tvst
Artificial Intelligence

Dynamic Prediction of Treatment Failure in Ocular Tuberculosis Using Machine Learning and Explainable AI

William Rojas-Carabali[1–3,*], Tristan Guérand[4,*], Carlos Cifuentes-González[2,3], John Abisheganaden[5], Palvannan RK[5], Yap Chun Wei[5], Germán Mejía-Salgado[6], Alejandra de-la-Torre[6], Justine R. Smith[7], John H. Kempen[8–11], Quan Dong Nguyen[12], Carlos Pavesio[13], Bernett Lee[1,14–16], Vishali Gupta[17], Thomas Peyrin[4], and Rupesh Agrawal[2,3,18–21]; for the Collaborative Ocular Tuberculosis Study (COTS) Group

[1] Centre for Biomedical Informatics, Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore
[2] Department of Ophthalmology, Tan Tock Seng Hospital, National Healthcare Group, Singapore, Singapore
[3] Programme for Ocular Inflammation & Infection Translational Research, National Healthcare Group, Singapore, Singapore
[4] Nanyang Technological University, Singapore, Singapore
[5] Health Services and Outcomes Research, National Healthcare Group, Singapore, Singapore
[6] Neuroscience Research Group (NEUROS), Neurovitae Center for Neuroscience, Institute of Translational Medicine (IMT), Escuela de Medicina y Ciencias de la Salud, Universidad del Rosario, Bogotá, Colombia
[7] Flinders University College of Medicine and Public Health, Adelaide, Australia; Queensland Eye Institute, Brisbane, Australia
[8] Department of Ophthalmology, Massachusetts Eye and Ear/Harvard Medical School; and Schepens Eye Research Institute; Boston, MA, USA
[9] Sight for Souls, Bellevue, WA, USA
[10] Addis Ababa University Department of Ophthalmology, Addis Ababa, Ethiopia
[11] MyungSung Christian Medical Center (MCM) Eye Unit, MCM Comprehensive Specialized Hospital, and MyungSung Medical School, Addis Ababa, Ethiopia
[12] Byers Eye Institute, Stanford University, Palo Alto, California, USA
[13] National Institute for Health Research Biomedical Research Centre, Moorfields Eye Hospital, UK
[14] Singapore Immunology Network (SIgN), Agency for Science, Technology and Research (A*STAR), Singapore, Singapore
[15] Infectious Disease Labs (ID Labs), Agency for Science, Technology and Research (A*STAR), Singapore, Singapore
[16] Khoo Teck Puat Hospital, Singapore, Singapore
[17] Post Graduate Institute of Medical Education and Research (PGIMER), Advance Eye Centre, Chandigarh, India
[18] Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore
[19] Singapore Eye Research Institute, Singapore, Singapore
[20] Duke NUS Medical School, Singapore, Singapore
[21] Moorfields Eye Hospital, NHUS Foundation Trust, London, UK

Correspondence: Rupesh Agrawal, Senior Consultant, National Healthcare Group Eye Institute, Tan Tock Seng Hospital, Singapore 308433, Singapore. e-mail: rupeshttsh@gmail.com

Received: April 30, 2025
Accepted: August 27, 2025
Published: October 24, 2025

Purpose: Ocular tuberculosis (OTB) poses significant challenges in treatment because of its complex diagnostic and therapeutic landscapes. Predicting treatment failure effectively is crucial for timely intervention and improving patient outcomes. We report the application of machine learning (ML) approaches to (i) allow predictions using baseline data and (ii) dynamically update predictions based on patient history and new observations.

Methods: The Collaborative Ocular Tuberculosis Study (COTS) was a multinational retrospective study encompassing data from 836 patients with tubercular uveitis across 27 international eye care centers. This study evaluated the performance of nine ML models to predict treatment failure at six, 12, and 24 months using baseline and longitudinal data. Metrics such as area under the curve (AUC), precision, accuracy, F1-score, and model complexity were reported. Top features and their importance were identified using XGBoost, with weight of evidence and information value calculated to enhance interpretability.

Keywords: ocular tuberculosis; explainable AI; COTS calculator; clinical decision support

Citation: Rojas-Carabali W, Guérand T, Cifuentes-González C, Abisheganaden J, RK P, Wei YC, Mejía-Salgado G, de-la-Torre A, Smith JR, Kempen JH, Nguyen QD, Pavesio C, Lee B, Gupta V, Peyrin T, Agrawal R. Dynamic prediction of treatment failure in ocular tuberculosis using machine learning and explainable AI. Transl Vis Sci Technol. 2025;14(10):31, https://doi.org/10.1167/tvst.14.10.31

Results: Data were collected from 836, 769, and 418 patients at six, 12, and 24 months, respectively. XGBoost and Random Forest (RF) models consistently showed superior performance across all timepoints. At 6 months, XGBoost achieved an AUC of 0.915 $\pm$ 0.019 and accuracy of 0.879 $\pm$ 0.027. At 12 months, RF outperformed with an AUC of 0.921 $\pm$ 0.011 and accuracy of 0.944 $\pm$ 0.022. At 24 months, RF maintained high accuracy (0.960 $\pm$ 0.029) despite a slight drop in AUC (0.888 $\pm$ 0.099). Deep Neural Networks and TT-net models were underfitted.

Conclusions: ML models like XGBoost and RF demonstrate promise for early and accurate prediction of treatment failure in OTB, with explainability tools enhancing clinical interpretability.

Translational Relevance: This study bridges basic ML research and clinical care by offering explainable, performance-driven models that support real-time, data-informed treatment decisions in managing OTB, potentially improving long-term outcomes.

