Navigating Severe Class Imbalance in Population Cohort Data

Joshua Fieggen[1], Bradley Segal[1], Emma C. Walker[2,3], Anshul Thakur[1]
Christopher C. Butler[4], David A. Clifton[1,5], and Lei Clifton[1,4]

***Abstract*— Class imbalance is a major challenge in predictive modelling for rare disease outcomes, particularly in large-scale population cohorts. Traditional machine learning models often struggle with imbalanced datasets, leading to biased performance metrics and poor generalisability. This study systematically evaluates multiple approaches to mitigate class imbalance in predicting Multiple myeloma using proteomic and clinical data from UK Biobank. We compare standard classification models (XGBoost and logistic regression) with synthetic resampling (SMOTE), anomaly detection techniques (isolation forests, local outlier factors, one-class SVM, and autoencoders), and a transformer-based foundation model (TabPFN), using standard classification performance metrics. Our results indicate that anomaly detection models generalise better than conventional classifiers (XGBoost and logistic regression), while SMOTE fails to improve, and may actively worsen, predictive performance. To address the precision-sensitivity trade-off, we introduce a sequential XGBoost ensemble classifier (SeqXGB) that prioritises high precision over sensitivity to minimise false positive predictions. Compared with a single XGBoost model, the SeqXGB approach successfully reduces false positives (420 vs 9), but significantly limits sensitivity (0.70 vs 0.15) in held-out test data. Our findings highlight that no single method is universally optimal for addressing class imbalance; rather, model selection should be guided by clinical application, balancing the risks of false positives and false negatives.**

***Clinical relevance*— This study highlights the challenges with using machine learning to predict diseases in highly imbalanced large population cohorts and underscores the need to consider clinical purpose (e.g. screening vs diagnosis) when evaluating models.**

I. INTRODUCTION

Class imbalance is a widespread challenge in medical data analysis. This is particularly true in large-scale population cohorts such as the UK Biobank [1]. Cohort studies provide unique opportunities to investigate causal relationships between predictors and outcomes as baseline measurements precede disease onset. However, their longitudinal nature makes it impossible to include more positive cases. This issue is further pronounced for diseases, such as cancers, that are individually rare at the population level [2]. As a result, conventional predictive models often struggle to accurately identify rare disease cases, leading to biased performance metrics and reduced generalisability [3].

[1]Computational Health Informatics Lab, Department of Engineering Sciences, University of Oxford, Oxford, UK [2]Nuffield Department of Women's and Reproductive Health, University of Oxford, Oxford, UK [3] Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, UK [4]Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, UK [5] Oxford Suzhou Centre for Advanced Research (OSCAR), University of Oxford, Suzhou 215123, Jiangsu, China joshua.fieggen@eng.ox.ac.uk

Most commonly employed methods to address class imbalance work at the data-level using various data resampling methods that either oversample the minority class or undersample the majority class. Of these, recent literature has suggested oversampling techniques perform better and are more stable in imbalanced medical data [4]. One of the most common oversampling methods is the Synthetic Minority Over-sampling Technique (SMOTE) [5]. However, in the context of severe class imbalance, these approaches have significant limitations where the synthetic data does not generalise to real-world distributions [6]. Another domain in machine learning where imbalanced data is common is anomaly detection, where in areas such as fraud detection imbalance can be severe [7]. Given this, it is plausible that methods typically used for anomaly detection, such as isolation forests, one-class support vector machines, local outlier factors, and autoencoders, may find utility in this parallel context. Finally, while deep learning has historically struggled to compete with gradient-boosted decision trees in tabular data due to challenges with overfitting, recent progress in developing foundation models for tabular data using in context learning, such as TabPFN, presents a third possible approach to address this challenge [8].

