11.3)	9.4 (7.1-12.6)	9.2 (6.9-12.5)	9.55(6.69-13.565)	<0.0001
PLATELETS (10^9l^{-1})	249.0 (199.0-307.0)	215.0 (163.0-283.5)	251.0 (198.0-314.0)	251.0 (199.0-312.0)	247.0 (196.0-311.0)	246.0 (196.0-310.0)	216.0(152.0-281.0)	<0.0001
HAEMATOCRIT	0.39 (0.35-0.42)	0.4 (0.35-0.44)	0.39 (0.35-0.43)	0.39 (0.34-0.42)	0.38 (0.34-0.42)	0.39 (0.35-0.43)	0.392(0.35-0.428)	<0.0001
SODIUM (mM)	138.0 (136.0-140.0)	136.0 (134.0-139.0)	138.0 (135.0-140.0)	138.0 (136.0-140.0)	137.0 (134.0-139.0)	138.0 (136.0-140.0)	134.0(130.0-137.0)	<0.0001
CREATININE (umol/L)	73.0 (60.0-93.0)	79.0 (65.0-106.0)	74.0 (60.0-97.0)	74.0 (60.0-96.0)	78.0 (62.0-105.0)	80.5 (65.75-104.0)	76.0(61.0-99.0)	<0.0001
POTASSIUM (mM)	4.0 (3.7-4.3)	4.0 (3.7-4.3)	4.0 (3.8-4.4)	4.2 (3.9-4.4)	4.1 (3.8-4.4)	4.3 (4.0-4.6)	3.72(3.35-4.09)	<0.0001
MEAN CELL VOL (fl)	89.6 (86.0-93.4)	90.2 (86.6-94.2)	90.2 (86.6-94.2)	89.0 (84.9-93.0)	89.9 (86.2-93.6)	88.0 (85.0-92.0)	89.75(85.675-93.4)	<0.0001
NEUTROPHILS (10^9l^{-1})	5.72 (3.99-8.36)	5.11 (3.48-7.49)	6.44 (4.4-9.55)	5.9 (4.2-8.6)	6.9 (4.7-10.0)	6.8 (4.7-9.73)	7.705(4.92-11.3)	<0.0001
LYMPHOCYTES (10^9l^{-1})	1.51 (1.0-2.13)	0.96 (0.65-1.38)	1.31 (0.85-1.89)	1.5 (0.97-2.2)	1.3 (0.9-1.9)	1.27 (0.86-1.83)	0.93(0.57-1.54)	<0.0001
MONOCYTES (10^9l^{-1})	0.64 (0.48-0.85)	0.49 (0.35-0.74)	0.66 (0.48-0.89)	0.63 (0.48-0.85)	0.7 (0.5-0.9)	0.66 (0.48-0.92)	0.46(0.28-0.69)	<0.0001
EOSINOPHILS (10^9l^{-1})	0.1 (0.04-0.2)	0.01 (0.0-0.06)	0.07 (0.02-0.16)	0.1 (0.02-0.2)	0.1 (0.0-0.2)	0.06 (0.02-0.16)	0.02(0.0-0.08)	<0.0001
BASOPHILS (10^9l^{-1})	0.04 (0.03-0.06)	0.02 (0.01-0.03)	0.04 (0.02-0.06)	0.04 (0.02-0.06)	0.1 (0.0-0.1)	0.05 (0.03-0.07)	0.01(0.01-0.03)	<0.0001
ALT (IU/L)	18.0 (13.0-28.0)	25.0 (17.0-41.0)	20.0 (13.0-33.0)	19.0 (13.0-30.0)	19.0 (13.0-30.0)	20.0 (13.0-31.0)	33.0(20.0-60.0)	<0.0001

E

Additional Case Studies

E.1 Software and Implementation

Models were implemented using Python (v3.6.9). Scikit Learn (v0.24.1) was used for standardization, median imputation, and dataset splitting. Performance metrics were calculated using Scikit Learn and manually programmed. Reinforcement learning was set up using Tensorflow (v2.6.2). The code for the imbalanced learning and bias mitigation reinforcement learning methods are available online at `https://github.com/yangjenny/ImbalancedLearningRL` and `https://github.com/yangjenny/BiasMitigationRL`, respectively. All models were run using an Intel Xeon E-2146G Processor (CPU: 6 cores, 4.50 GHz max frequency).

E.2 Final Hyperparameter Values

Table E.1: Final hyperparameter values used in COVID-19 status prediction (mitigating ethnicity bias).

Model	Hyperparameters
XGB	
	Learning rate = 0.1
	N estimators = 100
	Depth = 3
NN	
	Number hidden layers = 1
	Hidden nodes = 20
	Learning rate = 0.1
RL (debiasing)	
	Number hidden layers = 1
	Gamma = 0.1
	Learning Rate = 0.0001
	Epsilon range = [0.01,1]
	Hidden nodes = 150
ADV	
	Number hidden layers (predictor) = 1
	Number hidden layers (adversary) = 1
	Hidden nodes (predictor) = 100
	Hidden nodes (adversary) = 10
	Alpha = 10
	Learning Rate = 0.0001

Table E.2: Final hyperparameter values used in ICU patient discharge status prediction (mitigating ethnicity bias).

Model	Hyperparameters
XGB	
	Learning rate = 0.1
	N estimators = 100
	Depth = 3
NN	
	Number hidden layers = 1
	Hidden nodes = 200
	Learning rate = 0.1
RL (debiasing)	
	Number hidden layers = 1
	Gamma = 0.1
	Learning Rate = 0.00009
	Epsilon range = [0.01,1]
	Hidden nodes = 500
ADV	
	Number hidden layers (predictor) = 1
	Number hidden layers (adversary) = 1
	Hidden nodes (predictor) = 200
	Hidden nodes (adversary) = 10
	Alpha = 1
	Learning Rate = 0.0001

References

1. Moor, J. *The Turing test: the elusive standard of artificial intelligence* (Springer Science & Business Media, 2003).
2. Kaul, V., Enslin, S. & Gross, S. A. History of artificial intelligence in medicine. *Gastrointestinal Endoscopy* **92,** 807–812 (2020).
3. Malik, P., Pathania, M., Rathaur, V. K., *et al.* Overview of artificial intelligence in medicine. *Journal of Family Medicine and Primary Care* **8,** 2328 (2019).
4. Alowais, S. A. *et al.* Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Medical Education* **23,** 689 (2023).
5. Moran, M. E. Evolution of robotic arms. *Journal of Robotic Surgery* **1,** 103–111 (2007).
6. Russell, S. J. & Norvig, P. *Artificial intelligence: a modern approach* (Pearson, 2016).
7. Bonaccorso, G. *Machine learning algorithms* (Packt Publishing Ltd, 2017).
8. Busnatu, Ș. *et al.* Clinical applications of artificial intelligence—An updated overview. *Journal of Clinical Medicine* **11,** 2265 (2022).
9. Jiang, F. *et al.* Artificial intelligence in healthcare: past, present and future. *Stroke and Vascular Neurology* **2** (2017).
10. Vokinger, K. N., Feuerriegel, S. & Kesselheim, A. S. Mitigating bias in machine learning for medicine. *Communications Medicine* **1,** 25 (2021).
11. November, J. *et al.* Beginnings of artificial intelligence in medicine (AIM): computational artifice assisting scientific inquiry and clinical art–with reflections on present aim challenges. *Yearbook of Medical Informatics* **28,** 249–256 (2019).
12. Kulikowski, C. An opening chapter of the first generation of artificial intelligence in medicine: the first rutgers AIM workshop, June 1975. *Yearbook of Medical Informatics* **24,** 227–233 (2015).
13. Weiss, S., Kulikowski, C. A. & Safir, A. Glaucoma consultation by computer. *Computers in Biology and Medicine* **8,** 25–40 (1978).
14. Shortliffe, E. H. *et al.* Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system. *Computers and Biomedical Research* **8,** 303–320 (1975).
15. Bohr, A. & Memarzadeh, K. in *Artificial Intelligence in Healthcare* 25–60 (Elsevier, 2020).
16. Aggarwal, R. *et al.* Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *NPJ Digital Medicine* **4,** 65 (2021).

17. Liu, X. *et al.* A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health* **1,** e271–e297 (2019).

18. Smith, K. P. & Kirby, J. E. Image analysis and artificial intelligence in infectious disease diagnostics. *Clinical Microbiology and Infection* **26,** 1318–1323 (2020).

19. Karandikar, P. *et al.* Machine learning applications of surgical imaging for the diagnosis and treatment of spine disorders: current state of the art. *Neurosurgery* **90,** 372–382 (2022).

20. Meena, T. & Roy, S. Bone fracture detection using deep supervised learning from radiological images: A paradigm shift. *Diagnostics* **12,** 2420 (2022).