Introduction

Tuberculosis (TB) remains one of the most daunting infectious diseases in human history, inflicting significant morbidity and mortality across the globe. Each year, an estimated 10.4 million new cases of TB are reported, underscoring persistent challenges in managing and controlling the disease.[1] Despite significant strides in diagnostic and therapeutic methodologies, the causative bacterium, *Mycobacterium tuberculosis*, continues to challenge standard management protocols due to its persistent nature. This issue is particularly pronounced in ocular tuberculosis (OTB), because the eye's unique immune-privileged status complicates the infection response, leading to a wide array of complex ocular symptoms that can affect nearly every tissue of the eye.[2]

The standard treatment for TB, including OTB, involves a prolonged regimen of potent antimicrobials lasting from six to nine months.[3,4] Although these treatments are generally effective, they are not without substantial risks of significant toxicity.[5] Alarmingly, persistent inflammation occurs in one in 10 patients with OTB, necessitating the extended use of corticosteroids and immunomodulatory drugs.[2] This high rate of treatment failure underscores the critical need for more precise and effective therapeutic approaches, particularly strategies that can preemptively identify patients at risk of poor outcomes, thereby mitigating unnecessary exposure to toxic drugs.

In this context, the emergence of artificial intelligence (AI) presents promising opportunities for enhancing personalized medicine approaches. While AI has been leveraged in other infectious diseases and some forms of TB to predict therapeutic efficacy, its application in OTB remains underexplored.[6,7] Machine learning (ML) models such as support vector machines and convolutional neural networks have shown potential in predicting the required duration of treatment for pulmonary TB.[6] Meanwhile, models such as random forest, classification, and regression trees have demonstrated promise in prognosticating adverse reactions and treatment outcomes. These models have also been effective in predicting drug resistance in pulmonary TB, indicating their potential utility in management of OTB.[6]

This study aimed to assess the effectiveness of various ML models in predicting therapeutic failure among OTB patients using (i) baseline and (ii) longitudinal data, and to evaluate the complexity of these models. Leveraging the Collaborative Ocular Tuberculosis Study (COTS)[8] dataset, the largest collected to date in an international effort, we aimed to enhance the explainability of these models to provide a clearer understanding of their decision-making processes, ultimately improving the approach to diagnosing and treating OTB.

Methods

Design and Materials

This retrospective longitudinal study was approved by the ethics committee of the Postgraduate Institute of Medical Education and Research and was conducted in accordance with the principles outlined in the Declaration of Helsinki. The requirement for informed consent was waived because of the retrospective nature of the study. This study followed the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis in Artificial Intelligence (TRIPOD+AI) statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods

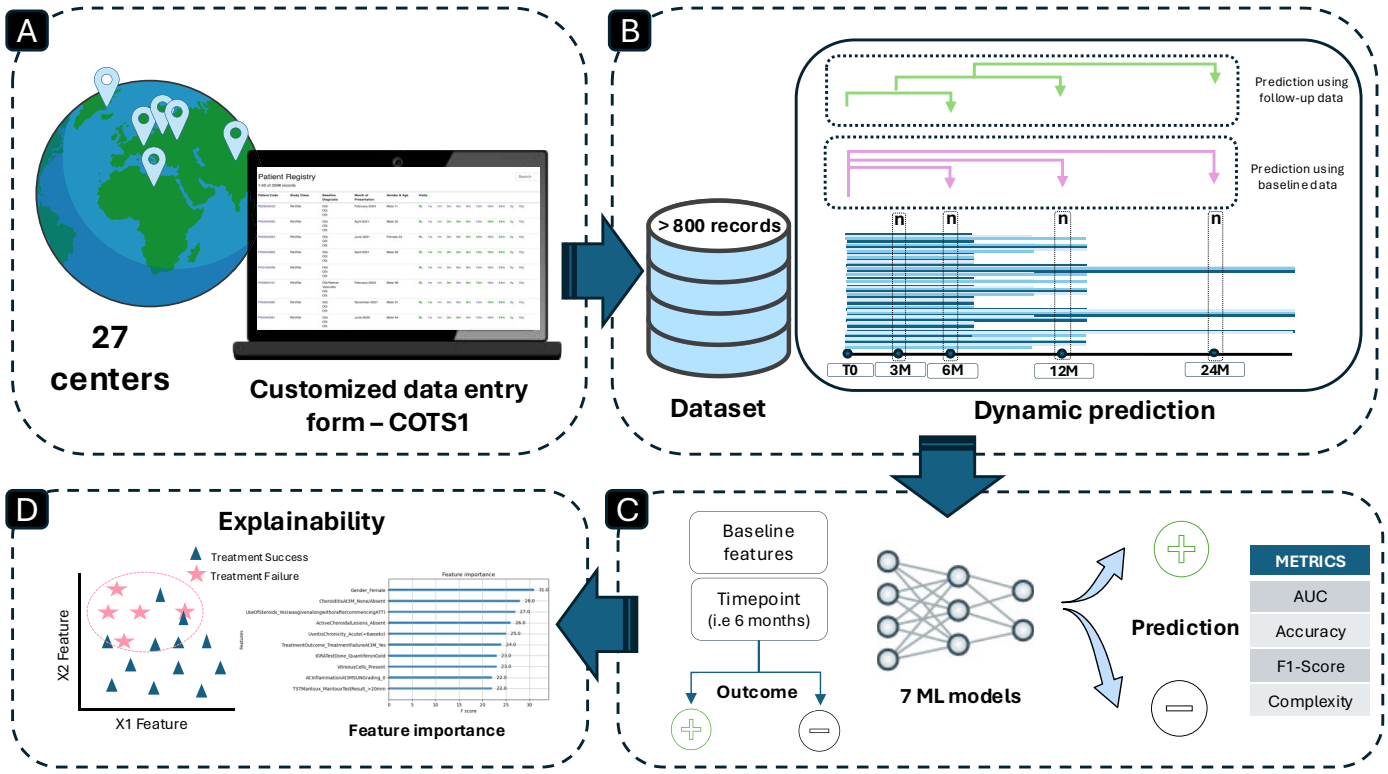

Figure 1. Overview of the multicenter study workflow for predicting treatment outcomes in uveitis. This flowchart provides a comprehensive overview of the structured workflow used in this multicenter study designed to predict treatment outcomes in ocular tuberculosis using over 836 patient records across 27 centers. The process begins with data collection through a customized form (COTS1), leading to the creation of a robust dataset. The data undergo stepwise prediction analyses at six, 12, and 24 months using baseline features and time-point data to generate predictions. Seven different machine learning models are evaluated based on metrics such as AUC, accuracy, F1-score, and complexity. Additionally, the study emphasizes explainability, as shown by the feature importance analysis, which helps in understanding the key predictors of treatment success or failure.