Previously, we developed a machine learning pipeline using XGBoost [9] to identify key predictors of Multiple myeloma in the UK Biobank [10]. Given that only 0.3% of the cohort develops myeloma, effective class imbalance management is crucial for ensuring robust and clinically meaningful predictions. In this study, we use this dataset to systematically evaluate and compare SMOTE, anomaly detection algorithms, and TabPFN to standard methods used for tabular data; namely gradient-boosted decision trees (XGBoost) and logistic regression. Upon finding that none of the approaches evaluated identified high-risk individuals with precision (i.e., low rates of false positives), we subsequently evaluated a sequential ensemble classifier, highlighting an alternate potential strategy for managing class imbalance in this context.

II. METHODS

A. Data

This study used data from a previously described machine learning feature selection pipeline applied to longitudinal population cohort data from UK Biobank [10]. In this pipeline an XGBoost algorithm with a Cox loss function was applied to plasma proteomic profiles for over 50k participants and SHAP values used to identify the top 10 predictive proteins. These 10 proteins were then combined

with 10 clinical predictors of myeloma, also derived from UK Biobank, yielding a dataset with 20 predictor features. Our disease outcome was defined as incident myeloma cases (diagnoses after the baseline date) ascertained via linked cancer registry, death registry and in-patient hospital records.

For this study the data was pre-processed by standardising continuous features using z-score normalisation and encoding categorical variables using one-hot encoding. The original 80:20 train:test data split was maintained from the previous analysis.

B. Experimental Design

We evaluated nine machine learning models to predict myeloma diagnoses using our 20 features. Each model was tuned using 5-fold stratified cross-validation, optimising the area under the receiver operating characteristic curve (AUC). The hyperparameters were selected via grid search.

As our baseline comparator models, we fitted an XGBoost and a logsitic regression model. For the XGBoost we tuned the hyperparameters column sample, ϕ; learning rate, η; maximum depth, $d_{\max}$; number of trees, $\mathcal{T}$; and subsample, ρ. For the logistic regression model we tuned the inverse regularisation strength, C, for the ℓ_1 and ℓ_2 regularisation penalties. Then to evaluate a resampling approach, we retained the logistic regression and XGBoost models within a pipeline incorporating SMOTE resampling utilising the same hyperparameter grids as in the baseline models. Next we evaluated four anomaly detection methods, namely an isolation forest (IF), a local outlier factor model (LOF), a one-class support vector machine (OCSVM), and an autoencoder (AE). For the IF we tuned the number of trees, $\mathcal{T}$; the maximum number of samples, $s_{\max}$; and the maximum number of features, $f_{\max}$. For the LOF the hyperparameters evaluated included the number of neighbours, k; the distance metric, δ; and the leaf size, λ and for the OCSVM the kernel scale parameter, γ was tuned with the kernel kept as κ = rbf. For the IF, LOF, and OCSVM models the contamination parameter was fixed at 0.003 based on prior domain knowledge. Next, we implemented a fully connected AE with a symmetric 15-4-15 bottleneck architecture, using ReLU activation in all layers. The model was trained using mean squared error (MSE) loss and the Adam optimiser with a learning rate, η of 10^{-3} for 50 epochs. Finally, TabPFN [8], a transformer-based probabilistic model, was trained on a resampled dataset from the training data, as it is optimised for 10,000 tabular rows. The training data was reprocessed by retaining all positive instances (n = 147) and randomly sampling 9850 negative instances. The classifier was trained on this processed dataset, and the predicted probabilities were threshold at the 99.7th percentile to define anomalies, based on prior domain knowledge.

C. Evaluation Metrics

The trained models are evaluated on the training and test data using AUC, precision, sensitivity, specificity, and F1-score. To derive performance metrics from confusion matrices, optimal classification thresholds were determined using Youden's J statistic [11] on the training set and applied to the test set.

