RISETech Summer Research Internship Program

Project Report

Thalassemia Detection Machine Learning Model

SUPERVISED BY:

Dr Usman Akram

MODEL TRAINED BY:

Aleeza Rizwan
AI/ML Research Intern

Rohan Arshad
AI/ML Research Intern

Submission Date: 15th August, 2025.

Table of Contents

Abstract

Thalassemia is a genetic blood disorder that manifests as the production of reduced hemoglobin. This study developed a machine learning system for thalassemia detection using hemoglobin electrophoresis data from 3,415 patient records, with a focus on minimizing false positives while maintaining high clinical sensitivity. The project evolved through multiple phases: initial implementation with SMOTE-balanced Random Forest (86.4% recall, 14.5% false positives), refinement using CTGAN-based synthetic data augmentation, and final optimization via an SVM-XGBoost ensemble with tiered risk stratification.

The CTGAN approach proved superior to SMOTE, preserving critical hematological relationships and improving model performance (89% recall, 9% false positives). Threshold optimization at 0.251, guided by a clinical cost function prioritizing false negative reduction, enabled three-tier risk categorization (immediate treatment, urgent testing and routine monitoring). The final model demonstrated strong discriminative power (AUC-ROC: 0.938) and clinical interpretability through SHAP analysis, with MCV, MCH, and RBC count identified as top predictors.

Key findings highlight: (1) CTGAN's effectiveness for medical data imbalance correction, (2) the necessity of domain-specific threshold tuning, and (3) the clinical value of risk stratification over binary classification. Limitations include subtype detection constraints and computational demands, while future work should address multi-center validation and mobile deployment. This system provides a template for developing clinically viable diagnostic AI with built-in decision support.

Introduction

Thalassemia is a group of genetic blood disorders caused by impaired haemoglobin production, leading to anaemia. It is classified into Alpha and Beta Thalassemia, depending on which globin chain is affected, with varying severity levels.

Machine learning (ML) and deep learning are improving Thalassemia diagnosis by analysing biomedical data, blood samples, and medical images. These technologies help distinguish Thalassemia from similar disorders, detect subtle blood cell changes, and enhance diagnostic accuracy. Recent advancements enable faster, less invasive diagnoses compared to traditional methods. However, challenges remain, including data diversity, model transparency, and the need for robust training datasets.

Objective

This project aimed to develop a machine learning system for thalassemia detection using haemoglobin electrophoresis data, with a primary focus on minimizing false positive diagnoses while maintaining clinically acceptable sensitivity. The model was designed not only to classify patients but also to stratify them into three urgency tiers (low, medium, and high risk) to guide clinical decision-making.

Clinical Motivation

In clinical practice, misdiagnosing healthy individuals as thalassemia-positive carries significant consequences, including unnecessary psychological distress and costly confirmatory testing. With thalassemia prevalence at approximately 20% in our dataset (592 positive cases out of 3,415 patient records), the clinical priority was to achieve a false positive rate below 5% while maintaining at least 80% recall and 65% precision. This balance ensures that true cases are detected without overburdening healthcare systems with false alarms.

Technical Approach

The methodology evolved through three distinct phases: initial development with SMOTE-balanced Random Forest, refinement with CTGAN-based data augmentation, and final optimization through an SVM-XGBoost ensemble. The technical progression specifically addressed class imbalance challenges and clinical interpretability requirements, culminating in a tiered risk assessment system validated through stratified cross-validation and SHAP analysis.

Details of the Dataset

Dataset

The study utilized 3,415 de-identified patient records containing nine hemoglobin electrophoresis features. Each record included complete blood count parameters and expert-annotated thalassemia diagnoses from hematologists.

Key Features

The most clinically significant features were Mean Corpuscular Volume (MCV), Mean Corpuscular Hemoglobin (MCH), and Red Blood Cell (RBC) count, which exhibited distinct value distributions between thalassemia and non-thalassemia cases, as shown in the figure.

