廣州醫科大學附屬第三醫院
The Third Affiliated Hospital of Guangzhou Medical University

Superior Performance of a Novel Thalassemia Screening Model Over Traditional Indices Using Automated Hematology Analyzers

Ping Yin, Jialin Tan, Honghai Hong

Department of Clinical Laboratory, Third Affiliated Hospital of Guangzhou Medical University, Guangzhou, China

✉*dymunacorp@gmail.com*

INTRODUCTION

Differentiating thalassemia carriers from individuals with other anemias - particularly microcytic types like iron deficiency - is a critical yet challenging task in laboratory hematology. Conventional screening indices, such as MCV, MCH, and the Mentzer index, are limited by suboptimal sensitivity and specificity. This study aimed to develop and validate an advanced analytical model using automated hematology analyzer data to overcome these limitations and provide a more accurate tool for thalassemia screening.

METHODS

A retrospective cohort of 968 participants was categorized into three groups:

1. **Thalassemia Group (n=442)**: Genetically confirmed, including α-thalassemia trait (n=160), α-thalassemia intermedia (n=7), β-thalassemia (n=103), and unspecified genotypes (n=172).
2. **Non-thalassemia Anemia Group (n=378)**: Included iron deficiency anemia (n=13), anemia of chronic disease (n=324), and hemolytic anemia (n=41).
3. **Healthy Controls (n=148)**.

Complete blood count and extended research-use parameters from a **DH-800CS automated hematology analyzer (Dymind Biotechnology)** were used. An advanced classification algorithm was developed and validated using a 7:3 stratified split and five-fold cross-validation. SHapley Additive exPlanations (SHAP) analysis was employed to interpret the predictive model.

Dymind DH-800CS

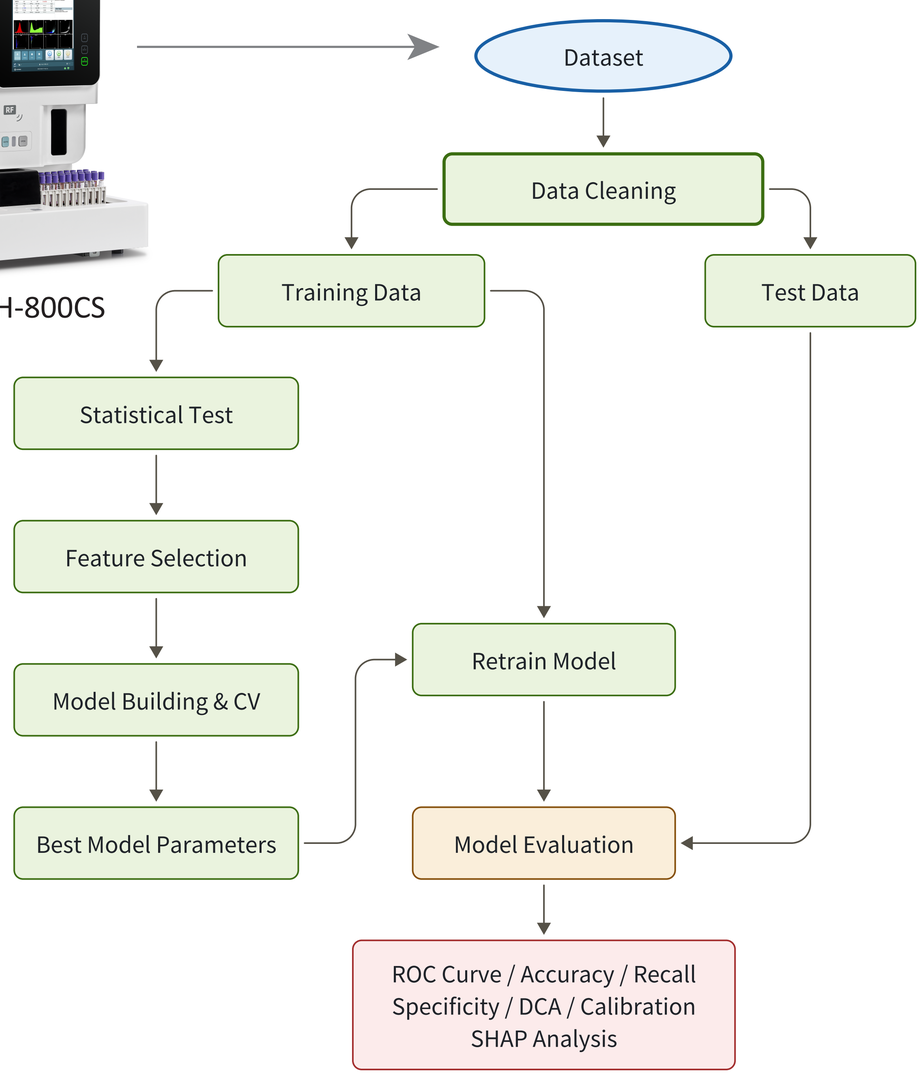

Figure 1: Scheme of model building. ROC: Receiver Operating Characteristic; DCA: decision curve analysis

RESULTS

Seven machine-learning classifiers - including Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), eXtreme Gradient Boosting (XGBoost), Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP) - were trained on a set of 12 selected variables. The best-performing model (Gradient Boosting) demonstrated high accuracy on an independent test set (AUC: 0.955, Sensitivity: 97.7%, Specificity: 88.6%)(Figure 2A,B). Its performance markedly exceeded that of the Mentzer index (Sensitivity: 42.7%, Specificity: 73.1%) and the conventional MCV/MCH rule (Sensitivity: 73.8%, Specificity: 37.6%)(Figure 2B). SHAP analysis identified the most influential predictive features as mean corpuscular volume (MCV), mean corpuscular hemoglobin (MCH), microcyte percentage (Micro%), and red blood cell count (RBC).