D. SeqXGB Ensemble Classifier

Upon evaluation of the performance metrics, it was notable that all of the models evaluated suffered from very low precision and therefore have commensurately poor F1-scores. As a potential solution to this, we designed and implemented a sequential two stage ensemble of XGBoost models (SeqXGB). In stage one, an XGBoost classifier was trained on the entire training dataset. A gating threshold was set as the smallest probability threshold that ensured a true positive rate of at least 50%. Samples exceeding this threshold were designated as high-risk and passed to a second XGBoost classifier, trained exclusively on the high-risk subset. For the final classification low-risk samples from the first model were directly assigned a negative class, while high-risk samples were reclassified using stage two. Both XGBoost models were tuned across the same hyperparameters and in the same way (using 5-fold cross validation) as previously described with the same performance metrics.

III. RESULTS

A. Training Performance

The performance of the machine learning models on the training dataset is summarised in Table I. Among all models, the XGBoost (XGB) and the sequential ensemble of XGBoosts (SeqXGB) exhibited the highest overall performance. The SeqXGB achieved perfect precision leading to a high F1 score (0.66). In contrast, the XGB model missed fewer true cases and had higher sensitivity (0.82) and AUC (0.93) but had poor precision (0.07) due to many false positives leading to a worse F1 score (0.13). SMOTE resulted in a worsening of performance of the XGBoost model. Logistic regression (LR) and its SMOTE-augmented counterpart (LR (S)) achieved relatively high sensitivity (0.73 and 0.84, respectively) but also suffered from low precision (0.02 and 0.01), leading to poor F1-scores.

The anomaly detection models, including IF, LOF, and OCSVM, demonstrated moderate sensitivity values ranging from 0.53 to 0.64 but had low precision again leading to low F1-scores, with the highest AUC among them being 0.76 (IF). The autoencoder (AE) exhibited the lowest training performance across almost all metrics, with an AUC of 0.73. The probabilistic transformer-based model, TabPFN, showed reasonable performance with an AUC of 0.81, sensitivity of 0.71, and an F1-score of 0.04.

B. Model Performance

Model performance on the test dataset is presented in Table II and Figure 1. Similar to the training results, SeqXGB achieved the highest precision (0.31) and F1-score (0.20), albeit substantially lower than its training values. This model had a notably poor sensitivity (0.15) on the test data leading to it achieving the lowest AUC (0.68). The XGB had the next best precision (0.04) although it was again nearly an order of magnitude worse. In contrast, logistic regression

TABLE I

MODEL PERFORMANCE ON TRAINING DATA SHOWING PERFORMANCE METRICS FOR DIFFERENT MODELS.

	LR	XGB	LR (S)	XGB (S)	IF	LOF	OCSVM	AE	TabPFN	SeqXGB
AUC	0.85	**0.93**	0.87	0.90	0.76	0.73	0.75	0.73	0.81	0.75
F1 Score	0.04	0.13	0.03	0.04	0.03	0.02	0.03	0.02	0.04	**0.66**
Precision	0.02	0.07	0.01	0.02	0.01	0.01	0.01	0.01	0.02	**1.00**
Sensitivity	0.73	**0.82**	0.84	0.76	0.54	0.64	0.53	0.57	0.71	0.50
Specificity	0.87	0.96	0.79	0.87	0.87	0.72	0.87	0.80	0.88	**1.00**

Best-performing models: XGBoost (XGB) achieved the highest AUC (0.93), indicating strong discrimination between classes. SeqXGB had the highest F1 Score (0.66) and prrecision (1.00), suggesting it correctly identified positive cases with very few false positives, but its lower sensitivity (0.50) indicates it missed many true positives. XGB also demonstrated the highest sensitivity (0.82), making it the best at capturing true positives, while SeqXGB had the highest specificity (1.00). The best values for each metric are highlighted in **bold**.