21. Rahman, R. & Reddy, C. K. Electronic Health Records: A Survey. *Healthcare Data Analytics* **36,** 21 (2015).

22. Jensen, P. B., Jensen, L. J. & Brunak, S. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics* **13,** 395–405 (2012).

23. Yang, J., Eyre, D. W., Lu, L. & Clifton, D. A. Interpretable machine learning-based decision support for prediction of antibiotic resistance for complicated urinary tract infections. *NPJ Antimicrobials and Resistance* **1,** 14 (2023).

24. Quazi, S. Artificial intelligence and machine learning in precision and genomic medicine. *Medical Oncology* **39,** 120 (2022).

25. Libbrecht, M. W. & Noble, W. S. Machine learning applications in genetics and genomics. *Nature Reviews Genetics* **16,** 321–332 (2015).

26. Zou, J. *et al.* A primer on deep learning in genomics. *Nature Genetics* **51,** 12–18 (2019).

27. Huang, K. *et al.* Machine learning applications for therapeutic tasks with genomics data. *Patterns* **2** (2021).

28. Pun, F. W., Ozerov, I. V. & Zhavoronkov, A. AI-powered therapeutic target discovery. *Trends in Pharmacological Sciences* (2023).

29. Vatansever, S. *et al.* Artificial intelligence and machine learning-aided drug discovery in central nervous system diseases: State-of-the-arts and future directions. *Medicinal Research Reviews* **41,** 1427–1473 (2021).

30. Seyhan, A. A. & Carini, C. Are innovation and new technologies in precision medicine paving a new era in patients centric care? *Journal of Translational Medicine* **17,** 114 (2019).

31. Biswas, N. & Chakrabarti, S. Artificial intelligence (AI)-based systems biology approaches in multi-omics data analysis of cancer. *Frontiers in Oncology* **10,** 588221 (2020).

32. Skeppstedt, M., Kvist, M., Nilsson, G. H. & Dalianis, H. Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study. *Journal of Biomedical Informatics* **49,** 148–158 (2014).

33. Spasic, I., Nenadic, G., *et al.* Clinical text data in machine learning: systematic review. *JMIR Medical Informatics* **8,** e17984 (2020).

34. Chodey, K. P. & Hu, G. *Clinical text analysis using machine learning methods* in *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)* (2016), 1–6.

35. Raja, K. & Jonnalagadda, S. Natural Language Processing and Data Mining for Clinical Text. *Healthcare Data Analytics* **36,** 219 (2015).

36. Sheikhalishahi, S. *et al.* Natural language processing of clinical notes on chronic diseases: systematic review. *JMIR Medical Informatics* **7,** e12239 (2019).

37. Iroju, O. G. & Olaleke, J. O. A systematic review of natural language processing in healthcare. *International Journal of Information Technology and Computer Science* **8,** 44–50 (2015).

38. Dymek, C. *et al.* Building the evidence-base to reduce electronic health record–related clinician burden. *Journal of the American Medical Informatics Association* **28,** 1057–1061 (2021).

39. Dunn, J. *et al.* Wearable sensors enable personalized predictions of clinical laboratory measurements. *Nature Medicine* **27,** 1105–1112 (2021).

40. Clifton, L., Clifton, D. A., Pimentel, M. A., Watkinson, P. J. & Tarassenko, L. Predictive monitoring of mobile patients by combining clinical observations with data from wearable sensors. *IEEE Journal of Biomedical and Health Informatics* **18,** 722–730 (2013).

41. Kubota, K. J., Chen, J. A. & Little, M. A. Machine learning for large-scale wearable sensor data in Parkinson's disease: Concepts, promises, pitfalls, and futures. *Movement Disorders* **31,** 1314–1326 (2016).

42. Nurmi, J. & Lohan, E. S. Systematic review on machine-learning algorithms used in wearable-based eHealth data analysis. *IEEE Access* **9,** 112221–112235 (2021).

43. Castaneda, D., Esparza, A., Ghamari, M., Soltanpur, C. & Nazeran, H. A review on wearable photoplethysmography sensors and their potential future applications in health care. *International Journal of Biosensors & Bioelectronics* **4,** 195 (2018).

44. Yang, J. *et al.* Machine learning-based risk stratification for gestational diabetes management. *Sensors* **22,** 4805 (2022).

45. Mishra, S. *et al.* A review: Recent advancements in sensor technology for non-invasive neonatal health monitoring. *Biosensors and Bioelectronics: X,* 100332 (2023).

46. McKinney, S. M. *et al.* International evaluation of an AI system for breast cancer screening. *Nature* **577,** 89–94 (2020).

47. Kim, H.-E. *et al.* Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *The Lancet Digital Health* **2,** e138–e148 (2020).

48. Yelin, I. *et al.* Personal clinical history predicts antibiotic resistance of urinary tract infections. *Nature Medicine* **25,** 1143–1152 (2019).

49. Kanjilal, S. *et al.* A decision algorithm to promote outpatient antimicrobial stewardship for uncomplicated urinary tract infection. *Science Translational Medicine* **12,** eaay5067 (2020).

50. Groh, M. *et al.* Deep learning-aided decision support for diagnosis of skin disease across skin tones. *Nature* **577,** 89–94 (2024).

51. Fernandez-Quilez, A. Deep learning in radiology: ethics of data and on the value of algorithm transparency, interpretability and explainability. *AI and Ethics* **3,** 257–265 (2023).

52. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine* **17,** 1–9 (2019).

53. McCradden, M. D., Joshi, S., Mazwi, M. & Anderson, J. A. Ethical limitations of algorithmic fairness solutions in health care machine learning. *The Lancet Digital Health* **2,** e221–e223 (2020).

54. Gaonkar, B., Cook, K. & Macyszyn, L. Ethical issues arising due to bias in training AI algorithms in healthcare and data sharing as a potential solution. *The AI Ethics Journal* **1** (2020).

55. Simundic, A.-M. Bias in research. *Biochemia Medica* **23,** 12–15 (2013).

56. Smith, J. & Noble, H. Bias in research. *Evidence-Based Nursing* **17,** 100–101 (2014).

57. Chen, I. Y., Szolovits, P. & Ghassemi, M. Can AI help reduce disparities in general medical and mental health care? *AMA Journal of Ethics* **21,** 167–179 (2019).

58. Oh, S. S. *et al.* Diversity in clinical and biomedical research: a promise yet to be fulfilled. *PLoS Medicine* **12,** e1001918 (2015).

59. Althubaiti, A. Information bias in health research: definition, pitfalls, and adjustment methods. *Journal of Multidisciplinary Healthcare,* 211–217 (2016).

60. Haneuse, S. Distinguishing selection bias and confounding bias in comparative effectiveness research. *Medical Care* **54,** e23–e29 (2016).

61. Skelly, A. C., Dettori, J. R. & Brodt, E. D. Assessing bias: the importance of considering confounding. *Evidence-Based Spine-Care Journal* **3,** 9–12 (2012).

62. McKinlay, J. B. Some contributions from the social system to gender inequalities in heart disease. *Journal of Health and Social Behavior,* 1–26 (1996).

63. Clerc Liaudat, C. *et al.* Sex/gender bias in the management of chest pain in ambulatory care. *Women's Health* **14,** 1745506518805641 (2018).

64. Arber, S. *et al.* Patient characteristics and inequalities in doctors' diagnostic and management strategies relating to CHD: a video-simulation experiment. *Social Science & Medicine* **62,** 103–115 (2006).

65. Green, A. R. *et al.* Implicit bias among physicians and its prediction of thrombolysis decisions for black and white patients. *Journal of general internal medicine* **22,** 1231–1238 (2007).

66. Lee, P. *et al.* Racial and ethnic disparities in the management of acute pain in US emergency departments: meta-analysis and systematic review. *The American Journal of Emergency Medicine* **37,** 1770–1777 (2019).

67. Ali, M., Salehnejad, R. & Mansur, M. Hospital heterogeneity: what drives the quality of health care. *The European Journal of Health Economics* **19,** 385–408 (2018).

68. Alston, L., Peterson, K. L., Jacobs, J. P., Allender, S. & Nichols, M. Quantifying the role of modifiable risk factors in the differences in cardiovascular disease mortality rates between metropolitan and rural populations in Australia: a macrosimulation modelling study. *BMJ Open* **7,** e018307 (2017).

69. Bradley, E. H. *et al.* Variation in hospital mortality rates for patients with acute myocardial infarction. *The American Journal of Cardiology* **106,** 1108–1112 (2010).

70. Dong, E. *et al.* Differences in regional distribution and inequality in health-resource allocation at hospital and primary health centre levels: a longitudinal study in Shanghai, China. *BMJ Open* **10,** e035635 (2020).

71. Seyyed-Kalantari, L., Zhang, H., McDermott, M. B., Chen, I. Y. & Ghassemi, M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine* **27,** 2176–2182 (2021).

72. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* **54,** 1–35 (2021).