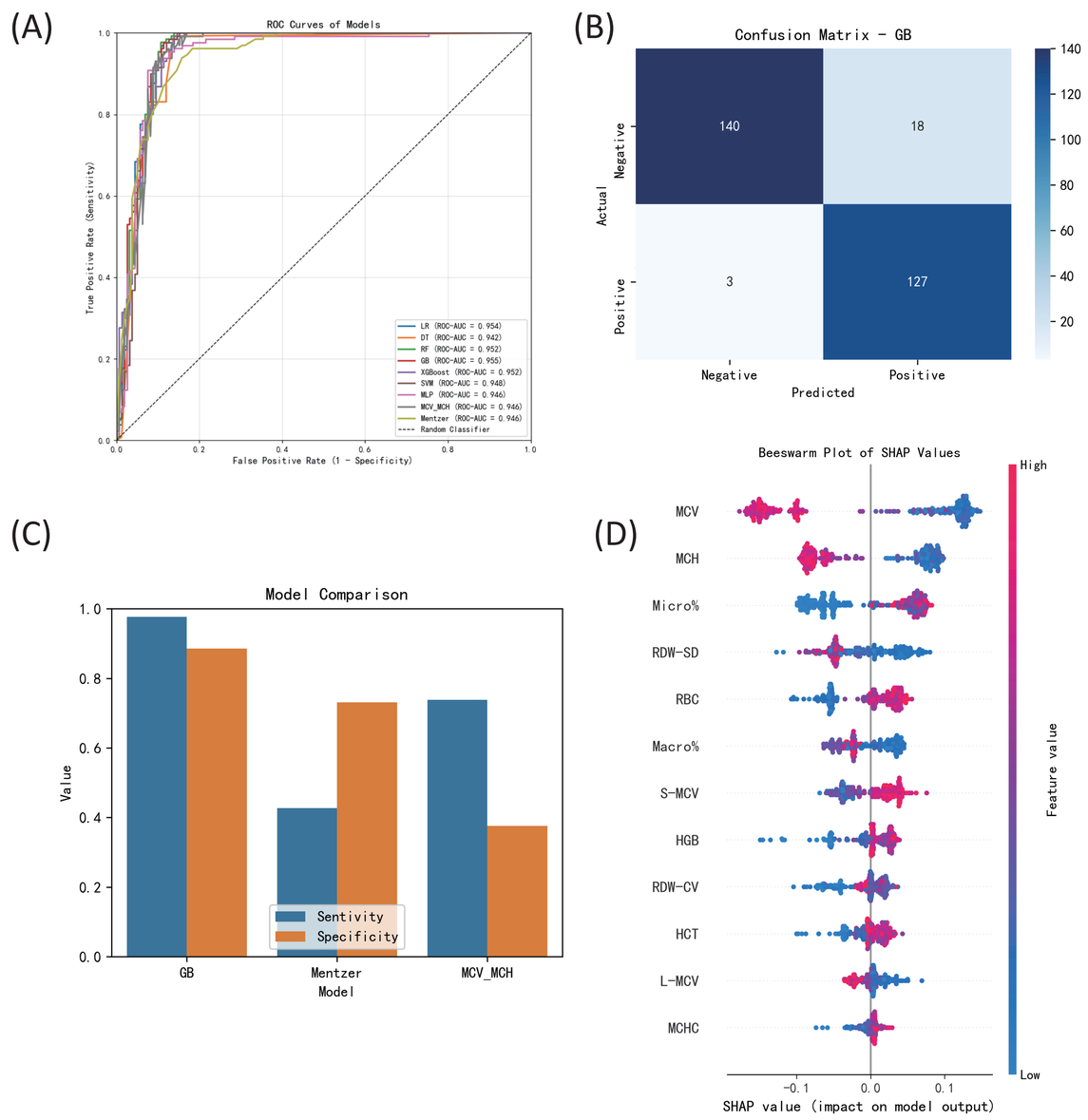

Figure 2: (A) ROC curves for screening thalassemia patients in the validation cohort. (B) Confusion matrix of the RF model. (C) Performance comparison of the GB model with MCV/MCH rule and the Mentzer Index. (D) Results of SHAP analysis.

CONCLISION

This high-performance screening model provides a substantial advancement for the clinical laboratory, enabling significantly more accurate identification of thalassemia carriers and their differentiation from other anemias directly from routine CBC data. Its implementation can streamline the analytical workflow, enhance the specificity of screening reports, and improve overall efficiency in the hematology laboratory.