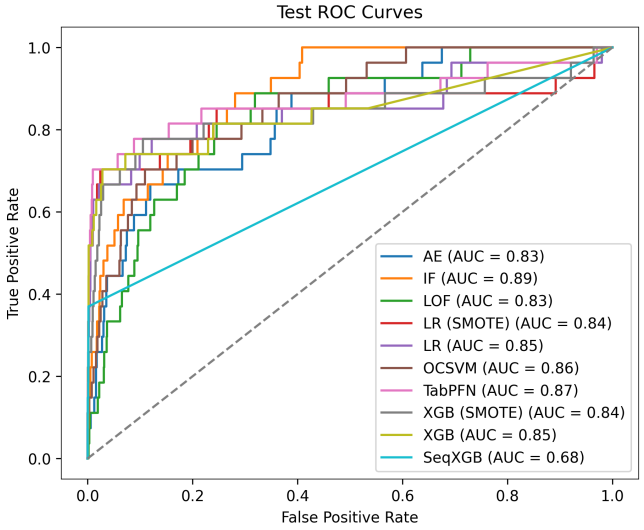

Fig. 1. Receiver Operator Characteristic (ROC) Curve plots for all models on held-out test data. AUC: Area under the curve

Fig. 2. Confusion matrices for Local Outlier Factor, Isolation Forest, SeqXGB, and XGBoost models on held-out test data. The labels "No Myeloma" and "Myeloma" indicate the predicted and actual classifications.

models retained high sensitivity (0.78) but exhibited very low precision, leading to a poor F1-scores (0.03). Similar to the results on the training data, SMOTE showed no benefit (LR) or meaningfully worsened performance (XGB) compared to the baseline models.

Notably, the isolation forest model exhibited an AUC of 0.89, surpassing XGB (0.85) and other anomaly detection models. The LOF model demonstrated the highest sensitivity (0.81) but with very low precision, resulting in an F1-score of 0.01. The autoencoder showed improved predictive capability, with an AUC of 0.83. Among the tested models, TabPFN achieved the second highest AUC on the test dataset (0.88). However, its sensitivity (0.74) was slightly lower than that of LOF and LR, and its F1-score remained low (0.02), suggesting that its predictive capability may be more effective in ranking cases rather than classification at the chosen threshold.

C. Generalisation

It is notable that all anomaly detection and deep learning-based methods improved performance (AUC) from the training to the test data suggesting good model generalisability. In contrast the XGBoost based methods all meaningfully lost performance on the held-out dataset. This highlights the substantial issue of overfitting common to these algorithms.

D. Select Confusion Matrices

Fig. 2 shows the test data confusion matrices constructed using Yoden's J-statistic (calculated on the training data) as a threshold on the LOF, IF, XGB, and SeqXGB models, respectively. This figure highlights the poor precision-sensitivity trade-off that is accentuated with severe class imbalance. The LOF, IF, and XGB models detect more true cases but also pick up many more false positive results. In contrast, SeqXGB has very few false positives, but misses many true cases.

IV. DISCUSSION

Our results underscore the challenges of severe class imbalance in large-scale cohort studies, especially for relatively rare disease endpoints such as Multiple myeloma. Precision

TABLE II

MODEL PERFORMANCE ON TEST DATA SHOWING PERFORMANCE METRICS FOR DIFFERENT MODELS.

	LR	XGB	LR (S)	XGB (S)	IF	LOF	OCSVM	AE	TabPFN	SeqXGB
AUC	0.85	0.85	0.84	0.84	**0.89**	0.83	0.86	0.83	0.87	0.68
F1 Score	0.03	0.08	0.02	0.03	0.03	0.01	0.03	0.02	0.02	**0.20**
Precision	0.02	0.04	0.01	0.02	0.01	0.01	0.01	0.01	0.01	**0.31**
Sensitivity	0.78	0.70	0.78	0.78	0.67	**0.81**	0.70	0.70	**0.81**	0.15
Specificity	0.87	0.96	0.79	0.87	0.87	0.70	0.87	0.81	0.82	**0.99**