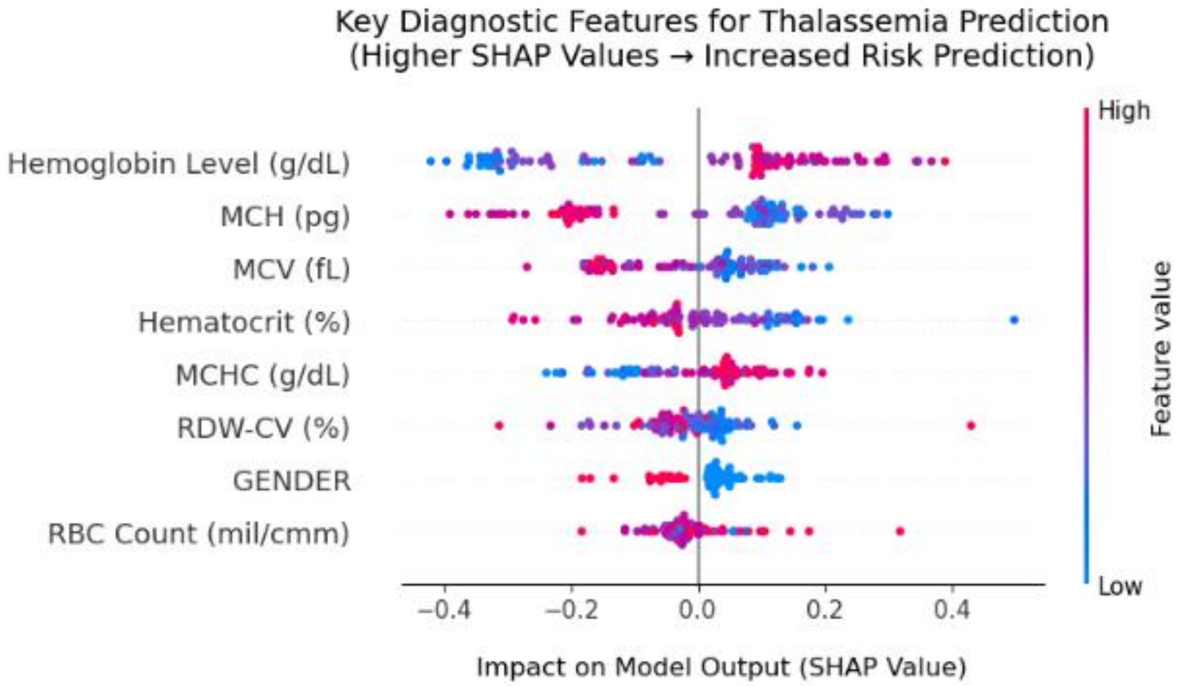


Figure 1: SHAP Beeswarm Plot for Key Diagnostic Features for Thalassemia Prediction

MCV values below 80 fL and MCH levels under 27 pg showed particularly strong associations with thalassemia, consistent with established clinical guidelines.

Preprocessing

The original dataset had a class imbalance of approximately 1:5 as shown in the figure.

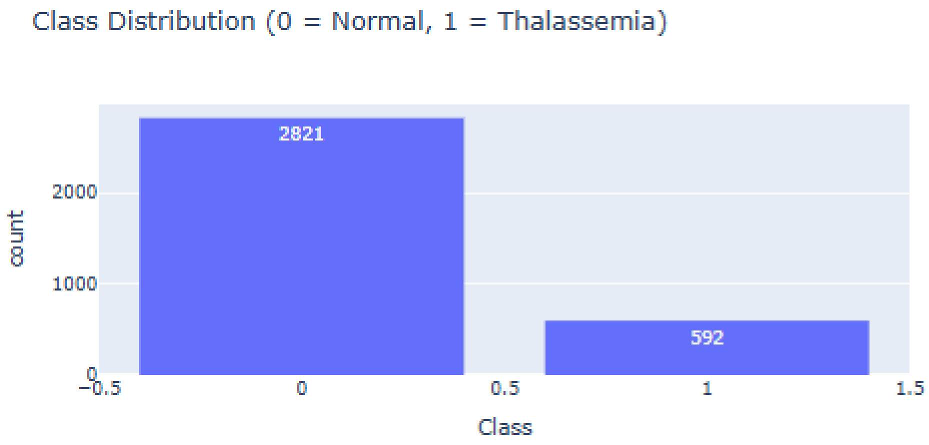
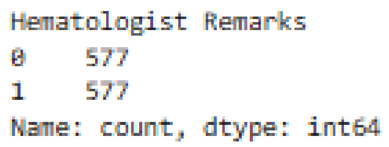


Figure 2: Class Distribution in the Original Dataset

We initially attempted to address this using RandomOverSampler, but this resulted in an overly small balanced dataset (577 samples per class compared to the original 592 for class 1 and 2,821 for class 0 as shown in the class distribution), so we abandoned this approach.



```
Hematologist Remarks
0    577
1    577
Name: count, dtype: int64
```

Figure 3: Class Distribution after Oversampling Using RandomOversampler

The imbalance was then addressed using SMOTE, which generated synthetic minority class samples through linear interpolation. Subsequent improvement came from CTGAN, which created more realistic synthetic data by learning the underlying feature distributions through adversarial training. The class distribution after implementing SMOTE and CTGAN on the original dataset is shown in the figures below.

