16.04.2025

Çağla Çağlar

Classification of Blood-Based Biomarkers: A Machine Learning Approach for Assessing HDL, LDL Cholesterol and Hemoglobin Levels Using Absorption Spectroscopy

Abstract

This study investigates the application of machine learning methods for classifying blood analytes including HDL cholesterol, LDL cholesterol, and hemoglobin (HGB), based on absorption spectra. The analysis was performed using a publicly available dataset provided by the Blood Spectroscopy Classification Challenge hosted on the Zindi Africa platform. Exploratory data analysis, data preprocessing, dimensionality reduction, model training, and evaluation steps were executed. Logistic Regression (LR) and XGBoost algorithms were compared in terms of performance. XGBoost models showed higher predictive accuracy as evidenced by statistical metrics including ROC AUC scores, confusion matrices, classification reports, and SHAP analysis.

Introduction and Dataset Description

The dataset analyzed was obtained from the Blood Spectroscopy Classification Challenge on the Zindi Africa platform (available at: *https://zindi.africa/competitions/bloodsai-blood-spectroscopy-classification-challenge/data*), and contains 13,140 samples. Each sample includes 170 absorbance measurements labeled absorbance0 to absorbance169, representing the spectral intensity response of blood to illumination. Additionally, each sample has measurements of temperature, humidity, and a unique identifier. The dataset defines three target analytes: HDL cholesterol, LDL cholesterol, and hemoglobin (HGB), categorized as low, ok, or high. Trimmed versions of datasets were provided to remove noisy spectral edges but were not used in this analysis.

Exploratory Data Analysis

Initial exploratory data analysis began with the standardization of spectral absorbance features. Principal Component Analysis (PCA) was conducted to assess data dimensionality. PCA results indicated that the first 49 principal components accounted for approximately 99.93% of the variance in the data, indicating significant redundancy. PCA scatter plots visualized clustering by analyte categories, showing clear differentiation. Analysis of mean absorbance spectra by analyte groups also displayed distinct patterns across different classes.

Data Cleaning and Feature Engineering

The preprocessing stage involved standardizing column names for uniformity. No missing values were present in the dataset, indicating robust initial data collection. The spectral absorbance data were standardized using StandardScaler. Outlier detection using the interquartile range (IQR) method identified numerous outliers, primarily concentrated in the

absorbance features with lower indices. These outliers were intentionally retained to maintain biologically relevant variability.

Model Training, Validation, and Evaluation

Two machine learning algorithms were evaluated: Logistic Regression and XGBoost. Logistic Regression used PCA-reduced data consisting of 49 principal components, with hyperparameters optimized through RandomizedSearchCV. Logistic Regression exhibited weak predictive performance, notably yielding ROC AUC scores around 0.5 for LDL cholesterol, indicating performance similar to random guessing. The HDL cholesterol and hemoglobin predictions were slightly better but insufficient for clinical diagnostics.

In contrast, XGBoost models trained on the full set of standardized spectral absorbance data without PCA dimensionality reduction displayed excellent performance. Extensive hyperparameter tuning and cross-validation confirmed the robustness of XGBoost models, achieving high ROC AUC scores: 0.9866 for HDL cholesterol, 0.9883 for LDL cholesterol, and 0.9868 for hemoglobin. Detailed classification reports and confusion matrices supported the superior predictive accuracy of XGBoost.

Model Interpretability and SHAP Analysis

Model interpretability was enhanced through the computation of SHapley Additive exPlanations (SHAP) values. SHAP analysis identified specific spectral features that significantly contributed to the classification outcomes, particularly distinguishing high-risk categories such as low HDL cholesterol, high LDL cholesterol, and low hemoglobin levels.

Conclusions and Future Directions

This study confirms the effectiveness of spectral absorbance data combined with XGBoost modeling for classifying HDL cholesterol, LDL cholesterol, and hemoglobin levels. The high predictive performance observed across all analytes, together with the model's interpretability through SHAP analysis, highlights its potential for clinical application. Although temperature and humidity measurements were available, these variables were not included in the current modeling framework.

Future studies may explore the impact of using trimmed spectral datasets, which aim to reduce noise by excluding the edge regions of the spectra. Additionally, validating the model on external and independent datasets would strengthen the generalizability of the findings. Further methodological enhancements could include the application of deep learning architectures tailored for spectral data, the implementation of advanced hyperparameter optimization techniques such as Bayesian search or Optuna, and the use of alternative feature importance frameworks beyond SHAP to capture model-agnostic interpretability.

Moreover, ensemble approaches that integrate complementary model predictions and interaction-aware SHAP analyses may provide further insights into nonlinear spectral patterns associated with high-risk biomarker levels. The supplementary dataset initially provided (Zindi_Contest_Spectra.xlsx) could not be incorporated into the present analysis due to inaccessibility at the time of processing.