73. Krasanakis, E., Spyromitros-Xioufis, E., Papadopoulos, S. & Kompatsiaris, Y. *Adaptive sensitive reweighting to mitigate bias in fairness-aware classification* in *Proceedings of the 2018 World Wide Web Conference* (2018), 853–862.

74. Angwin, J., Larson, J., Kirchner, L. & Mattu, S. *Machine bias* Last accessed Feb. 8, 2024. May 2016. `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`.

75. Tolan, S., Miron, M., Gómez, E. & Castillo, C. *Why machine learning may lead to unfairness: Evidence from risk assessment for juvenile justice in catalonia* in *Proceedings of the 17th International Conference on Artificial Intelligence and Law* (2019), 83–92.

76. Zhang, B. H., Lemoine, B. & Mitchell, M. *Mitigating unwanted biases with adversarial learning* in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (2018), 335–340.

77. Dwork, C., Hardt, M., Pitassi, T., Reingold, O. & Zemel, R. *Fairness through awareness* in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (2012), 214–226.

78. Hardt, M., Price, E. & Srebro, N. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems* **29** (2016).

79. Beutel, A. *et al. Putting fairness principles into practice: Challenges, metrics, and improvements* in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (2019), 453–459.

80. Hort, M., Chen, Z., Zhang, J. M., Sarro, F. & Harman, M. Bia mitigation for machine learning classifiers: A comprehensive survey. *arXiv preprint arXiv:2207.07068* (2022).

81. Chen, Z., Zhang, J. M., Sarro, F. & Harman, M. A Comprehensive Empirical Study of Bias Mitigation Methods for Machine Learning Classifiers. *ACM Transactions on Software Engineering and Methodology* **32,** 1–30 (2023).

82. Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C. & Venkatasubramanian, S. *Certifying and removing disparate impact* in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2015), 259–268.

83. Zemel, R., Wu, Y., Swersky, K., Pitassi, T. & Dwork, C. *Learning fair representations* in *International Conference on Machine Learning* (2013), 325–333.

84. Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* **5,** 153–163 (2017).

85. Manrai, A. K. *et al.* Genetic misdiagnoses and the potential for health disparities. *New England Journal of Medicine* **375,** 655–665 (2016).

86. Pessach, D. & Shmueli, E. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)* **55,** 1–44 (2022).

87. Dhabliya, D. *et al. Addressing Bias in Machine Learning Algorithms: Promoting Fairness and Ethical Design* in *E3S Web of Conferences* **491** (2024), 02040.

88. Hasanzadeh, F. *et al.* Bias recognition and mitigation strategies in artificial intelligence healthcare applications. *NPJ Digital Medicine* **8,** 154 (2025).

89. Siddique, S. *et al.* Survey on machine learning biases and mitigation techniques. *Digital* **4,** 1–68 (2023).

90. Chheda, K. J., Beckel, J. L. & Gardner, D. M. When equal isn't equal: Contrasting equity and equality perspectives in supporting female professors. *Industrial and Organizational Psychology* **16,** 248–251 (2023).

91. Venkateswaran, N. *et al.* Bringing an Equity-Centered Framework to Research: Transforming the Researcher, Research Content, and Practice of Research. Occasional Paper. RTI Press Publication OP-0085-2301. *RTI International* (2023).

92. Mohottige, D., Olabisi, O. & Boulware, L. E. Use of race in kidney function estimation: lessons learned and the path toward health justice. *Annual review of medicine* **74,** 385–400 (2023).

93. Williams, P. Retaining race in chronic kidney disease diagnosis and treatment. *Cureus* **15** (2023).

94. Pagano, T. P. *et al.* Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big Data and Cognitive Computing* **7,** 15 (2023).

95. Friedler, S. A. *et al. A comparative study of fairness-enhancing interventions in machine learning* in *Proceedings of the Conference on Fairness, Accountability, and Transparency* (2019), 329–338.

96. Calders, T., Kamiran, F. & Pechenizkiy, M. *Building classifiers with independency constraints* in *2009 IEEE International Conference on Data Mining Workshops* (2009), 13–18.

97. Iosifidis, V. & Ntoutsi, E. Dealing with bias via data augmentation in supervised learning scenarios. *Jo Bates Paul D. Clough Robert Jäschke* **24** (2018).

98. Wang, H., Ustun, B., Calmon, F. P. & Harvard, S. *Avoiding disparate impact with counterfactual distributions* in *NeurIPS Workshop on Ethical, Social and Governance Issues in AI* (2018).

99. Hajian, S. & Domingo-Ferrer, J. A methodology for direct and indirect discrimination prevention in data mining. *IEEE Transactions on Knowledge and Data Engineering* **25,** 1445–1459 (2012).

100. Jung, C. *et al.* An algorithmic framework for fairness elicitation. *arXiv preprint arXiv:1905.10660* (2019).

101. Wadsworth, C., Vera, F. & Piech, C. Achieving fairness through adversarial learning: an application to recidivism prediction. *arXiv preprint arXiv:1807.00199* (2018).

102. Beutel, A., Chen, J., Zhao, Z. & Chi, E. H. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075* (2017).

103. Kim, M., Reingold, O. & Rothblum, G. Fairness through computationally-bounded awareness. *Advances in Neural Information Processing Systems* **31** (2018).

104. Mehrabi, N., Gupta, U., Morstatter, F., Steeg, G. V. & Galstyan, A. Attributing fair decisions with attention interventions. *arXiv preprint arXiv:2109.03952* (2021).

105. Du, M. *et al.* Fairness via representation neutralization. *Advances in Neural Information Processing Systems* **34,** 12091–12103 (2021).

106. Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J. & Weinberger, K. Q. On fairness and calibration. *Advances in Neural Information Processing Systems* **30** (2017).

107. Kamiran, F., Karim, A. & Zhang, X. *Decision theory for discrimination-aware classification* in *2012 IEEE 12th International Conference on Data Mining* (2012), 924–929.

108. Menon, A. K. & Williamson, R. C. *The cost of fairness in binary classification* in *Conference on Fairness, Accountability and Transparency* (2018), 107–118.

109. Lohia, P. K. *et al. Bias mitigation post-processing for individual and group fairness* in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2019), 2847–2851.

110. Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V. & Kalai, A. T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems* **29** (2016).

111. Caliskan, A., Bryson, J. J. & Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science* **356,** 183–186 (2017).

112. Buolamwini, J. & Gebru, T. *Gender shades: Intersectional accuracy disparities in commercial gender classification* in *Conference on fairness, accountability and transparency* (2018), 77–91.

113. Krishnan, A. & Rattani, A. A novel approach for bias mitigation of gender classification algorithms using consistency regularization. *Image and Vision Computing* **137,** 104793 (2023).

114. Kamiran, F. & Calders, T. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* **33,** 1–33 (2012).

115. Berk, R. *et al.* A convex framework for fair regression. *arXiv preprint arXiv:1706.02409* (2017).

116. Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G. & Chin, M. H. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine* **169,** 866–872 (2018).

117. Pfohl, S. R., Duan, T., Ding, D. Y. & Shah, N. H. *Counterfactual reasoning for fair clinical risk prediction* in *Machine Learning for Healthcare Conference* (2019), 325–358.

118. Zech, J. R. *et al.* Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine* **15,** e1002683 (2018).

119. Seyyed-Kalantari, L., Liu, G., McDermott, M., Chen, I. Y. & Ghassemi, M. *CheXclusion: Fairness gaps in deep chest X-ray classifiers* in *BIOCOMPUTING 2021: proceedings of the Pacific symposium* (2020), 232–243.

120. Okwor, I. A. *et al.* Digital technologies impact on healthcare delivery: a systematic review of artificial intelligence (AI) and machine-learning (ML) adoption, challenges, and opportunities. *AI* **5,** 1918–1941 (2024).

121. Wong, A. *et al.* External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA internal medicine* **181,** 1065–1070 (2021).

122. Adams, I., Adi-Dako, O., Boafo, E., Ofori, E. K. & Amponsah, S. K. Recent Trends and Possible Future Trajectory of COVID-19. *Rising Contagious Diseases: Basics, Management, and Treatments,* 7–19 (2024).

123. Sreepadmanabh, M., Sahu, A. K. & Chande, A. COVID-19: Advances in diagnostic tools, treatment strategies, and vaccine development. *Journal of Biosciences* **45,** 1–20 (2020).

124. Basu, A., Banerjee, S., Samanta, A., Chowdhury, R. & Panda, S. in *Computational Approaches for Novel Therapeutic and Diagnostic Designing to Mitigate SARS-CoV-2 Infection* 97–114 (Elsevier, 2022).

125. Bchetnia, M., Girard, C., Duchaine, C. & Laprise, C. The outbreak of the novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2): A review of the current global status. *Journal of Infection and Public Health* **13,** 1601–1610 (2020).