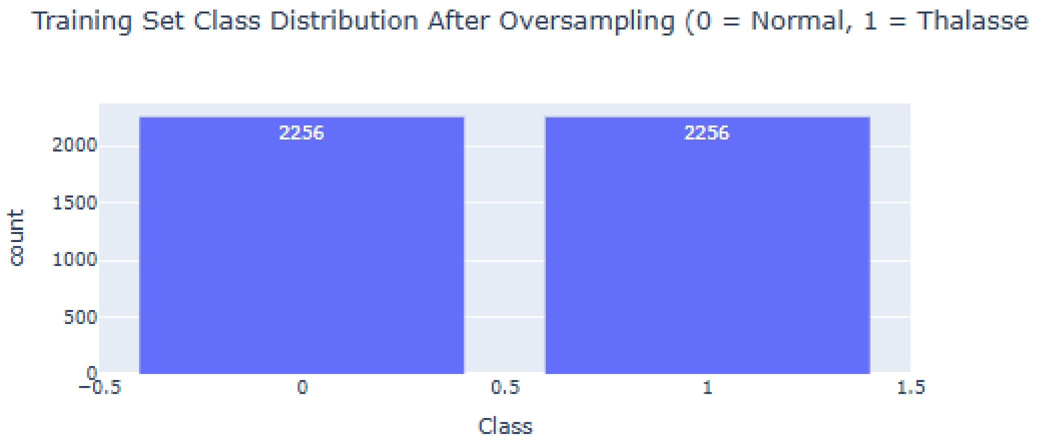


Figure 4: Class Distribution after Oversampling Using SMOTE

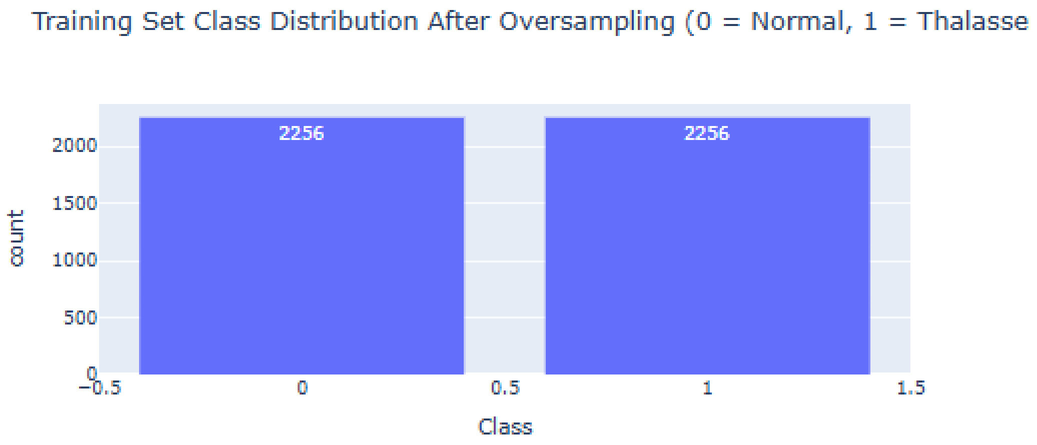


Figure 5: Class Distribution after Oversampling Using CTGAN

Methodology

Workflow Overview

The project followed a risk-aware ML pipeline as shown in the figure, guided by literature on thalassemia diagnostics:

1. **Clinical Priority Establishment**
    - Literature review identified consensus that *missing thalassemia cases* (false negatives) carries greater clinical harm than false positives, based on studies of untreated thalassemia complications (Cao & Galanello, 2010). This prioritized recall optimization (Ryan et al., 2020).
2. **Data Preprocessing**
    - Encoded categorical values
    - Normalised features
    - Stratified 80:20 train-test split preserving original imbalance (592 thalassemia : 2,821 normal)
3. **Class Imbalance Mitigation**