Best-performing models: The Isolation Forest (IF) model achieved the highest AUC (0.89), indicating the best overall discrimination between classes. SeqXGB exhibited the highest F1 Score (0.20) and precision (0.31), suggesting it performed better in correctly identifying positive cases despite its lower AUC. LOF and TabPFN had the highest sensitivity (0.81), meaning they captured the most true positives. SeqXGB also had the highest specificity (0.99), implying strong ability to correctly classify negatives. The best values for each metric are highlighted in **bold**.

metrics are most impacted with models developed in populations with low disease prevalences, presenting a particular challenge in oncological research where false positive cancer diagnoses carry significant psychological burden. The results of this study highlight the systemic challenges in interpreting metrics like the AUC which obfuscate the underlying class distribution bias [2], and the difficulty of achieving a task-specific balance between sensitivity and precision.

A. Anomaly Detection vs Resampling

A notable outcome of our study is that all anomaly detection methods showed stronger generalisation on the held-out dataset than standard machine learning approaches such as XGBoost, which exhibited clear overfitting. This aligns with evidence from other medical domains, such as imaging and signal processing, that unsupervised anomaly detection models demonstrate superior generalisation capabilities [12]. The relative success of anomaly detection methods in this study highlight a gap for further research around practical and clinically relevant applications for tabular medical data. This finding emphasise that the intrinsic approach -treating positive cases as "outliers" in a high-dimensional space- may be more appropriate for rare disease contexts than forcibly balancing training data through resampling. This concept may also scale more effectively to high dimensional data where fully supervised methods are at higher risk of overfitting [13].

In contrast, SMOTE-based resampling actually degraded performance in our experiments, particularly for XGBoost. While perhaps surprising, this finding is in keeping with recent literature that has shown that the use of resampling generally results in lower calibration performance and no improvement in the discrimination performance of clinical risk prediction models [14]. Similarly, van den Goorbergh et al. found that training logistic regression models on imbalance corrected data, including using SMOTE, led to "strong and systematic overestimation of the probability for the minority class" [15]. Our finding re-emphasise this caution in the application of SMOTE or related synthetic oversampling approaches to real-world medical data, particularly as the event rate becomes extremely small.

The relative performance of the different models presents few surprises. XGBoost generally performs well across all evaluated metrics, likely due to its effectiveness in capturing complex interactions prevalent in high-dimensional biomarker data. While anomaly detection methods typically generalise better, their performance here may be limited by the highly variable and often prolonged lead times (up to 15 years) between sample collection and disease onset, making clear decision boundaries challenging. Nevertheless, like linear models (LR), these methods appear less prone to overfitting. The strong performance of TabPFN is particularly notable, as it suggests that attention-based mechanisms might capture underlying biological relationships beyond merely establishing statistical decision boundaries. If validated, this would not only improve generalisation but could also extend the utility of such models beyond risk prediction alone. This highlights exciting potential for future expansion into larger feature spaces (e.g., incorporating a greater number of proteins) and enhancing interpretable biological insights.

B. High Precision is Possible

The results of our SeqXGB model demonstrate that achieving high precision in rare disease prediction in large population cohort studies is indeed possible. However, this came at the significant cost of model sensitivity with the model missing 85% of true test cases (Fig. 2). This highlights the fundamental challenge of this trade-off in predictive modelling in imbalanced datasets[15]. Nonetheless, this result suggests that in scenarios where false positive results carry significant clinical consequence, hybrid approaches, such as multi-stage risk stratification, may be a promising avenue for future work. This could be extended to methods, like support vector machines, to subset cases near the decision boundary and then use an XGBoost model to classify those cases.

