

Original Article

Subgroup-Aware Evaluation and Interpretability in Myopia Prediction using Deep ROC and XAI

Subrina Akter¹, Rahasina Hossain², Kaniz Fatema Tanni³, Samsun Nahar Wafia⁴, Sumayea Benta Hasan⁵,
Abdullah Man⁶, Nor Hidayati Abdul Aziz⁷

^{1,2,3,4}Department of Computer Science and Engineering, International Islamic University Chittagong, Chattogram, Bangladesh.

¹Department of Computer Science and Engineering, Chittagong University of Engineering and Technology, Chattogram, Bangladesh.

³M.Sc. Artificial Intelligence, Brandenburgische Technische Universität Cottbus-Senftenberg, Cottbus, Germany.

^{5,7}Faculty of Engineering and Technology, Multimedia University, Melaka, Malaysia.

⁶Center of Robotic and Sensing Technology, TM R&D, Multimedia University, Cyberjaya, Malaysia.

⁷Centre for Advanced Analytics, CoE for Artificial Intelligence, Multimedia University, Melaka, Malaysia.

⁷Corresponding Author : hidayati.aziz@mmu.edu.my

Received: 26 February 2026

Revised: 11 July 2026

Accepted: 29 August 2026

Published: 27 October 2026

Abstract - Machine Learning (ML) models are familiar for their great success in classification challenges within different domains. For medical diagnostics, those models have numerous applications as well. Basically, these ML models generate good predictions. However, sometimes, for clinicians, they are unable to depend on and use those models in a timely manner while making clinical decisions. Moreover, this factor arises because of the "black box" nature of these models. In medicine, it is a matter of concern. Because it is crucial to understand the base for predictions of patients' safety and treatment assurance. Besides, recording key performance differences between risk groups is significant for informed clinical decision-making. In that case, traditional measures like the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) show some limitations. Hence, to defeat such matters, this study proposes a unified framework for experimenting with Myopia prediction (known as nearsightedness) and evaluating the results using some advanced techniques. This approach integrates Deep ROC analysis with Explainable AI (XAI) methods, where for XAI, Local Interpretable Model-agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP) are applied. Here, the Deep ROC analysis is more refined and has clinical relevance, as the model performance is generated within subgroups defined by False Positive Rate (FPR). On the other hand, it allows for an in-depth understanding of model behavior in high-risk, medium-risk, and low-risk patient populations. Meanwhile, LIME and SHAP offer XAI value by explaining how individual attributes contribute to model output. While doing myopia prediction, this proposed evaluation structure is applied to the machine learning models, named Support Vector Machine (SVM), Random Forest (RF), Multi-Layer Perceptron (MLP), and Decision Tree (DT). Firstly, these models are trained and then evaluated to ensure robustness and reduce bias. After the evaluation, Random Forest (RF) outperforms the other three models across all subgroups, with an AUC of 98.92% in high-risk data 99.49% in low-risk data. The Brier score further confirmed the reliability of the model and improved its predictive power. The union of Deep ROC and XAI techniques significantly enhances the explainability and clinical utility of RF, with proof of how it forecasts for different risk levels. The approach is a leap forward towards the production of explainable, accurate, and clinically valuable AI models in medicine.

Keywords - Deep Roc Analysis, Explainable AI (XAI), Myopia prediction, Machine Learning, Model interpretability.

1. Introduction

New advances in Machine Learning (ML) have greatly improved its application to medical diagnosis and disease categorization. However, most of these models produce accurate results without providing an understanding of how the decisions are reached. This lack of transparency is a critical issue in the healthcare domain. Because the doctors and medical practitioners need clear and trusted explanations behind predictions to use them safely in the treatment plan [1].

The use of ML models in clinical settings is not easy without the willingness. Besides, ML, with its usual model performance metrics, faces healthcare challenges [2]. Common measures, such as the Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC), provide a good performance. However, they do not offer detailed information on the model's performance for various patient groups or risk levels [3]. Basically, this limitation is important since missing a high-risk disease (false negative)



will have serious consequences. Here, misclassifying a low-risk patient (false positive) will not certainly be dangerous, but possibly cause unreasonable treatment or anxiety [4]. In addition, to improve model transparency and trustworthiness, Explainable Artificial Intelligence (XAI) techniques, such as LIME and SHAP, are being widely used in medical prediction systems [32]. However, these approaches are mostly applied separately from subgroup-aware evaluation strategies. Existing research mainly focuses either on predictive performance or on model interpretability. However, very little attention has been given to how prediction reliability changes among clinically relevant patient subgroups. Therefore, an elaborate and more advanced evaluation of model performance is essential to guiding well-informed clinical decisions. As a case study, this study applies the proposed framework to the task of myopia detection. The word Myopia refers to short-sightedness, an international health issue, which can be seen especially in children and youth [5-12]. The popularity of myopia has risen considerably over the past few decades, specifically in East and Southeast Asia. Its outgrowth suggests that by 2050, half of the world's population could be myopic [11]. Severe myopia is related to a higher risk of developing serious eye conditions such as myopic macular degeneration, cataracts, retinal detachment, and even blindness [13, 14]. So, the necessity of early detection and intervention is critical in preventing these complications [15].