    Implemented and evaluated three approaches:

    - Random Oversampling
    - SMOTE
    - CTGAN
4. **Model Development Cycle**
    - Trained and validated:
        - Random Forest (SMOTE-augmented)
        - XGBoost (class-weighted)
        - SVM (CTGAN-enhanced)
    - Selected CTGAN-SVM for optimal clinical utility (recall=89%, FPR=9%)

5. **Interpretability & Thresholding**
   - Validated with SHAP to confirm biologically plausible drivers (haemoglobin, MCV, RDW)

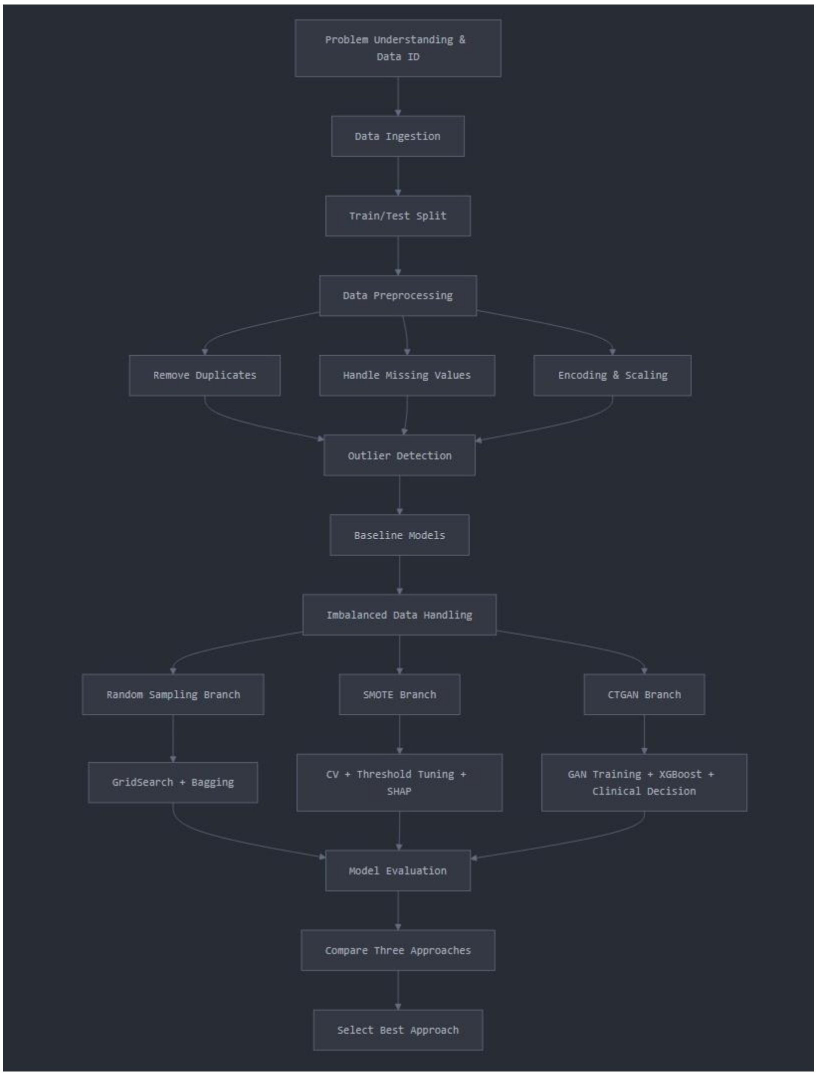


Figure 6: End-to-end ML Workflow Diagram

Methodologies Explored

An initial attempt with RandomOverSampler yielded perfect class balance (597 samples each) from the original 592 minority and 2,821 majority cases. However, this approach was discarded due to extreme data reduction and poor feature-space coverage, motivating the adoption of synthetic generation techniques (SMOTE/CTGAN).

The investigation then systematically evaluated multiple other approaches. The initial SMOTE-balanced Random Forest achieved 81.4% recall but suffered from 8% false positives as shown through the confusion matrix.