126. Mehta, O. P., Bhandari, P., Raut, A., Kacimi, S. E. O. & Huy, N. T. Coronavirus disease (COVID-19): comprehensive review of clinical presentation. *Frontiers in Public Health* **8,** 582932 (2021).

127. Sharma, A., Tiwari, S., Deb, M. K. & Marty, J. L. Severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2): a global pandemic and treatment strategies. *International Journal of Antimicrobial Agents* **56,** 106054 (2020).

128. England, N. *Guidance for emergency departments: initial assessment* Last accessed Feb. 12, 2024. Aug. 2022. `https://www.england.nhs.uk/guidance-for-emergency-departments-initial-assessment/`.

129. England, N. *Same day emergency care* Last accessed Feb. 12, 2024. May 2016. https://www.england.nhs.uk/urgent-emergency-care/same-day-emergency-care/.

130. England, N. *Clinical guide for the management of emergency department patients during the COVID-19 pandemic* Last revised May 2021, Last accessed Feb. 13, 2024. Nov. 2020. https://www.nice.org.uk/media/default/about/covid-19/specialty-guides/management-emergency-department-patients.pdf.

131. England, N. *Standard operating procedure: Lateral flow device testing for emergency department patient pathways* Last accessed Feb. 13, 2024. Dec. 2020. https://www.england.nhs.uk/coronavirus/documents/standard-operating-procedure-lateral-flow-device-testing-for-emergency-department-patient-pathways/.

132. Trust, U. H. P. N. *Ward and department—Coronavirus infectious disease-19 zone status.* Last accessed Feb. 12, 2024. 2020. https://www.plymouthhospitals.nhs.uk/download.cfm?doc=docm93jijm4n8571.pptx&ver=12974.

133. Soltan, A. A. *et al.* Real-world evaluation of rapid and laboratory-free COVID-19 triage for emergency care: external validation and pilot deployment of artificial intelligence driven screening. *The Lancet Digital Health* **4,** e266–e278 (2022).

134. For Health, N. I. & Excellence, C. *COVID-19 rapid guideline: managing COVID-19* Last revised Jan. 2024, Last accessed Feb. 13, 2024. Mar. 2021. https://www.nice.org.uk/guidance/ng191/chapter/2-Assessment.

135. England, N. *Novel coronavirus (COVID-19) standard operating procedure: Testing for inpatients* Last accessed Feb. 13, 2024. Apr. 2022. https://www.england.nhs.uk/coronavirus/wp-content/uploads/sites/52/2022/04/C1624-Novel-coronavirus-COVID-19-standard-operating-procedure-testing-for-inpatients-April-2022.pdf.

136. England, N. *COVID-19 standard operating procedure: testing for elective care pre-admission patient* Last accessed Feb. 13, 2024. https://www.england.nhs.uk/covid-19standard-operating-procedure-testing-for-elective-care-pre-admission-patient/.

137. England, N. & Improvement, N. *Guidance and Standard Operating Procedure COVID-19 Virus Testing in NHS Laboratories* Last accessed Feb. 13, 2024. https://www.leedsth.nhs.uk/assets/COVID-19/eadfc5a62b/Guidance-and-SOP-COVID-19-Testing-NHS-E-Laboratories-final-for-Regional-Labs.pdf.

138. England, N. *NHS England and NHS Improvement rollout of lateral flow devices for asymptomatic staff testing for SARS CoV-2 (phase 2: trusts)* Last accessed Feb. 13, 2024. Nov. 2020. https://www.england.nhs.uk/coronavirus/documents/nhs-england-and-nhs-improvement-rollout-of-lateral-flow-devices-for-asymptomatic-staff-testing-for-sars-cov-2-phase-2-trusts/#Lateral%20flow%20antigen%20testing.

139. Neamah, S. R. Comparison between symptoms of COVID-19 and other respiratory diseases. *Electronic Journal of Medical and Educational Technologies* **13,** em2014 (2020).

140. Shah, S. J. *et al.* Clinical features, diagnostics, and outcomes of patients presenting with acute respiratory illness: a comparison of patients with and without COVID-19. *medRxiv* (2020).

141. Mistry, D. A., Wang, J. Y., Moeser, M.-E., Starkey, T. & Lee, L. Y. A systematic review of the sensitivity and specificity of lateral flow devices in the detection of SARS-CoV-2. *BMC Infectious Diseases* **21,** 1–14 (2021).

142. Eyre, D. W. *et al.* Performance of antigen lateral flow devices in the UK during the alpha, delta, and omicron waves of the SARS-CoV-2 pandemic: a diagnostic and observational study. *The Lancet Infectious Diseases* **23,** 922–932 (2023).

143. Taylor, A., Calvez, R., Atkins, M. & Fink, C. G. Comparing lateral flow testing with a rapid RT-PCR method for SARS-CoV-2 detection in the United Kingdom—A retrospective diagnostic accuracy study. *Health Science Reports* **5,** e811 (2022).

144. Soltan, A. A. *et al.* Rapid triage for COVID-19 using routine clinical data for patients attending hospital: development and prospective validation of an artificial intelligence screening test. *The Lancet Digital Health* **3,** e78–e87 (2021).

145. Panpradist, N. *et al.* Simpler and faster Covid-19 testing: Strategies to streamline SARS-CoV-2 molecular assays. *EBioMedicine* **64** (2021).

146. Kim, C. K. *et al.* An automated COVID-19 triage pipeline using artificial intelligence based on chest radiographs and clinical data. *NPJ Digital Medicine* **5,** 5 (2022).

147. Jang, S. B. *et al.* Deep-learning algorithms for the interpretation of chest radiographs to aid in the triage of COVID-19 patients: A multicenter retrospective study. *PLoS One* **15,** e0242759 (2020).

148. Dev, K., Khowaja, S. A., Bist, A. S., Saini, V. & Bhatia, S. Triage of potential COVID-19 patients from chest X-ray images using hierarchical convolutional networks. *Neural Computing and Applications* **35,** 23861–23876 (2023).

149. Liang, W. *et al.* Early triage of critically ill COVID-19 patients using deep learning. *Nature Communications* **11,** 3543 (2020).

150. Williams, T. C. *et al.* Sensitivity of RT-PCR testing of upper respiratory tract samples for SARS-CoV-2 in hospitalised patients: a retrospective cohort study. *Wellcome Open Research* **5** (2020).

151. Miller, T. E. *et al.* Clinical sensitivity and interpretation of PCR and serological COVID-19 diagnostics for patients presenting to the hospital. *The FASEB Journal* **34,** 13877 (2020).

152. Lewis, M. J. & Jawad, A. S. The effect of ethnicity and genetic ancestry on the epidemiology, clinical features and outcome of systemic lupus erythematosus. *Rheumatology* **56,** i67–i77 (2017).

153. Caprio, S. *et al.* Influence of race, ethnicity, and culture on childhood obesity: implications for prevention and treatment: a consensus statement of Shaping America's Health and the Obesity Society. *Diabetes Care* **31,** 2211 (2008).

154. Xie, H.-G., Kim, R. B., Wood, A. J. & Stein, C. M. Molecular basis of ethnic differences in drug disposition and response. *Annual Review of Pharmacology and Toxicology* **41,** 815–850 (2001).

155. Bamshad, M. Genetic influences on health: does race matter? *Journal of the American Medical Association* **294,** 937–946 (2005).

156. Biener, A. I. & Zuvekas, S. H. Do racial and ethnic disparities in health care use vary with health? *Health Services Research* **54,** 64–74 (2019).

157. Taylor, Y. J., Spencer, M. D., Mahabaleshwarkar, R. & Ludden, T. Racial/ethnic differences in healthcare use among patients with uncontrolled and controlled diabetes. *Ethnicity & Health* **24,** 245–256 (2019).

158. Egede, L. E. Race, ethnicity, culture, and disparities in health care. *Journal of General Internal Medicine* **21,** 667 (2006).

159. Curry Jr, W. T., Carter, B. S., Barker, F. G., *et al.* Racial, ethnic, and socioeconomic disparities in patient outcomes after craniotomy for tumor in adult patients in the United States, 1988–2004. *Neurosurgery* **66,** 427–438 (2010).

160. Ghodeswar, B. & Vaidyanathan, J. Organisational adoption of medical technology in healthcare sector. *Journal of Services Research* **7,** 57 (2007).

161. Krumholz, H. M. Variations in health care, patient preferences, and high-quality decision making. *Journal of the American Medical Association* **310,** 151–152 (2013).

162. Atsma, F., Elwyn, G. & Westert, G. Understanding unwarranted variation in clinical practice: a focus on network effects, reflective medicine and learning health systems. *International Journal for Quality in Health Care* **32,** 271–274 (2020).

163. Lucero, R. J., Lake, E. T. & Aiken, L. H. Variations in nursing care quality across hospitals. *Journal of Advanced Nursing* **65,** 2299–2310 (2009).