(Supplemental Material S1, TRIPOD+AI Checklist).[9] Figure 1 summarizes the design for the current study.

Data

This study used all data from 836 patients diagnosed with presumed OTB, obtained from the existing COTS database that drew from 27 international eye care centers. Dataset variables and the data collection process have been reported in detail elsewhere.[8] Data collected from January 2004 to December 2014 were analyzed. Further information about data preparation is available in Supplementary Material S2.

Experimental Setup

Eight cores Intel Core i7-8650U CPU was clocked at 1.90 GHz and 16 GB RAM. All experiments were seeded and completed with Python (version 3.11.5) with a fivefold cross-validation.

Main Outcome

This study included patients with clinical data collected at baseline (static observations) and during follow-up (longitudinal observations) to predict treatment failure at six, 12, and 24 months (Fig. 1B). Treatment failure was defined as the persistence or recurrence of inflammation within six months after antitubercular therapy (ATT), failure to taper oral corticosteroids (prednisone <10 mg/day) or topical corticosteroid eye drops (<2 drops/day), or the need for corticosteroid-sparing immunosuppressive therapy because of recalcitrant inflammation. For longitudinal analysis, each dataset included clinical features and test results obtained before the specified time points. For example, the six-month dataset encompassed all data up to, but excluding, the six-month mark. Records lacking endpoint data were excluded.

Approach

We used different methods from scikit learn library and IBM XAI[10,11] that focused on either giving rules or

interpretable models, that is that explained the decision process globally and not locally with methods like LIME, "a novel explanation technique that explains the predictions of any classifier in an interpretable and faithful manner, by learning an interpretable model locally around the prediction."[12,13] We aimed to explain the model so that medical practitioners could understand the decision process. We chose different models that explained the decision process with rules or trees, including deep neural network (DNN), logistic regression (LR), column generation (CG),[14] decision trees (DT),[15] random forest (RF),[16] general linear rule model (GLRM),[17] Repeated Incremental Pruning to Produce Error Reduction or Ripper,[18] TT-rules,[19] and eXtreme Gradient Boosting (XGBoost).[20] Detailed information about these models is available in Supplementary Material S2.

Statistical Analysis

We evaluated each model's performance using classic metrics such as the area under the curve (AUC), F1-score, and accuracy. The scores reported represent the mean and standard deviation derived from a fivefold cross-validation process using 80–20 splits. Additionally, to ensure the decision-making process would be comprehensible to medical practitioners, we assessed model complexity by counting the number of Boolean operations required to compute the model. Lower complexity indicated easier model interpretability. We also identified the top features and their importance for the best-performing algorithm using XGBoost. Moreover, weight of evidence (WoE) and information value (IV) were calculated for the top features to improve explainability. WoE helps interpret how features contribute to different outcomes, whereas IV highlights their importance in predicting the target variable.

Results

For the six-month dataset, 836 patients were recorded. Each record had 797 binary features, and 24.3% of the patients were registered as failing treatment. For the 12-month dataset, 769 patients with 833 binary features were recorded. 17.4% of these had failed treatment. Finally, the 24-month dataset had 418 patients with 717 binary features, and 9.57% were reported as failing treatment (Table 1). Table 2 shows the prediction at each time point, with results as mean plus standard deviation from the 5-fold cross-validation. Figures 2A–C show the plot of AUC against complexity for the different models.

Prediction Using Baseline Data

At six months, XGBoost stands out with an AUC of 0.899 ± 0.018, complemented by a precision of 0.790 ± 0.040 and an F1 score of 0.714 ± 0.022, indicating strong classification capability (Table 2). In contrast, other models, like TT-net, DNN, LR, CG and GLRN, showed weaker precision and AUCs, highlighting a potential compromise in their predictions, all of them with F1 scores below 0.5 (Supplementary Material S3). By the 12-month interval, enhancements in model metrics are evident. RF achieves an accuracy of 0.903 ± 0.017 and an AUC of 0.901 ± 0.023, showcasing significant improvements in classifying capabilities over time (Table 2). Conversely, models such as DNN and LR showed minimal evolution, with DNN recorded a persistent F1 of 0.000 ± 0.000 across all intervals, indicating a failure in predictive adequacy for the dataset. The 24-month predictions further delineate model maturation (Supplementary Material S3). XGBoost maintained robustness with an AUC of 0.870 ± 0.064, although a slight dip compared to earlier intervals. Remarkably, DTs show considerable growth, achieving an AUC of 0.783 ± 0.094, which is an increment from its six-month performance. Table 2 shows the results for the top 3 models. Results for the other models are available in Figures 2A–F and Supplementary Material S3 and S4.

Prediction Using Longitudinal Data

Using longitudinal data, the models demonstrated varying levels of effectiveness across the six-month, 12-month, and 24-month prediction timepoints, with a consistent trade-off between accuracy and complexity. The XGBoost model excelled at six months with the highest AUC of 0.915 ± 0.019. However, its complexity was considerably higher at 1.3k ± 10.2 Boolean operations compared to simpler models like Ripper, which showed significantly inferior performance metrics despite a much lower complexity of 9.4 ± 5.24. DTs and RF provided a more balanced approach with moderate accuracy and complexity. By the 12-month evaluation, RFs led in both AUC (0.944 ± 0.022) while also showing a reduction in complexity (19k ± 245), with XGBoost maintaining strong performance and slightly reduced complexity. Simpler models like TT-rules continued to lag in performance, highlighting the ongoing challenge of achieving high accuracy with low complexity. At the 24-month mark, the XGBoost model's performance slightly decreased in terms of AUC (0.856 ± 0.115) compared to earlier timepoints. Likewise, RF maintained a high AUC