C. Utility in Clinical Contexts: Myeloma

The persistent difficulty in achieving both high sensitivity and high precision, as seen in Fig. 2, reinforces that the "optimal" classifier threshold depends strongly on clinical purpose. This is especially relevant in Multiple myeloma where clinical disease is always preceded by a non-specific precursor state; monoclonal gamopathy of uncertain significance (MGUS). While MGUS can be screened for, there is much debate around the utility of this due to low rates of annual progression to clinical disease ($\tilde{1}\%$) [16]. Thus if one considers a risk model for myeloma, its position in relation

to MGUS screening is exceptionally relevant. For instance if the XGBoost model in Fig. 2 was used before screening, and only high risk individuals were screened, you would increase screening yield from 0.3% to 4.5%. In contrast if you wanted to use a model to decide which patients with minimally symptomatic MGUS needed a bone marrow biopsy, using SeqXGB would be the only viable option. This underscores that no single model is universally "best," but rather that the choice must be guided by the relative costs of false positives vs false negatives. This is further advanced by Birch et al. who argue that even decision thresholds within models that are not tailored to the patient's values and attitudes about risk are ethically questionable [17].

D. Limitations and Future Work

In terms of data, one of the limitations of this work was the lack of availability of an external validation cohort. While we attempted to reduce overfitting through hyperparameter tuning and cross-validation, prospective studies or validation in distinct populations would help unpack the variations in generalisability across the different approaches noted in this study. In addition, we only used 20 predictor features (10 proteomic, 10 clinical). A higher dimensional feature space might enhance predictive capacity or change the relative performance of anomaly detection, zero-shot learning, or conventional classification.

With respect to model architectures, we only evaluated one sequential ensemble model as a proof of concept. To fully understand the potential benefits of this approach, further combinations of classifiers should be considered. Finally, given the rapid advancement of foundation models and the broader movement towards large-scale pre-training and fine-tuning paradigms, new avenues exist for tabular data. Zero-shot or few-shot methods may sidestep some of the overfitting challenges of heavily parametrised models (e.g. XGBoost) in highly imbalanced settings.

V. CONCLUSIONS

This work highlights that no single "best" solution exists for handling class imbalance in large-scale cohort studies. While anomaly detection approaches offered better generalisation than resampling and significantly outperformed SMOTE, models were forced to sacrifice precision for sensitivity (or vice versa). Our sequential XGBoost architecture (SeqXGB) demonstrated the ability to tune this trade-off differently from a single classifier, although it remains particularly susceptible to over-fitting. Ultimately, deciding among these approaches depends crucially on the clinical end goal. For cancer screening contexts where missing as few cases as possible is paramount, a high sensitivity approach with more false positives may be optimal, whereas in more invasive diagnostic settings a high precision approach could be preferable.

ACKNOWLEDGMENTS

This research has been conducted using the UK Biobank Resource under Application Number 83801. This work uses data provided by patients and collected by the NHS as part of their care and support. We thank the participants of the UK Biobank study without whom this research would not have been possible. Computing for this study used the Oxford Biomedical Research Computing (BMRC) facility, a joint development between the Wellcome Centre for Human Genetics and the Big Data Institute supported by Health Data Research UK and the NIHR Oxford Biomedical Research Centre. JF and BS are supported by Rhodes Scholarships. ECW is supported by the EPSRC Centre for Doctoral Training in Health Data Science (grant EP/S02428X/1). DAC is supported by the Pandemic Sciences Institute at the University of Oxford; the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC); an NIHR Research Professorship; a Royal Academy of Engineering Research Chair; the Wellcome Trust funded VITAL project (grant 204904/Z/16/Z); the EPSRC (grant EP/W031744/1); and the InnoHK Hong Kong Centre for Cerebro-cardiovascular Engineering (COCHE). The ADH group at the Nuffield Department of Primary Care Health Sciences is supported by the National Institute for Health and Care Research (NIHR) Applied Research Collaboration Oxford and Thames Valley at Oxford Health NHS Foundation Trust. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

DATA AND CODE AVAILABILITY

The data reported in this paper are available via application directly to the UK Biobank, https://www.ukbiobank.ac.uk. All code is available at github.com/jfieggen/ad4mm

REFERENCES

[1] N. Allen, C. Sudlow, P. Downey, T. Peakman, J. Danesh, P. Elliott, J. Gallacher, J. Green, P. Matthews, J. Pell *et al.*, "Uk biobank: Current status and what it means for epidemiology," *Health Policy and Technology*, vol. 1, no. 3, pp. 123–126, 2012.