164. Kearnes, S. Pursuing a prospective perspective. *Trends in chemistry* **3,** 77–79 (2021).

165. LeCun, Y. A., Bottou, L., Orr, G. B. & Müller, K.-R. Efficient backProp BT-neural networks: Tricks of the trade. *Neural Networks: Tricks of the Trade* (2012).

166. Laurent, C., Pereyra, G., Brakel, P., Zhang, Y. & Bengio, Y. *Batch normalized recurrent neural networks* in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2016), 2657–2661.

167. Hosseinzadeh, M. *et al.* A multiple multilayer perceptron neural network with an adaptive learning algorithm for thyroid disease diagnosis in the internet of medical things. *The Journal of Supercomputing* **77,** 3616–3637 (2021).

168. Yang, J., Soltan, A. A., Eyre, D. W. & Clifton, D. A. Algorithmic fairness and bias mitigation for clinical machine learning with deep reinforcement learning. *Nature Machine Intelligence* **5,** 884–894 (2023).

169. Arevalo-Rodriguez, I. *et al.* False-negative results of initial RT-PCR assays for COVID-19: a systematic review. *PloS one* **15,** e0242958 (2020).

170. Woloshin, S., Patel, N. & Kesselheim, A. S. False negative tests for SARS-CoV-2 infection—challenges and implications. *New England Journal of Medicine* **383,** e38 (2020).

171. Zhang, L. A pattern-recognition-based ensemble data imputation framework for sensors from building energy systems. *Sensors* **20,** 5947 (2020).

172. Mohan, K., Pearl, J. & Jin, T. *Missing data as a causal inference problem* in *Proceedings of the Neural Information Processing Systems Conference* (2013).

173. Bang, H. & Robins, J. M. Doubly robust estimation in missing data and causal inference models. *Biometrics* **61,** 962–973 (2005).

174. Dalvi, N., Domingos, P., Mausam, Sanghai, S. & Verma, D. *Adversarial classification* in *Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2004), 99–108.

175. Adel, T., Valera, I., Ghahramani, Z. & Weller, A. *One-network adversarial fairness* in *Proceedings of the AAAI Conference on Artificial Intelligence* **33** (2019), 2412–2420.

176. Delobelle, P. *et al.* Ethical adversaries: Towards mitigating unfairness with adversarial machine learning. *ACM SIGKDD Explorations Newsletter* **23,** 32–41 (2021).

177. Birzhandi, P. & Cho, Y.-S. Application of fairness to healthcare, organizational justice, and finance: a survey. *Expert Systems with Applications* **216,** 119465 (2023).

178. Finocchiaro, J. *et al. Bridging machine learning and mechanism design towards algorithmic fairness* in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2021), 489–503.

179. Muhammad, I. & Yan, Z. SUPERVISED MACHINE LEARNING APPROACHES: A SURVEY. *ICTACT Journal on Soft Computing* **5** (2015).

180. Shetty, S. H., Shetty, S., Singh, C. & Rao, A. Supervised machine learning: algorithms and applications. *Fundamentals and Methods of Machine and Deep Learning: Algorithms, Tools and Applications,* 1–16 (2022).

181. Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A. & Aljaaf, A. J. A systematic review on supervised and unsupervised machine learning algorithms for data science. *Supervised and Unsupervised Learning for Data Science,* 3–21 (2020).

182. Celebi, M. E. & Aydin, K. *Unsupervised learning algorithms* (Springer, 2016).

183. Hastie, T. *et al.* Unsupervised learning. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction,* 485–585 (2009).

184. Wiering, M. A. & Van Otterlo, M. Reinforcement learning. *Adaptation, Learning, and Optimization* **12,** 729 (2012).

185. Kaelbling, L. P., Littman, M. L. & Moore, A. W. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research* **4,** 237–285 (1996).

186. Sutton, R. S. & Barto, A. G. *Reinforcement learning: An introduction* (MIT press, 2018).

187. Jung, C. *et al. An Algorithmic Framework for Fairness Elicitation* in *2nd Symposium on Foundations of Responsible Computing* **31** (2021), 21.

188. Oneto, L., Doninini, M., Elders, A. & Pontil, M. *Taking advantage of multitask learning for fair classification* in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (2019), 227–237.

189. Calders, T. & Verwer, S. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* **21,** 277–292 (2010).

190. Chen, Z., Zhang, J. M., Sarro, F. & Harman, M. *MAAT: a novel ensemble approach to addressing fairness and performance bugs for machine learning software* in *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (2022), 1122–1134.

191. Mishler, A. & Kennedy, E. H. *FADE: FAir Double Ensemble Learning for Observable and Counterfactual Outcomes* in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (2022), 1053–1053.

192. Romano, Y., Bates, S. & Candes, E. Achieving equalized odds by resampling sensitive attributes. *Advances in Neural Information Processing Systems* **33,** 361–371 (2020).

193. Kamishima, T., Akaho, S., Asoh, H. & Sakuma, J. *Fairness-aware classifier with prejudice remover regularizer* in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23* (2012), 35–50.

194. Liu, W. *et al.* Fair differential privacy can mitigate the disparate impact on model accuracy (2020).

195. Pérez-Suay, A. *et al. Fair kernel learning* in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (2017), 339–355.

196. Gretton, A., Bousquet, O., Smola, A. & Schölkopf, B. *Measuring statistical dependence with Hilbert-Schmidt norms* in *International Conference on Algorithmic Learning Theory* (2005), 63–77.

197. Komiyama, J., Takeda, A., Honda, J. & Shimao, H. *Nonconvex optimization for regression with fairness constraints* in *International Conference on Machine Learning* (2018), 2737–2746.

198. Jiang, H. & Nachum, O. *Identifying and correcting label bias in machine learning* in *International Conference on Artificial Intelligence and Statistics* (2020), 702–712.

199. Li, P. & Liu, H. *Achieving fairness at no utility cost via data reweighing with influence* in *International Conference on Machine Learning* (2022), 12917–12930.

200. Ashokan, A. & Haas, C. Fairness metrics and bias mitigation strategies for rating predictions. *Information Processing & Management* **58,** 102646 (2021).

201. Foulds, J. R. & Pan, S. Are Parity-Based Notions of {AI} Fairness Desirable? *A Quarterly Bulletin of the Computer Society of the IEEE Technical Committee on Data Engineering* **43** (2020).

202. Yang, Y., Zhang, C., Fan, C., Mostafavi, A. & Hu, X. Towards fairness-aware disaster informatics: an interdisciplinary perspective. *IEEE Access* **8,** 201040–201054 (2020).

203. Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J. & Wallach, H. *A reductions approach to fair classification* in *International Conference on Machine Learning* (2018), 60–69.

204. Verma, S. & Rubin, J. *Fairness definitions explained* in *Proceedings of the International Workshop on Software Fairness* (2018), 1–7.

205. Gajane, P. & Pechenizkiy, M. On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184* (2017).

206. Zafar, M. B., Valera, I., Gomez Rodriguez, M. & Gummadi, K. P. *Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment* in *Proceedings of the 26th International Conference on World Wide Web* (2017), 1171–1180.

207. Corbett-Davies, S., Pierson, E., Feller, A., Goel, S. & Huq, A. *Algorithmic decision making and the cost of fairness* in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2017), 797–806.

208. Raftopoulos, G., Davrazos, G. & Kotsiantis, S. Evaluating fairness strategies in educational data mining: A comparative study of bias mitigation techniques. *Electronics* **14,** 1856 (2025).

209. Kecki, V. & Said, A. *Understanding fairness in recommender systems: a healthcare perspective* in *Proceedings of the 18th ACM Conference on Recommender Systems* (2024), 1125–1130.

210. Van der Meijden, S. *et al.* Navigating Fairness in AI-based Prediction Models: Theoretical Constructs and Practical Applications. *medRxiv,* 2025–03 (2025).

211. Zehlike, M., Loosley, A., Jonsson, H., Wiedemann, E. & Hacker, P. Beyond incompatibility: Trade-offs between mutually exclusive fairness criteria in machine learning and law. *Artificial Intelligence* **340,** 104280 (2025).

212. Ni, H., Han, L., Chen, T., Sadiq, S. & Demartini, G. *Fairness without sensitive attributes via knowledge sharing* in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (2024), 1897–1906.

213. Chen, J., Kallus, N., Mao, X., Svacha, G. & Udell, M. *Fairness under unawareness: Assessing disparity when protected class is unobserved* in *Proceedings of the conference on fairness, accountability, and transparency* (2019), 339–348.

214. Anderson, J. W. & Visweswaran, S. Algorithmic individual fairness and healthcare: a scoping review. *JAMIA open* **8,** ooae149 (2025).

215. Kusner, M. J., Loftus, J., Russell, C. & Silva, R. Counterfactual fairness. *Advances in neural information processing systems* **30** (2017).

216. Kuppler, M., Kern, C., Bach, R. L. & Kreuter, F. Distributive justice and fairness metrics in automated decision-making: How much overlap is there? *arXiv preprint arXiv:2105.01441* (2021).