Table 1. Baseline Characteristics of Patients Included in the Analysis

	6M		12M		24M	
	N = 836	%	*N* = 769	%	*N* = 418	%
Country						
Countries with high TB (endemic)	549	65.68%	511	66.44%	260	62.21%
Countries with mid-to-low TB (non-endemic)	287	34.33%	258	33.54%	158	37.8%
Gender						
Male	399	47.73%	362	47.7%	189	45.22%
Female	384	45.93%	357	46.42%	207	49.52%
Missing data	53	6.34%	50	6.5%	22	5.26%
Phenotype						
Anterior uveitis	99	11.84%	99	12.87%	56	13.4%
Intermediate uveitis	95	11.36%	106	13.78%	52	12.44%
Panuveitis	292	34.93%	257	33.42%	149	35.65%
Retinal vasculitis	193	23.9%	173	22.5%	92	22.01%
Choroiditis						
Multifocal	5	0.6%	4	0.52%	0	0
Serpiginous	121	14.47%	111	14.43%	65	15.55%
Tuberculoma	33	3.95%	25	3.25%	15	3.59%
Mantoux						
Not done/unknown	225	26.91%	189	24.58%	123	29.43%
Positive	541	64.71%	515	66.97%	264	63.16%
Negative	70	8.37%	65	8.45%	31	7.42%
IGRA test						
Not done/unknown	343	41.03%	310	40.31%	145	34.69%
Positive	40	4.78%	34	4.42%	29	6.94%
Negative	2	0.24%	2	0.26%	2	0.48%
CXR						
Not done/unknown	192	22.97%	157	20.42%	91	21.77%
Positive	143	17.11%	137	17.82%	67	16.03%
Negative	479	57.3%	455	59.17%	252	60.29%
Treatment failure (Yes)	203	24.2%	134	17.4%	40	9.6%

(0.888 ± 0.099) with further complexity reduction (6k $\pm$ 261). The complexity of most models generally decreased, with models like DT showing a substantial reduction from 172.2 to 40.2 Boolean operations, suggesting the models are more efficient for a 24-month prediction compared to a six-month or 12-month prediction. Table 2 shows the results for the top 3 models. Results for the other models are available in Figure 2 and Supplementary Material S3 and S4.

Features Importance

At six months, gender, choroiditis at three months, use of steroids, and active choroidal lesions were identified as the most relevant predictors of treatment failure. Being female (WoE = 0.225, IV = 0.048) and the presence of active choroidal lesions at three months (WoE = 0.112, IV = 0.024) were weakly associated with a higher likelihood of treatment failure. In contrast, steroid use was moderately associated with treatment success (WoE = −0.501, IV = 0.242). Several variables, including choroiditis and uveitis chronicity, could not be evaluated because of infinite values. At 12 months, gender, retinal vasculitis, systemic TB, and treatment failure at six months emerged as the most significant predictors. Being female (WoE = 0.214, IV = 0.046) and the absence of retinal vasculitis (WoE = 0.052, IV = 0.000) were weakly associated with treatment failure. Steroid use before ATT was again moderately linked to treatment success (WoE = −0.731, IV = 0.216). Uveitis chronicity and additional investigations could not be assessed because of infinite values. At 24 months, the most relevant predictors were treatment failure at 12

Table 2. Results of the Top Three Methods on the Six-Month, 12-Month, and 24-Month Prediction

	DT Baseline	DT Longitudinal	RF Baseline	RF Longitudinal	XGBoost Baseline	XGBoost Longitudinal
6 months						
Precision	0.692 ± 0.062	0.706 ± 0.059	0.935 ± 0.062	0.939 ± 0.036	0.790 ± 0.040	0.820 ± 0.052
Recall	0.625 ± 0.051	0.644 ± 0.055	0.505 ± 0.070	0.558 ± 0.054	0.654 ± 0.042	0.667 ± 0.062
Accuracy	0.833 ± 0.021	0.843 ± 0.017	0.864 ± 0.019	0.879 ± 0.023	0.868 ± 0.017	0.879 ± 0.027
AUC	0.764 ± 0.017	0.777 ± 0.027	0.903 ± 0.016	0.913 ± 0.006	0.899 ± 0.018	0.915 ± 0.019
F1-score	0.653 ± 0.031	0.672 ± 0.046	0.650 ± 0.046	0.698 ± 0.047	0.714 ± 0.022	0.735 ± 0.054
Complexity	—	172.2 ± 8.63	—	27k ± 303	—	1.3k ± 10.2
12 months						
Precision	0.591 ± 0.098	0.637 ± 0.047	0.978 ± 0.044	0.867 ± 0.125	0.751 ± 0.080	0.753 ± 0.098
Recall	0.677 ± 0.084	0.698 ± 0.076	0.364 ± 0.056	0.570 ± 0.034	0.626 ± 0.045	0.661 ± 0.060
Accuracy	0.882 ± 0.026	0.894 ± 0.010	0.903 ± 0.017	0.921 ± 0.011	0.913 ± 0.007	0.914 ± 0.010
AUC	0.797 ± 0.050	0.814 ± 0.034	0.901 ± 0.023	0.944 ± 0.022	0.890 ± 0.012	0.924 ± 0.032
F1-score	0.631 ± 0.091	0.661 ± 0.025	0.528 ± 0.063	0.683 ± 0.041	0.681 ± 0.048	0.697 ± 0.033
Complexity	—	119.4 ± 6.86	—	19k ± 245	—	1k ± 16.3
24 months						
Precision	0.636 ± 0.241	0.750 ± 0.232	0.800 ± 0.400	0.800 ± 0.400	0.800 ± 0.163	0.745 ± 0.287
Recall	0.598 ± 0.199	0.716 ± 0.199	0.132 ± 0.093	0.516 ± 0.293	0.517 ± 0.137	0.635 ± 0.189
Accuracy	0.936 ± 0.016	0.952 ± 0.025	0.921 ± 0.040	0.960 ± 0.029	0.943 ± 0.032	0.955 ± 0.024
AUC	0.783 ± 0.094	0.848 ± 0.094	0.862 ± 0.080	0.888 ± 0.099	0.870 ± 0.064	0.856 ± 0.115
F1-score	0.587 ± 0.155	0.691 ± 0.153	0.221 ± 0.147	0.618 ± 0.329	0.608 ± 0.115	0.672 ± 0.218
Complexity	—	40.2 ± 4.49	—	6k ± 261	—	372 ± 14.8

months (WoE = 0.124, IV = 0.033) and uveitis chronicity with relapsing episodes beyond 12 weeks (WoE = 1.330, IV = 0.050), both weakly associated with treatment failure. Systemic TB diagnosis and the absence of pulmonary lesions showed no predictive value (Fig. 3).