[2] E. Tasci, Y. Zhuge, K. Camphausen, and A. V. Krauze, "Bias and class imbalance in oncologic data—towards inclusive and transferrable ai in large scale oncology data sets," *Cancers*, vol. 14, no. 12, p. 2897, 2022.

[3] A. Visibelli, B. Roncaglia, O. Spiga, and A. Santucci, "The impact of artificial intelligence in the odyssey of rare diseases," *Biomedicines*, vol. 11, no. 3, p. 887, 2023.

[4] M. Khushi, K. Shaukat, T. M. Alam, I. A. Hameed, S. Uddin, S. Luo, X. Yang, and M. C. Reyes, "A comparative performance analysis of data resampling methods on imbalance medical data," *IEEE Access*, vol. 9, pp. 109 960–109 975, 2021.

[5] A. Fernández, S. Garcia, F. Herrera, and N. V. Chawla, "Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary," *Journal of artificial intelligence research*, vol. 61, pp. 863–905, 2018.

[6] A. S. Tarawneh, A. B. Hassanat, G. A. Altarawneh, and A. Almuhaimeed, "Stop oversampling for class imbalance learning: A review," *IEEE Access*, vol. 10, pp. 47 643–47 660, 2022.

[7] Q. Wang, "Research on the application of machine learning in financial anomaly detection," *iBusiness*, vol. 16, no. 4, pp. 173–183, 2024.

[8] N. Hollmann, S. Müller, L. Purucker, A. Krishnakumar, M. Körfer, S. B. Hoo, R. T. Schirrmeister, and F. Hutter, "Accurate predictions on small data with a tabular foundation model," *Nature*, vol. 637, no. 8045, pp. 319–326, 2025.

[9] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: ACM, 2016, pp. 785–794. [Online]. Available: http://doi.acm.org/10.1145/2939672.2939785

[10] J. J. Fieggen, A. Thakur, C. C. Butler, K. Ramasamy, A. Thakurta, D. A. Clifton, and L. Clifton, "Dysregulated immune proteins in plasma in the uk biobank predict multiple myeloma 12 years before clinical diagnosis," *medRxiv*, 2025. [Online]. Available: https://www.medrxiv.org/content/early/2025/02/05/2025.02.04.25321690

[11] W. J. Youden, "Index for rating diagnostic tests," *Cancer*, vol. 3, no. 1, pp. 32–35, 1950.

[12] T. Fernando, H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, "Deep learning for medical anomaly detection–a survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 7, pp. 1–37, 2021.

[13] P. D. Talagala, R. J. Hyndman, and K. Smith-Miles, "Anomaly detection in high-dimensional data," *Journal of Computational and Graphical Statistics*, vol. 30, no. 2, pp. 360–374, 2021.

[14] M. Piccininni, M. Wechsung, B. Van Calster, J. L. Rohmann, S. Konigorski, and M. van Smeden, "Understanding random resampling techniques for class imbalance correction and their consequences on calibration and discrimination of clinical risk prediction models," *Journal of Biomedical Informatics*, p. 104666, 2024.

[15] R. van den Goorbergh, M. van Smeden, D. Timmerman, and B. Van Calster, "The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression," *Journal of the American Medical Informatics Association*, vol. 29, no. 9, pp. 1525–1534, 2022.

[16] I. M. Ghobrial and F. Chabrun, "Is it time to screen for multiple myeloma?" *Blood*, vol. 145, no. 3, pp. 253–255, 2025.

[17] J. Birch, K. A. Creel, A. K. Jha, and A. Plutynski, "Clinical decisions using ai must consider patient values," *Nature medicine*, vol. 28, no. 2, pp. 229–232, 2022.