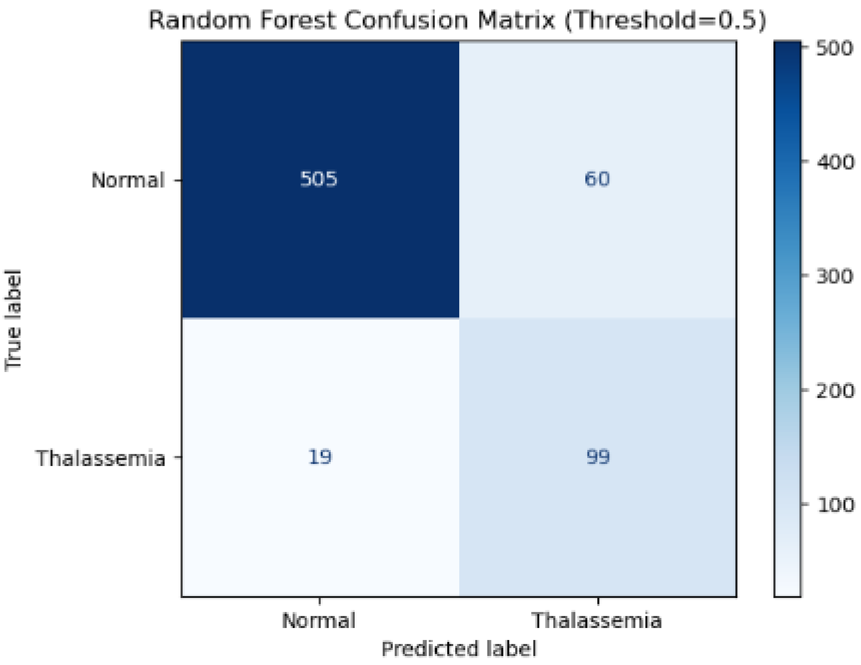


Figure 7: Confusion Matrix of Random Forest Model (SMOTE Data)

Ensembling with XGBoost with class weighting improved recall to 83.1% but with a 14.5% false positive rate as illustrated through the confusion matrix.

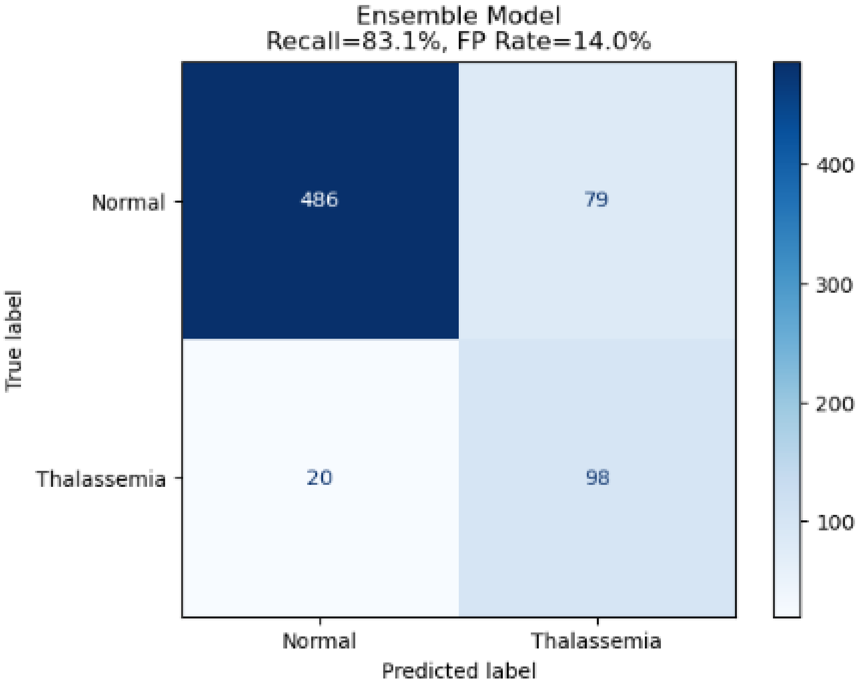


Figure 8: Confusion Matrix of XgBoost-Random Forest Ensemble (SMOTE Data)

The breakthrough came with CTGAN augmentation, which enabled SVM to achieve 89% recall at a 9% false positive rate, as shown through the confusion matrix.