Discussion

Ocular tuberculosis poses significant challenges in diagnosis and treatment, leading to potential treatment failures. The disease can result in severe visual impairment and blindness if not adequately managed.[21] Treatment failure in OTB is particularly concerning due to the rise in drug-resistant strains of *M. tuberculosis*, which complicates therapeutic strategies and represents a substantial public health threat.[4] Variability in presentation and subtleties involved in treatment demand precise and personalized therapeutic strategies. Artificial intelligence, which can analyze large datasets and uncover patterns that may not be immediately apparent to humans, is ideally suited for such tasks. However, the development and deployment of AI in this area have been slow, primarily because of the scarcity of high-quality, comprehensive datasets that are crucial for training robust AI models.[22]

In response to these challenges, our study leveraged a large, multi-center dataset from the COTS,[23] which includes data from more than 800 patients diagnosed with presumed OTB across 27 international eye care centers. Key findings include the predominance of posterior uveitis (36.3%) and panuveitis (35.3%), with choroidal involvement (64.4%) and vitreous haze (45.4%) as common clinical features. Only 23.3% of patients had a known history of systemic TB, confirming that ocular TB often presents without systemic symptoms. Management practices varied globally, with 84.6% receiving ATT and 85% treated with corticosteroids. Treatment failure was observed in 12.7% of cases, particularly in patients with choroidal involvement and vitreous haze. Regional and ethnic variations were noted in phenotypes of retinal vasculitis and choroiditis, highlighting the need for standardized diagnostic and therapeutic protocols for ocular TB.[8]

Studies have shown that factors like drug sensitivity patterns, imaging findings, and demographic variables can predict treatment failure effectively.[24] By providing insights into these predictive features, ML models can assist clinicians in tailoring treatment plans and improving patient monitoring, thereby enhancing treatment success rates and mitigating the impact of drug resistance.[25] Thus leveraging EML in OTB can be pivotal in addressing the complexities of treatment failure and improving patient outcomes.

The objective of using EML models in predicting treatment failure in OTB was to develop a method-