217. Chen, T. & Guestrin, C. *Xgboost: A scalable tree boosting system* in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016), 785–794.

218. Ramraj, S., Uzir, N., Sunil, R. & Banerjee, S. Experimenting XGBoost algorithm for prediction and classification of different datasets. *International Journal of Control Theory and Applications* **9,** 651–662 (2016).

219. Liu, Z. *et al. How do adam and training strategies help bnns optimization* in *International Conference on Machine Learning* (2021), 6936–6946.

220. Iscen, A., Valmadre, J., Arnab, A. & Schmid, C. *Learning with neighbor consistency for noisy labels* in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), 4672–4681.

221. Li, Y. & Vasconcelos, N. *Repair: Removing representation bias by dataset resampling* in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019), 9572–9581.

222. Freeman, E. A. & Moisen, G. G. A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. *Ecological modelling* **217,** 48–58 (2008).

223. Maity, S., Mukherjee, D., Yurochkin, M. & Sun, Y. There is no trade-off: enforcing fairness can improve accuracy (2020).

224. Li, X., Wu, P. & Su, J. *Accurate fairness: Improving individual fairness without trading accuracy* in *Proceedings of the AAAI Conference on Artificial Intelligence* **37** (2023), 14312–14320.

225. Li, Y. Deep reinforcement learning: An overview. *arXiv preprint arXiv:1701.07274* (2017).

226. Wiering, M. A., Van Hasselt, H., Pietersma, A.-D. & Schomaker, L. *Reinforcement learning algorithms for solving classification problems* in *2011 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)* (2011), 91–96.

227. Lin, E., Chen, Q. & Qi, X. Deep reinforcement learning for imbalanced classification. *Applied Intelligence* **50,** 2488–2502 (2020).

228. Watkins, C. J. & Dayan, P. Q-learning. *Machine Learning* **8,** 279–292 (1992).

229. Sutton, R. S. & Barto, A. G. The reinforcement learning problem. *Reinforcement Learning: An Introduction,* 51–85 (1998).

230. Mnih, V. *et al.* Human-level control through deep reinforcement learning. *Nature* **518,** 529–533 (2015).

231. Almasan, P., Suárez-Varela, J., Rusek, K., Barlet-Ros, P. & Cabellos-Aparicio, A. Deep reinforcement learning meets graph neural networks: Exploring a routing optimization use case. *Computer Communications* **196,** 184–194 (2022).

232. Shen, Y., Zhao, N., Xia, M. & Du, X. A deep q-learning network for ship stowage planning problem. *Polish Maritime Research* **24,** 102–109 (2017).

233. Wang, Z. *et al. Dueling network architectures for deep reinforcement learning* in *International Conference on Machine Learning* (2016), 1995–2003.

234. Yang, J. *et al.* Deep reinforcement learning for multi-class imbalanced training: applications in healthcare. *Machine Learning,* 1–20 (2023).

235. Thrun, S. & Schwartz, A. *Issues in using function approximation for reinforcement learning* in *Proceedings of the 1993 Connectionist Models Summer School* (2014), 255–263.

236. Van Hasselt, H., Guez, A. & Silver, D. *Deep reinforcement learning with double q-learning* in *Proceedings of the AAAI conference on artificial intelligence* **30** (2016).

237. Sui, Z., Pu, Z., Yi, J. & Tan, X. *Path planning of multiagent constrained formation through deep reinforcement learning* in *2018 International Joint Conference on Neural Networks (IJCNN)* (2018), 1–8.

238. Labrique, A. B. *et al.* Best practices in scaling digital health in low and middle income countries. *Globalization and Health* **14,** 1–8 (2018).

239. Wang, D. *et al. "Brilliant AI doctor" in rural clinics: Challenges in AI-powered clinical decision support system deployment* in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021), 1–18.

240. Ciecierski-Holmes, T., Singh, R., Axt, M., Brenner, S. & Barteit, S. Artificial intelligence for strengthening healthcare systems in low-and middle-income countries: a systematic scoping review. *NPJ Digital Medicine* **5,** 162 (2022).

241. Alami, H. *et al.* Artificial intelligence in health care: laying the Foundation for Responsible, sustainable, and inclusive innovation in low-and middle-income countries. *Globalization and Health* **16,** 1–6 (2020).

242. Schwalbe, N. & Wahl, B. Artificial intelligence and the future of global health. *The Lancet* **395,** 1579–1586 (2020).

243. Zhou, N. *et al.* Concordance study between IBM Watson for oncology and clinical practice for patients with cancer in China. *The Oncologist* **24,** 812–819 (2019).

244. Ndabarora, E., Chipps, J. A. & Uys, L. Systematic review of health data quality management and best practices at community and district levels in LMIC. *Information Development* **30,** 103–120 (2014).

245. Odhiambo-Otieno, G. W. Evaluation of existing district health management information systems: a case study of the district health systems in Kenya. *International Journal of Medical Informatics* **74,** 733–744 (2005).

246. Mate, K. S., Bennett, B., Mphatswe, W., Barker, P. & Rollins, N. Challenges for routine health system data management in a large public programme to prevent mother-to-child HIV transmission in South Africa. *PloS One* **4,** e5483 (2009).

247. Bosch-Capblanch, X., Ronveaux, O., Doyle, V., Remedios, V. & Bchir, A. Accuracy and quality of immunization information systems in forty-one low income countries. *Tropical Medicine & International Health* **14,** 2–10 (2009).

248. Heunis, C. *et al.* Accuracy of tuberculosis routine data and nurses' views of the TB-HIV information system in the free state, South Africa. *Journal of the Association of Nurses in AIDS Care* **22,** 67–73 (2011).

249. Harper, S., Edge, V., Schuster-Wallace, C., Ar-Rushdi, M. & McEwen, S. Improving Aboriginal health data capture: evidence from a health registry evaluation. *Epidemiology & Infection* **139,** 1774–1783 (2011).

250. Abdul-Rahman, T. *et al.* Inaccessibility and low maintenance of medical data archive in low-middle income countries: Mystery behind public health statistics and measures. *Journal of Infection and Public Health* **16,** 1556–1561 (2023).

251. Akhlaq, A., McKinstry, B., Muhammad, K. B. & Sheikh, A. Barriers and facilitators to health information exchange in low-and middle-income country settings: a systematic review. *Health Policy and Planning* **31,** 1310–1325 (2016).

252. Oduoye, M. O. *et al.* Impacts of the advancement in artificial intelligence on laboratory medicine in low-and middle-income countries: Challenges and recommendations—A literature review. *Health Science Reports* **7,** e1794 (2024).

253. López, D. M., Rico-Olarte, C., Blobel, B. & Hullin, C. Challenges and solutions for transforming health ecosystems in low-and middle-income countries through artificial intelligence. *Frontiers in Medicine* **9,** 958097 (2022).

254. Wahl, B., Cossy-Gantner, A., Germann, S. & Schwalbe, N. R. Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings? *BMJ Global Health* **3,** e000798 (2018).

255. Ngwa, W., Olver, I. & Schmeler, K. M. The use of health-related technology to reduce the gap between developed and undeveloped regions around the globe. *American Society of Clinical Oncology Educational Book* **40,** 227–236 (2020).

256. Beran, D. *et al.* Research capacity building—obligations for global health partners. *The Lancet Global Health* **5,** e567–e568 (2017).

257. Rickard, J., Ntirenganya, F., Ntakiyiruta, G. & Chu, K. Global health in the 21st century: equity in surgical training partnerships. *Journal of Surgical Education* **76,** 9–13 (2019).

258. Baumgartner, R. *et al.* Fair and equitable AI in biomedical research and healthcare: Social science perspectives. *Artificial Intelligence in Medicine* **144,** 102658 (2023).

259. Whittlestone, J., Nyrup, R., Alexandrova, A., Dihal, K. & Cave, S. Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research. *London: Nuffield Foundation* (2019).

260. Bonmatí, L. M. *et al.* CHAIMELEON project: creation of a pan-European Repository of health imaging data for the development of AI-powered cancer management tools. *Frontiers in Oncology* **12,** 742701 (2022).

261. Durieux, M. E. & Naik, B. I. Scientia potentia est: striving for data equity in clinical medicine for low-and middle-income countries. *Anesthesia & Analgesia* **135,** 209–212 (2022).

262. Souza, R., Stanley, E. A. & Forkert, N. D. *On the Relationship Between Open Science in Artificial Intelligence for Medical Imaging and Global Health Equity* in *Workshop on Clinical Image-Based Procedures* (2023), 289–300.

263. Luo, Y., Jin, H. & Li, P. *A blockchain future for secure clinical data sharing: A position paper* in *Proceedings of the ACM international workshop on security in software defined networks & network function virtualization* (2019), 23–27.

264. For Genomics, G. A. & Health*. A federated ecosystem for sharing genomic, clinical data. *Science* **352,** 1278–1280 (2016).