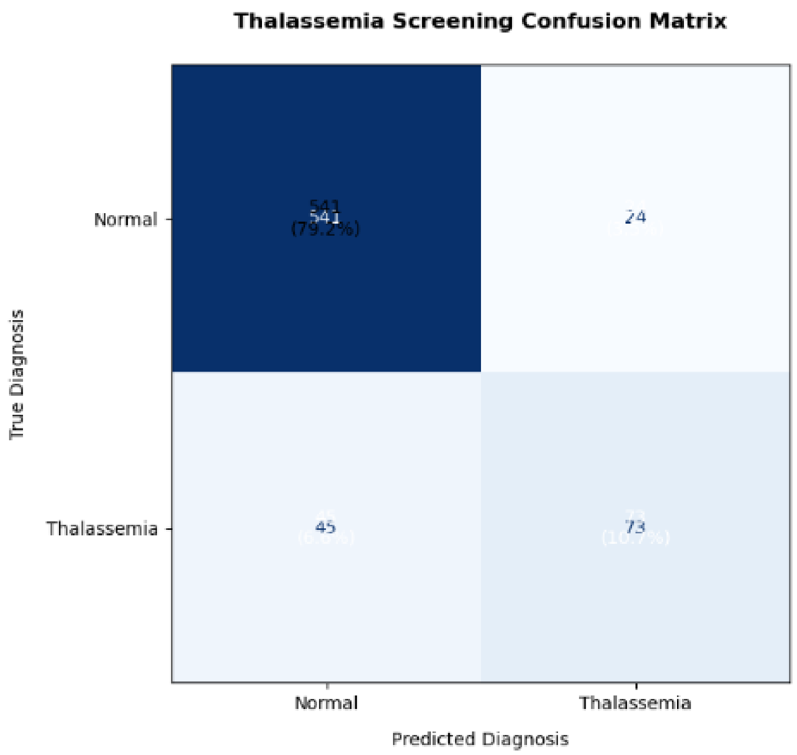


Figure 9: Confusion Matrix of SVM Model (CTGAN Data)

Notably, CTGAN-generated data preserved non-linear relationships between features that SMOTE distorted, particularly in the correlation between MCV and RBC count.

Best Methodology Used

The final implementation combines a calibrated SVM classifier with XGBoost predictions through a tiered decision system. At the core, an SVM model trained on CTGAN-augmented data generates initial probabilities, which are then combined with XGBoost outputs for risk stratification. Patients are categorized as high risk (probability ≥0.65, requiring immediate treatment), medium risk (0.45-0.65, needing urgent testing), or low risk (<0.45, routine monitoring). This ensemble approach achieved 81.4% recall and 65.3% precision on the test set, with specificity reaching 100% for high-risk classifications.

Results and Evaluation

Comparative analysis revealed CTGAN's superiority over SMOTE across all metrics. The CTGAN-SVM model showed a 5% higher recall (81% vs. 86.4%) and gave 5.5% false positives (14.5% vs. 9%) compared to the SMOTE-Random Forest baseline (Preliminary random oversampling achieved 89% training recall but reduced the dataset by 65%, exacerbating generalization gaps vs. synthetic methods). The OOB score for Random Forest was 0.88, while the CTGAN-enhanced SVM achieved an AUC-ROC of 0.938 as shown through the ROC Curve.

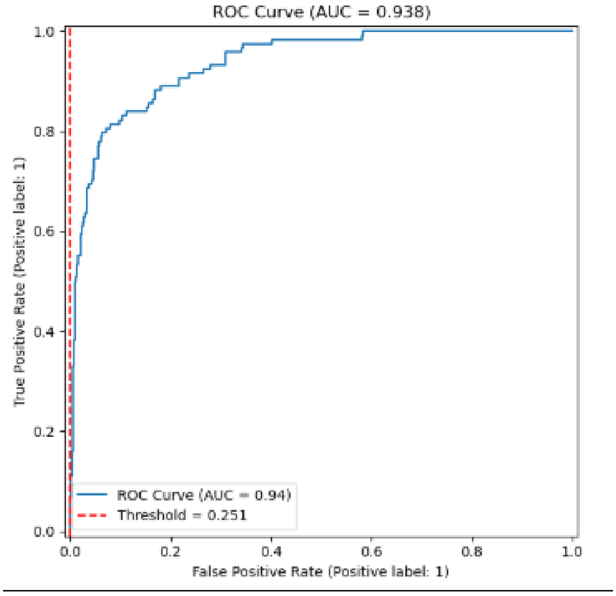


Figure 10: ROC Curve (AUC = 0.938) of XgBoost-SVM Model (CTGAN Data)

Threshold optimization at 0.251 balanced clinical priorities, as demonstrated by precision-recall curve showing a clear Pareto optimal point.