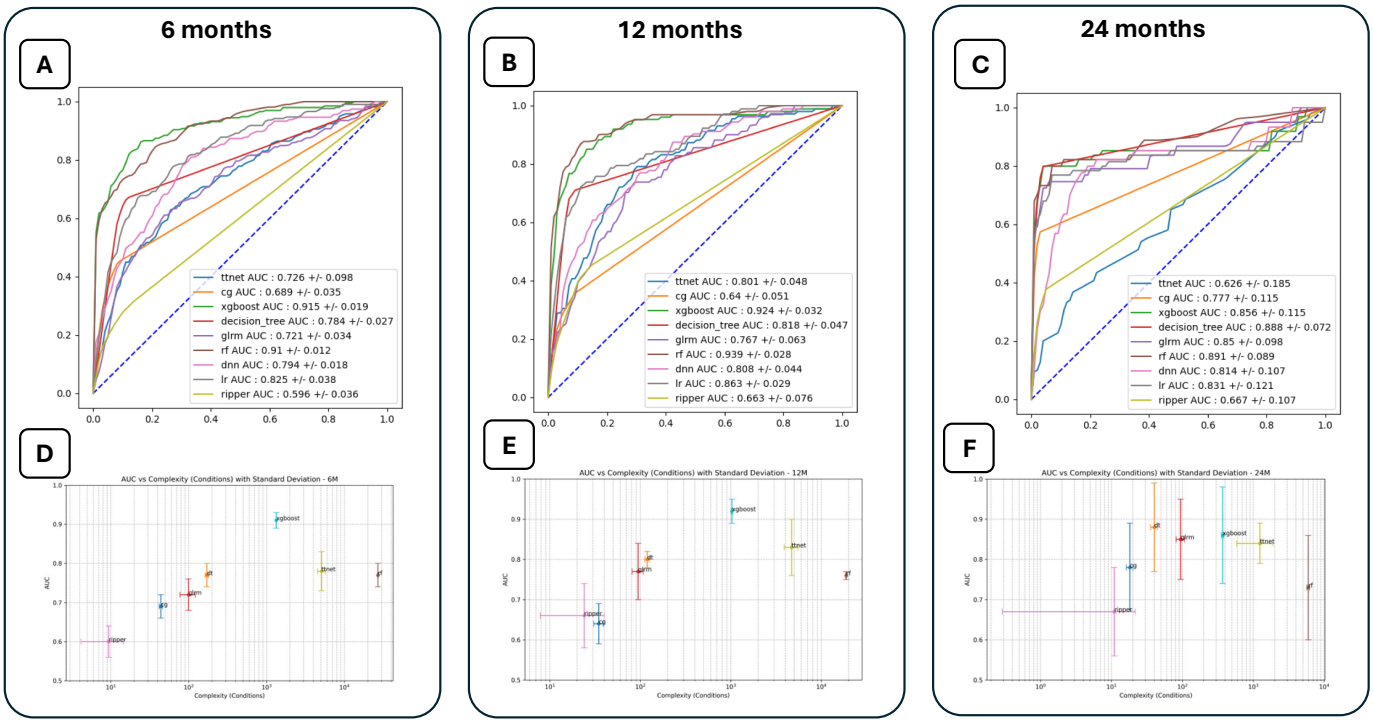

Figure 2. ROC curves with AUC values and comparative analysis of algorithm performance by AUC across complexity levels at six, 12, and 24 Months. AUC performance of several algorithms—XGBoost, RIPPER, CG, GLRM, DT, TT-net, RF, DNN, and LR—over three time points: six months (**A**), 12 months (**B**), and 24 months (**C**). Each panel presents ROC curves with AUC values and standard deviations for each algorithm, highlighting their predictive accuracy over time. In **D–F**, each panel displays the AUC against the logarithm of model complexity (number of conditions in a log scale), along with *error bars* representing standard deviations, illustrating the trade-off between model accuracy and complexity. Notably, XGBoost consistently shows high AUC across all time points and complexities, suggesting its robustness in handling varying complexities while maintaining predictive accuracy.

ology that is both effective in its predictive accuracy and transparent in its decision-making processes. This dual requirement is critical in medical settings where understanding the rationale behind a prediction is as important as the prediction itself. The explainability of a model facilitates a deeper trust and reliance among medical practitioners, allowing them to make more informed decisions regarding patient treatment plans.[26]

Our results underscore the varied trajectories of model performance over time, with certain models like Random Forest and XGBoost demonstrating adaptability and sustained efficacy, whereas others like DNN and LR exhibit stagnation or diminished returns. Such longitudinal data provides critical insights into the temporal dynamics of predictive modeling, essential for optimizing long-term application strategies in real-world scenarios. The results also highlighted a common challenge in ML: the trade-off between model complexity and interpretability. For instance, XGBoost and RF, while providing the highest accuracy as indicated by AUC metrics (Fig. 3), presented significant complexity because of the numerous decision nodes. This complexity can be a barrier in clinical practice, where decisions need to be made quickly and justified transparently to patients and other healthcare providers. XGBoost models, despite their high performance, often require detailed explanations to ensure transparency in clinical decision-making processes, which can be challenging in high-stakes environments such as coronary care units or for predicting adverse outcomes in diseases like community-acquired pneumonia.[27,28]

Conversely, simpler models like DT and the Ripper algorithm, although less accurate, offer clearer insights into their decision processes through more straightforward rules or decision paths.[15,18] These models are particularly valuable in clinical settings where explanations are required to justify medical decisions to non-specialist stakeholders. The DT, with its flowchart-like structure, illustrates this balance particularly well, offering a compromise between accuracy and simplicity. Furthermore, applying CG and GLRM introduces interesting possibilities for feature selection and rule-based predictions, respectively.[14,17] These methods contribute to a more granular understanding of feature importance and the relationships between