265. Beutler, E. & Waalen, J. The definition of anemia: what is the lower limit of normal of the blood hemoglobin concentration? *Blood* **107,** 1747–1750 (2006).

266. Thomas, C. & Lumb, A. B. Physiology of haemoglobin. *Continuing Education in Anaesthesia, Critical Care & Pain* **12,** 251–256 (2012).

267. Maidstone & Trust, T. W. N. *Reference Ranges (RWF-BS-Haem-LI34 Revision 2.0)* Last accessed April 2, 2024. 2020. `https://www.mtw.nhs.uk/wp-content/uploads/2020/11/Haematology-reference-ranges.pdf`.

268. Smith, A. & Milnthorpe, J. Haematological malignancies: a guide to novel therapies. *Prescriber* **31,** 9–15 (2020).

269. Smiti, A. A critical overview of outlier detection methods. *Computer Science Review* **38,** 100306 (2020).

270. Tropsha, A. Best practices for QSAR model development, validation, and exploitation. *Molecular Informatics* **29,** 476–488 (2010).

271. Van Calster, B., Steyerberg, E. W., Wynants, L. & Van Smeden, M. There is no such thing as a validated prediction model. *BMC Medicine* **21,** 70 (2023).

272. Futoma, J., Simons, M., Panch, T., Doshi-Velez, F. & Celi, L. A. The myth of generalisability in clinical research and machine learning in health care. *The Lancet Digital Health* **2,** e489–e492 (2020).

273. Yang, J. *et al.* Mitigating machine learning bias between high income and low–middle income countries for enhanced model fairness and generalizability. *Scientific Reports* **14,** 13318 (2024).

274. Nong, P., Hamasha, R., Singh, K., Adler-Milstein, J. & Platt, J. *How academic medical centers govern AI prediction tools in the context of uncertainty and evolving regulation* 2024.

275. Apa, H. *et al.* Clinical accuracy of tympanic thermometer and noncontact infrared skin thermometer in pediatric practice: an alternative for axillary digital thermometer. *Pediatric Emergency Care* **29,** 992–997 (2013).

276. Berk, R., Heidari, H., Jabbari, S., Kearns, M. & Roth, A. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* **50,** 3–44 (2021).

277. Nelson, A. Unequal treatment: confronting racial and ethnic disparities in health care. *Journal of the National Medical Association* **94,** 666 (2002).

278. Maina, I. W., Belton, T. D., Ginzberg, S., Singh, A. & Johnson, T. J. A decade of studying implicit racial/ethnic bias in healthcare providers using the implicit association test. *Social Science & Medicine* **199,** 219–229 (2018).

279. Van Ryn, M. & Burke, J. The effect of patient race and socio-economic status on physicians' perceptions of patients. *Social Science & Medicine* **50,** 813–828 (2000).

280. Johnson, R. L., Saha, S., Arbelaez, J. J., Beach, M. C. & Cooper, L. A. Racial and ethnic differences in patient perceptions of bias and cultural competence in health care. *Journal of General Internal Medicine* **19,** 101–110 (2004).

281. Of Health, U. D., Services, H., *et al.* National Health Care Quality and Disparities Report and 5th Anniversary Update on the National Quality Strategy. *Agency for Health Care Research and Quality* (2015).

282. Ho, S. Y., Phua, K., Wong, L. & Goh, W. W. B. Extensions of the external validation for checking learned model interpretability and generalizability. *Patterns* **1** (2020).

283. Rush, B., Celi, L. A. & Stone, D. J. Applying machine learning to continuously monitored physiological data. *Journal of Clinical Monitoring and Computing* **33,** 887–893 (2019).

284. Johnson, A. E. *et al.* Machine learning and decision support in critical care. *Proceedings of the IEEE* **104,** 444–466 (2016).

285. Chaudhry, F. *et al.* Machine learning applications in the neuro ICU: a solution to big data mayhem? *Frontiers in Neurology* **11,** 554633 (2020).

286. Feldman, K., Faust, L., Wu, X., Huang, C. & Chawla, N. V. *Beyond volume: The impact of complex healthcare data on the machine learning pipeline* in *Towards Integrative Machine Learning and Knowledge Extraction: BIRS Workshop, Banff, AB, Canada, July 24-26, 2015, Revised Selected Papers* (2017), 150–169.

287. Hassan, S., Dhali, M., Zaman, F. & Tanveer, M. Big data and predictive analytics in healthcare in Bangladesh: regulatory challenges. *Heliyon* **7** (2021).

288. Ruiz, V. M. *et al.* Early prediction of clinical deterioration using data-driven machine-learning modeling of electronic health records. *The Journal of Thoracic and Cardiovascular Surgery* **164,** 211–222 (2022).

289. Kaji, D. A. *et al.* An attention based deep learning model of clinical events in the intensive care unit. *PloS One* **14,** e0211057 (2019).

290. Calvert, J. S. *et al.* A computational approach to early sepsis detection. *Computers in Biology and Medicine* **74,** 69–73 (2016).

291. Wong, A.-K. I., Cheung, P. C., Kamaleswaran, R., Martin, G. S. & Holder, A. L. Machine learning methods to predict acute respiratory failure and acute respiratory distress syndrome. *Frontiers in Big Data* **3,** 579774 (2020).

292. Mehta, R. L. *et al.* Refining predictive models in critically ill patients with acute renal failure. *Journal of the American Society of Nephrology* **13,** 1350–1357 (2002).

293. Mohammed, A. *et al.* Using machine learning to predict early onset acute organ failure in critically ill intensive care unit patients with sickle cell disease: retrospective study. *Journal of Medical Internet Research* **22,** e14693 (2020).

294. Iwase, S. *et al.* Prediction algorithm for ICU mortality and length of stay using machine learning. *Scientific Reports* **12,** 12912 (2022).

295. Dan, T. *et al. Machine learning to predict ICU admission, ICU mortality and survivors' length of stay among COVID-19 patients: toward optimal allocation of ICU resources* in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (2020), 555–561.

296. Kim, S., Kim, W. & Park, R. W. A comparison of intensive care unit mortality prediction models through the use of data mining techniques. *Healthcare Informatics Research* **17,** 232 (2011).

297. Bhattacharya, S., Rajan, V. & Shrivastava, H. *ICU mortality prediction: a classification algorithm for imbalanced datasets* in *Proceedings of the AAAI Conference on Artificial Intelligence* **31** (2017).

298. Johnson, A. E. & Mark, R. G. *Real-time mortality prediction in the Intensive Care Unit* in *AMIA Annual Symposium Proceedings* **2017** (2017), 994.

299. Karbouband, K. & Tabaa, M. Bed Allocation Optimization Based on Survival Analysis, Treatment Trajectory and Costs Estimations. *IEEE Access* **11,** 31699–31715 (2023).

300. Schiele, J., Koperna, T. & Brunner, J. O. Predicting intensive care unit bed occupancy for integrated operating room scheduling via neural networks. *Naval Research Logistics (NRL)* **68,** 65–88 (2021).

301. Calvert, J. *et al.* Cost and mortality impact of an algorithm-driven sepsis prediction system. *Journal of Medical Economics* **20,** 646–651 (2017).

302. Badawi, O. & Breslow, M. J. Readmissions and death after ICU discharge: development and validation of two predictive models. *PloS One* **7,** e48758 (2012).

303. Ouanes, I. *et al.* A model to predict short-term death or readmission after intensive care unit discharge. *Journal of Critical Care* **27,** 422–e1 (2012).

304. Sharma, A., Shukla, A., Tiwari, R. & Mishra, A. *Mortality Prediction of ICU patients using Machine Leaning: A survey* in *Proceedings of the International Conference on Compute and Data Analysis* (2017), 49–53.

305. Goldberger, A. L. *et al.* PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* **101,** e215–e220 (2000).

306. Pollard, T. J. *et al.* The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Scientific Data* **5,** 1–13 (2018).

307. Sheikhalishahi, S., Balaraman, V. & Osmani, V. Benchmarking machine learning models on multi-centre eICU critical care dataset. *PloS One* **15,** e0235424 (2020).

308. Henson, V. & Vickery, D. Patient self discharge from the emergency department: who is at risk? *Emergency Medicine Journal* **22,** 499–501 (2005).

309. Alfandre, D. J. *"I'm going home": discharges against medical advice* in *Mayo Clinic Proceedings* **84** (2009), 255–260.

310. Ashrafi, E. *et al.* Discharge against medical advice (DAMA): Causes and predictors. *Electronic Physician* **9,** 4563 (2017).

311. Blazer, D. G. & Hernandez, L. M. Genes, behavior, and the social environment: Moving beyond the nature/nurture debate (2006).

312. Marmot, M. Social determinants of health inequalities. *The Lancet* **365,** 1099–1104 (2005).

313. Paulus, J. K. & Kent, D. M. Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. *NPJ Digital Medicine* **3,** 99 (2020).