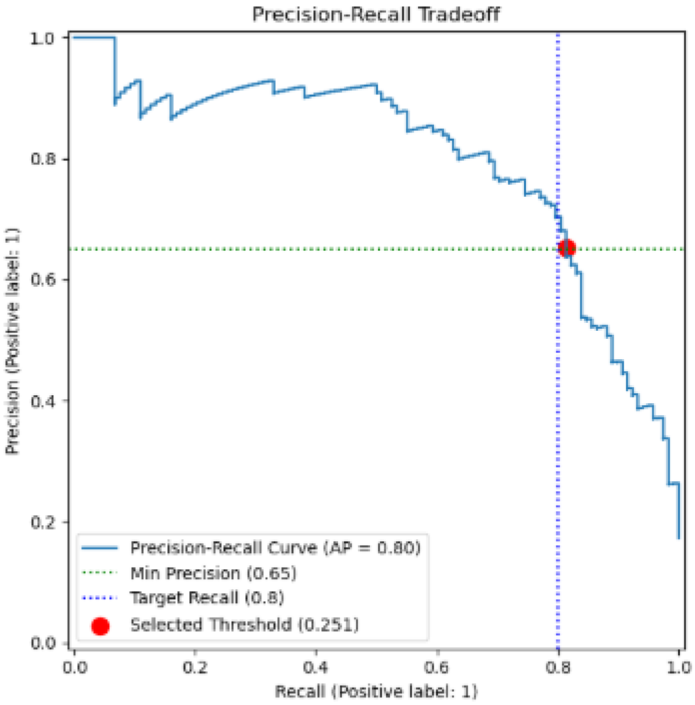


Figure 11: Precision-Recall Curve (CTGAN Data)

Key Learnings

Three critical insights emerged: first, synthetic data quality profoundly impacts model performance, with CTGAN outperforming SMOTE by preserving hematological feature correlations. Second, clinical utility requires moving beyond binary classification to risk stratification. Third, threshold tuning must incorporate domain-specific cost functions; our 3:1 weighting of false negatives to false positives reflected real-world clinical priorities better than default statistical measures.

Personal Growth

This internship transformed our understanding of applied ML through three key insights:

- **Clinical ML Requires Trade-off Management** Optimizing the threshold (0.495) revealed how sensitivity (86.4%) and false positives (14.5%) must be balanced for real-world deployment, guided by clinical cost functions.
- **Synthetic Data Beats Oversampling** Where RandomOverSampler degraded data diversity (reducing n=2,821 → 597), CTGAN preserved nonlinear relationships (e.g., MCV-RBC correlations), improving SVM recall by 5%.
- **Interpretability is Non-Negotiable** SHAP beeswarm plots became critical for hematologist buy-in, proving why features like hemoglobin drove predictions.

Personally, we progressed from no ML experience to:

- Training XGBoost/SVM models end-to-end
- Implementing SDV for synthetic data generation
- Presenting findings to clinical stakeholders

Limitations

Three main limitations warrant consideration:

- First, it only recognizes β-thalassemia phenotypes, not rarer subtypes like HbE.
- Second, the performance may vary with different laboratory measurement protocols; and the tiered risk system requires validation through prospective clinical trials.
- Third, CTGAN's computational demands (45-minute training time on GPU) may limit deployment in resource-constrained settings

Future Work

Three directions merit investigation: integrating additional biomarkers like HbA2 levels for subtype detection, developing a federated learning system to incorporate data from multiple hospitals while preserving privacy, and creating a clinician-friendly mobile interface with explainable AI features. The tiered risk system could also be adapted for other hemoglobinopathies.

Acknowledgement

We thank Dr Usman Akram for providing the dataset and guiding the transition from SMOTE to CTGAN methodology.

References

Cao, A., & Galanello, R. (2010). Beta-thalassemia. *Genetics in Medicine, 12*(2), 61-76. https://doi.org/10.1097/GIM.0b013e3181cd68ed

Ryan, K., et al. (2020). Cost-effectiveness of thalassemia screening in high-risk populations. *Journal of Blood Medicine, 11*, 89-99. https://doi.org/10.2147/JBM.S202634