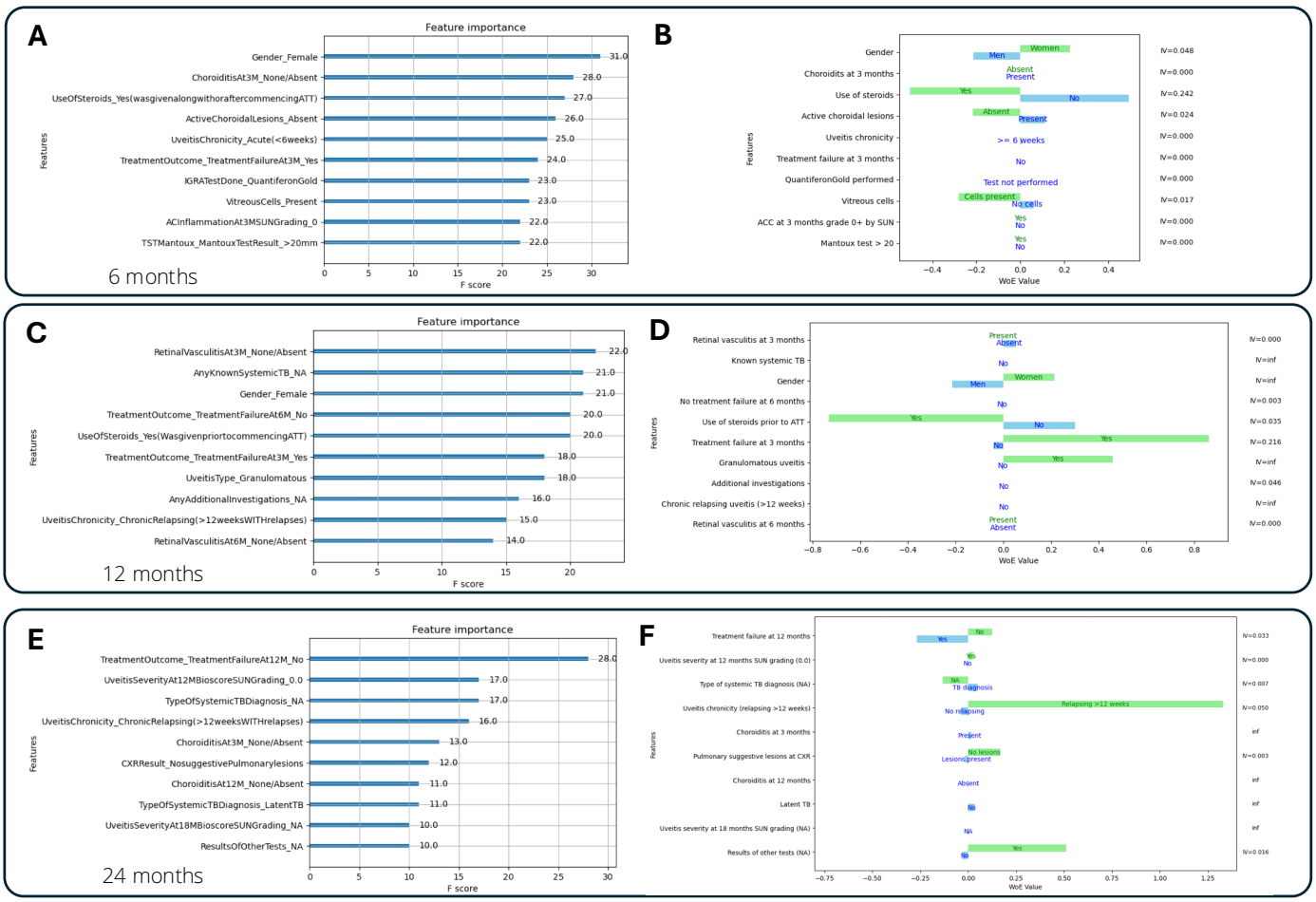

Figure 3. **Feature importance analysis for predicting treatment failure using XGBoost.** Each panel (**A–C**) corresponds to a different time point, illustrating the shift in feature importance over time. Notably, female gender and treatment failure in a previous visit are variables consistently important across the first two time points, while the absence of treatment failure at 12 months emerges as the most significant predictor at 24 months. This analysis employs explainable algorithms to extract these features, providing insights into factors that significantly impact the treatment outcomes in uveitis. **D–F** show the WoE and the IV for each feature.

variables, which is crucial for diagnosing and predicting outcomes in OTB.

XGBoost gave the best results or was remarkably close regarding AUC, but its complexity, counted as the number of nodes across trees, was among the highest. Other methods with low complexity, such as Ripper or CG, cannot compete in terms of AUC. DT and GLRM were quite similar, with DT always reaching better AUC. Random Forest was too complex with a similar AUC, whereas the TT-rules framework reached a similar AUC to DNN but with higher complexity than other ML methods. Variability of the predictive performance of models over time underscores the importance of dynamic modeling approaches that adapt to changes in patient condition and treatment responses. The significant drop in the positive class prevalence from 24.3% at six months to 9.57% at 24 months after treatment suggests a need for models that can adjust to varying degrees of class imbalance.

Some limitations should be acknowledged in the current study. First, the reliance on data from the COTS database, although comprehensive, introduces potential biases inherent to retrospective studies. These biases include selection bias, because patients included in the database may not represent the broader population affected by OTB, and information bias because of the variability in data collection methods across different centers. The dataset was predominantly composed of cases with posterior uveitis and choroiditis, with fewer anterior and intermediate uveitis cases, which may limit the generalizability of the findings to other uveitis subtypes.

The heterogeneity in data quality, particularly with clinical assessments, poses significant challenges. Variability in diagnostic criteria (i.e., high-resolution

computed tomography is not included in the baseline diagnostic workup, which may have reduced diagnostic precision), measurement techniques, and observer expertise can introduce bias and error into the dataset, potentially skewing the model outputs. The definition of treatment failure was assessed across 27 international centers, introducing potential interobserver variability that could affect consistency despite enhancing overall generalizability. This variability underscores the necessity for standardized protocols and training to ensure data quality and consistency.

Additionally, the models used, particularly those with higher complexity, such as XGBoost, although providing superior predictive accuracy, still present challenges regarding their deployment and interpretability in clinical practice. Attrition in patient numbers was notable, with nearly a 50% reduction by the 24-month follow-up compared to the six-month cohort, which could introduce bias in long-term prediction accuracy despite the strategies used to address missing data. Although efforts were made to use explainable models, the depth of explanation required to fully understand the decision-making process in complex models might still be beyond the grasp of routine clinical use. This could limit the practical deployment of such models in everyday clinical settings, where time constraints and the need for clear, straightforward explanations prevail. Moreover, the models' performance might be influenced by the uneven distribution of treatment outcomes across different time points, which reflects the varying stages of the disease and treatment response. This imbalance can skew the predictive performance and might affect the generalizability of the models to other populations or settings.

Although our model was developed using variables collected within the COTS study, which may not fully overlap with routinely available electronic health record data, its potential clinical utility lies in adapting these predictors into parameters that can be feasibly captured in real-world settings. Once harmonized with local data collection practices, the model could be embedded into clinical workflows through integration with electronic health record systems as a decision-support tool. Dynamic predictions generated at baseline and updated with follow-up data could alert clinicians to patients at higher risk of treatment failure, enabling timely adjustments in therapy. Such integration would facilitate real-time, individualized management and support evidence-based decision-making without adding significant burden to clinical practice. For example, consider a 32-year-old woman with presumed OTB presenting with posterior uveitis and active choroidal lesions. At baseline, our model classified her as *high risk* for treatment failure. During follow-up at three months, despite antitubercular therapy, she continued to require high-dose corticosteroids to control inflammation. The model dynamically updated her risk estimate, confirming persistent high risk of treatment failure. This prediction could prompt the clinician to escalate care earlier, such as initiating corticosteroid-sparing immunomodulatory therapy or closer monitoring, thereby potentially preventing long-term vision-threatening complications.

Conclusions

This study has established that ML models, particularly XGBoost and RF, possess robust predictive capabilities for treatment failure in OTB. These models were consistently superior in performance across various timeframes, although they initially required a greater operational complexity. By the 24-month follow-up, this complexity had been significantly mitigated, suggesting an improvement in model efficiency over time. The findings emphasize the need to select ML models carefully, balancing predictive performance with operational complexity, especially in clinical settings where both accuracy and explainability are critical to effective treatment planning. Further research should aim at improving the interpretability of these ML models, thereby not compromising their high predictive accuracy. Developing hybrid models that adapt to evolving clinical data while remaining transparent could significantly close the gap between clinical applicability and ML performance. Moreover, the consistent collection of high-quality data across diverse clinical environments remains essential. Future initiatives must prioritize refining data capture techniques and enhancing model interpretability.

Acknowledgments

The authors thank all COTS-1 study collaborators for their contributions to data collection. Beyond usual salary, no one received additional financial compensation for their contributions.