314. Islam, R., Pan, S. & Foulds, J. R. *Can we obtain fairness for free?* in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (2021), 586–596.

315. Chen, I., Johansson, F. D. & Sontag, D. Why is my classifier discriminatory? *Advances in Neural Information Processing Systems* **31** (2018).

316. Zhao, H. & Gordon, G. J. Inherent tradeoffs in learning fair representations. *Journal of Machine Learning Research* **23,** 1–26 (2022).

317. Zhou, H. *et al.* Multimodal Data Integration for Precision Oncology: Challenges and Future Directions. *arXiv preprint arXiv:2406.19611* (2024).

318. Steyaert, S. *et al.* Multimodal data fusion for cancer biomarker discovery with deep learning. *Nature Machine Intelligence* **5,** 351–362 (2023).

319. Schneider, L. *et al.* Integration of deep learning-based image analysis and genomic data in cancer pathology: A systematic review. *European Journal of Cancer* **160,** 80–91 (2022).

320. Mazzaschi, G. *et al.* Integrated MRI–Immune–Genomic Features Enclose a Risk Stratification Model in Patients Affected by Glioblastoma. *Cancers* **14,** 3249 (2022).

321. Boehm, K. M. *et al.* Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer. *Nature Cancer* **3,** 723–733 (2022).

322. Lobato-Delgado, B., Priego-Torres, B. & Sanchez-Morillo, D. Combining molecular, imaging, and clinical data analysis for predicting cancer prognosis. *Cancers* **14,** 3215 (2022).

323. Sheng, J. *et al.* Predictive classification of Alzheimer's disease using brain imaging and genetic data. *Scientific Reports* **12,** 2405 (2022).

324. Wu, J. *et al.* Integrating Transcriptomics, Genomics, and Imaging in Alzheimer's Disease: A Federated Model. *Frontiers in Radiology* **1,** 777030 (2022).

325. Adewale, Q., Khan, A. F., Carbonell, F., Iturria-Medina, Y. & Initiative, A. D. N. Integrated transcriptomic and neuroimaging brain model decodes biological mechanisms in aging and Alzheimer's disease. *Elife* **10,** e62589 (2021).

326. Ramírez, J. *et al.* Sudden cardiac death and pump failure death prediction in chronic heart failure by combining ECG and clinical markers in an integrated risk model. *PloS One* **12,** e0186152 (2017).

327. Hemingway, H. *et al.* Big data from electronic health records for early and late translational cardiovascular research: challenges and potential. *European Heart Journal* **39,** 1481–1495 (2018).

328. Soto, J. T. *et al.* Multimodal deep learning enhances diagnostic precision in left ventricular hypertrophy. *European Heart Journal-Digital Health* **3,** 380–389 (2022).

329. Amal, S. *et al.* Use of multi-modal data and machine learning to improve cardiovascular disease care. *Frontiers in Cardiovascular Medicine* **9,** 840262 (2022).

330. Bhattacharya, A. *et al.* Multi-modal fusion model for predicting adverse cardiovascular outcome post percutaneous coronary intervention. *Physiological Measurement* **43,** 124004 (2022).

331. Manure, A. & Bengani, S. in *Introduction to Responsible AI: Implement Ethical AI Using Python* 61–106 (Springer, 2023).

332. Longo, L., Goebel, R., Lecue, F., Kieseberg, P. & Holzinger, A. *Explainable artificial intelligence: Concepts, applications, research challenges and visions* in *International cross-domain conference for machine learning and knowledge extraction* (2020), 1–16.

333. Arrieta, A. B. *et al.* Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* **58,** 82–115 (2020).

334. Linardatos, P., Papastefanopoulos, V. & Kotsiantis, S. Explainable ai: A review of machine learning interpretability methods. *Entropy* **23,** 18 (2020).

335. Samek, W. & Müller, K.-R. Towards explainable artificial intelligence. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning,* 5–22 (2019).

336. Ribeiro, M. T., Singh, S. & Guestrin, C. *" 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), 1135–1144.

337. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems* **30** (2017).

338. Konig, R., Johansson, U. & Niklasson, L. *G-REX: A versatile framework for evolutionary data mining* in *2008 IEEE International Conference on Data Mining Workshops* (2008), 971–974.

339. Robnik-Šikonja, M. & Kononenko, I. Explaining classifications for individual instances. *IEEE Transactions on Knowledge and Data Engineering* **20,** 589–600 (2008).

340. Strumbelj, E. & Kononenko, I. An efficient explanation of individual classifications using game theory. *The Journal of Machine Learning Research* **11,** 1–18 (2010).

341. Quinlan, J. R. Simplifying decision trees. *International Journal of Man-Machine Studies* **27,** 221–234 (1987).

342. Laurent, H. & Rivest, R. L. Constructing optimal binary decision trees is NP-complete. *Information Processing Letters* **5,** 15–17 (1976).

343. Ustun, B. & Rudin, C. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning* **102,** 349–391 (2016).

344. Berg, D. Bankruptcy prediction by generalized additive models. *Applied Stochastic Models in Business and Industry* **23,** 129–143 (2007).

345. Hastie, T. & Tibshirani, R. Generalized additive models: some applications. *Journal of the American Statistical Association* **82,** 371–386 (1987).

346. Taylan, P., Weber, G.-W. & Beck, A. New approaches to regression by generalized additive models and continuous optimization for modern applications in finance, science and technology. *Optimization* **56,** 675–698 (2007).

347. Samek, W. in *Explainable Deep Learning AI* 7–33 (Elsevier, 2023).

348. Myers, D. *et al.* Foundation and large language models: fundamentals, challenges, opportunities, and social impacts. *Cluster Computing* **27,** 1–26 (2024).

349. Zhou, C. *et al.* A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *International Journal of Machine Learning and Cybernetics,* 1–65 (2024).

350. Yu, Y. *et al.* Large language model as attributed training data generator: A tale of diversity and bias. *Advances in Neural Information Processing Systems* **36** (2024).

351. Bommasani, R. *et al.* On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

352. Song, P., Ojo, A. & Curry, E. Towards Trustworthy Foundation Models: A Survey. *Available at SSRN 4985376* (2024).

353. Moor, M. *et al.* Foundation models for generalist medical artificial intelligence. *Nature* **616,** 259–265 (2023).

354. Gallegos, I. O. *et al.* Bias and fairness in large language models: A survey. *Computational Linguistics,* 1–79 (2024).

355. Achiam, J. *et al.* Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

356. Radford, A. *et al. Learning transferable visual models from natural language supervision* in *International Conference on Machine Learning* (2021), 8748–8763.

357. Ferrara, E. Should chatgpt be biased? challenges and risks of bias in large language models. *arXiv preprint arXiv:2304.03738* (2023).

358. Liu, Z. Cultural Bias in Large Language Models: A Comprehensive Analysis and Mitigation Strategies. *Journal of Transcultural Communication* (2024).

359. Vosoughi, A. *et al.* Cross Modality Bias in Visual Question Answering: A Causal View with Possible Worlds VQA. *IEEE Transactions on Multimedia* (2024).

360. Guo, Y. *et al.* On modality bias recognition and reduction. *ACM Transactions on Multimedia Computing, Communications and Applications* **19,** 1–22 (2023).

361. Patel, P. & Uddin, M. N. AI for algorithmic auditing: mitigating bias and improving fairness in big data systems. *International Journal of Social Analytics* **7,** 39–48 (2022).

362. Raji, I. D. *et al. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing* in *Proceedings of the 2020 conference on fairness, accountability, and transparency* (2020), 33–44.

363. Olorunfemi, O. L. *et al.* Towards a conceptual framework for ethical AI development in IT systems. *Computer Science & IT Research Journal* **5,** 616–627 (2024).

364. Jobin, A., Ienca, M. & Vayena, E. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* **1,** 389–399 (2019).

365. Taddeo, M. & Floridi, L. in *Ethics, Governance, and Policies in Artificial Intelligence* 91–96 (Springer, 2021).

366. De Almeida, P. G. R., dos Santos, C. D. & Farias, J. S. Artificial intelligence regulation: a framework for governance. *Ethics and Information Technology* **23,** 505–525 (2021).

367. Peters, D., Vold, K., Robinson, D. & Calvo, R. A. Responsible AI—two frameworks for ethical design practice. *IEEE Transactions on Technology and Society* **1,** 34–47 (2020).

368. Vesnic-Alujevic, L., Nascimento, S. & Polvora, A. Societal and ethical impacts of artificial intelligence: Critical notes on European policy frameworks. *Telecommunications Policy* **44,** 101961 (2020).

369. Nikolinakos, N. T. in *EU Policy and Legal Framework for Artificial Intelligence, Robotics and Related Technologies-The AI Act* 101–166 (Springer, 2023).

370. Bellamy, R. K. *et al.* AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* **63,** 4–1 (2019).