Supported by grants awarded by the National Medical Research Council (NMRC), Ministry of Health, Republic of Singapore grant number NRMC/CSAINV22jul-0004, NMRC/CSAINV19nov-0007, and NMRC/CIRG21nov-0023. The funders had

no role in study design, data collection and analysis, publication decisions, or manuscript preparation.

Disclosure: **W. Rojas-Carabali**, None; **T. Guérand**, None; **C. Cifuentes-González**, None; **J. Abisheganaden**, None; **P. R.K.**, None; **Y.C. Wei**, None; **G. Mejía-Salgado**, None; **A. de-la-Torre**, None; **J.R. Smith**, None; **J.H. Kempen**, None; **Q.D. Nguyen**, None; **C. Pavesio**, None; **B. Lee**, None; **V. Gupta**, None; **T. Peyrin**, None; **R. Agrawal**, None

* WRC and TG are co-first authors.

References

1. Barberis I, Bragazzi NL, Galluzzo L, Martini M. The history of tuberculosis: from the first historical records to the isolation of Koch's bacillus. *J Prev Med Hyg*. 2017;58:E9–E12.
2. Agrawal R, Gunasekeran DV, Grant R, et al. Clinical features and outcomes of patients with tubercular uveitis treated with antitubercular therapy in the Collaborative Ocular Tuberculosis Study (COTS)–1. *JAMA Ophthalmol*. 2017;135:1318–1327.
3. Agrawal R, Testi I, Bodaghi B, et al. Collaborative Ocular Tuberculosis Study Consensus Guidelines on the Management of Tubercular Uveitis—Report 2. *Ophthalmology*. 2021;128:277–287.
4. Bagcchi S. WHO's global tuberculosis report 2022. *Lancet Microbe*. 2023;4:e20.
5. Sekaggya-Wiltshire C, Von Braun A, Scherrer AU, et al. Anti-TB drug concentrations and drug-associated toxicities among TB/HIV-coinfected patients. *J Antimicrob Chemother*. 2017;72:1172–1177.
6. Zhang F, Zhang F, Li L, Pang Y. Clinical utilization of artificial intelligence in predicting therapeutic efficacy in pulmonary tuberculosis. *J Infect Public Health*. 2024;17:632–641.
7. Park H, Lo-Ciganic WH, Huang J, et al. Machine learning algorithms for predicting direct-acting antiviral treatment failure in chronic hepatitis C: an HCV-TARGET analysis. *Hepatology*. 2022;76:483–491.
8. Testi I, Agrawal R, Mahajan S, et al. Tubercular uveitis: nuggets from Collaborative Ocular Tuberculosis Study (COTS)-1. *Ocular Immunol Inflamm*. 2020;28(Suppl 1):8–16.
9. Collins GS, Moons KG, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378.
10. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825–2830.
11. Arya V, Bellamy RK, Chen PY, et al. One explanation does not fit all: a toolkit and taxonomy of AI explainability techniques. *arXiv*. Preprint posted online September 6, 2019. doi:10.48550/ARXIV.1909.03012.
12. Tjoa E, Guan C. A Survey on explainable artificial intelligence (XAI): toward medical XAI. *IEEE Trans Neural Netw Learning Syst*. 2021;32:4793–4813.
13. Ribeiro MT, 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*. New York: Association for Computing Machinery; 2016:1135–1144.
14. Dash S, Günlük O, Wei D. Boolean decision rules via column generation. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems* Red Hook, NY: Curran Associates Inc., 2018:4660–4670.
15. Quinlan JR. Induction of decision trees. *Mach Learn*. 1986;1:81–106.
16. Breiman L. Random forests. *Mach Learn*. 2001;45:5–32.
17. Wei D, Dash S, Gao T, Günlük O. Generalized linear rule models. In: International Conference on Machine Learning. New York: PMLR. 2019:6687–6696.
18. Cohen WW. Fast effective rule induction. In: *Proceedings of the Twelfth International Conference on International Conference on Machine Learning*. San Francisco: Morgan Kaufmann Publishers Inc.; 1995:115–123.
19. Benamira A, Guérand T, Peyrin T, Soegeng H. Neural Network-Based Rule Models With Truth Tables. *arXiv*. Preprint posted online September 18, 2023. doi.org/10.48550/ARXIV.2309.09638.
20. 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*. New York: Association for Computing Machinery; 2016:785–794.
21. Agrawal R, Gunasekeran DV, Agarwal A, et al. Visual morbidity in ocular tuberculosis – Collaborative Ocular Tuberculosis Study (COTS)-1: Report #6. *Ocular Immunol Inflamm*. 2020;28:49–57.
22. Rojas-Carabali W, Cifuentes-González C, Gutierrez-Sinisterra L, et al. Managing a patient

with uveitis in the era of artificial intelligence: current approaches, emerging trends, and future perspectives. *Asia Pac J Ophthalmol*. 2024;13(4):100082. doi:10.1016/j.apjo.2024.100082.

23. Agrawal R, Testi I, Mahajan S, et al. The Collaborative Ocular Tuberculosis Study (COTS) Consensus (CON) Group Meeting Proceedings. *Ocular Immunol Inflamm*. 2020;28:85–95.
24. Sauer CM, Sasson D, Paik KE, et al. Feature selection and prediction of treatment failure in tuberculosis. *PLoS ONE*. 2018;13:e0207491.
25. Kanesamoorthy K, Dissanayake M. Prediction of treatment failure of tuberculosis using support vector machine with genetic algorithm. *Int J Mycobacteriol*. 2021;10:279.
26. Reddy GP, Kumar YVP. Explainable AI (XAI): explained. In: *2023 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream)*. Piscataway, NJ: IEEE. 2023:1–6.
27. Xu Z, Guo K, Chu W, Lou J, Chen C. Performance of machine learning algorithms for predicting adverse outcomes in community-acquired pneumonia. *Front Bioeng Biotechnol*. 2022;10:903426.
28. Wang X, Zhu T, Xia M, et al. Predicting the prognosis of patients in the coronary care unit: a novel multi-category machine learning model using XGBoost. *Front Cardiovasc Med*. 2022;9:764629.