In recent years, the authors have applied ML to predict myopia, focusing on models such as Support Vector Machines (SVM), Decision Trees, Neural Networks, and Ensemble Methods [16-21]. For example, Rauf et al. [16] investigated 400 labeled training images with the mobile version of the VGG16 model and demonstrated performance by achieving an AUC score of 0.9845. Similarly, Noori et al. [22] mentioned that a VGG16-based model scored 96% accuracy rate in diagnosing myopia cases. Here, LIME was used to explain classification decisions. On the other side, Li et al. [17] developed a Random Forest model by using a large longitudinal dataset to predict myopia progression in children. Though it identified uncorrected distance visual acuity, spherical equivalents, and parents' myopia as significant predictors. Prashar et al. [23] did a meta-analysis on AI-based pathological myopia diagnosis by using fundus images. It reported excellent discriminative performance, with convolutional neural networks which outperforms SVM-based models. Other studies have emphasized the function of environmental and behavioral components in myopia development. Like Grzybowski et al. [24] highlighted, limited physical activity and excessive near work are major contributors to myopia in school-aged children. Moreover, Ba et al. [25] further demonstrated that genetic tendency magnifies the effects of an unhealthy lifestyle. Besides, it is recommended to increase outdoor activity as a preventive measure. Recent deep learning research has concentrated heavily on retinal image-based detection of pathological

myopia. Lu et al. [26] proposed deep learning models for detecting pathological myopia and classifying myopic maculopathy. It demonstrated strong real-world performance. In their extended work, Lu et al. [27] reported AUC values that exceeded 0.97, along with high accuracy, sensitivity, and specificity across multiple classification tasks. Next, Guo et al. [28] proposed a MobileNet version 2 with a transfer learning approach for differentiating multiple eye diseases with limited training data. A deep learning platform was trained by Cen et al. [29] over 249,000 fundus images by achieving an AUC of 99.84% for multi-label classification.

In addition, Li et al. [30] designed a self-supervised learning approach for retinal disease classification, which reduced dependence on labeled data. Besides, a T-LSTM model was introduced by Huang et al. [31] to predict long-term changes in refraction and visual acuity for early myopia management. Some recent literature has investigated the use of ML and XAI methods for myopia prediction. Specifically, Zou et al. [42] proposed an interpretable XGBoost model utilizing retinal structural metrics, in which the SHAP analysis provided critical information about retinal biomarkers that are highly correlated with a high risk of myopia.

Similarly, Sifat et al. [44] employed SHAP interpretability to extract risk factors associated with myopia, such as family history, screen time, and outdoor physical activities, along with the use of ML models by Feng et al. [43] on ocular biometric and refractive parameters. Although previous studies reported strong predictive performance using machine learning and deep learning approaches, some important limitations still remain. Most studies have mainly focused on improving prediction accuracy using retinal images, clinical measurements, or behavioral data. However, less attention has been given to how model performance varies among different patient risk groups. More importantly, research that integrates fairness-oriented subgroup evaluation and explainable model interpretation together in myopia prediction is still very limited.

To overcome these gaps, this study proposes a unified framework for myopia prediction. The novelty of this study is the integration of subgroup-aware Deep ROC evaluation and XAI techniques within a single unified framework for myopia prediction. Compared with Conventional ROC-based evaluation, DRA can perform subgroup-aware assessment because it analyzes prediction performance according to False Positive Rate (FPR)-based risk category [33]. These categories include high-risk, medium-risk, and low-risk subgroups. Moreover, XAI techniques were applied separately to the three subgroups to identify and compare the influential features driving model predictions across different risk categories. This subgroup-level analysis provides more clinically meaningful insight into the model's behavior and supports informed medical decision-making.

Mainly, this research contributes to the following:

- Deep ROC analysis to assess model performance, which offers better evaluation details than traditional ROC curves.
- Subgroup-specific explainability analysis to identify influential clinical, behavioral, and genetic predictors through XAI analysis.
- Clinical validation of the proposed Deep ROC risk groups through statistical significance testing.

2. Methods

This research proposes an evaluation approach in predicting myopia by ML models by using Deep ROC Analysis and Explainable AI (XAI) techniques.

Initially, data preprocessing has been completed by handling outliers and normalizing features to ensure the model's consistency and efficiency. Then the dataset is separated into 80% and 20% for training and testing, respectively. It ensures enough data for model learning while maintaining a set for unbiased evaluation. For further improvement, 10-fold cross-validation is applied over the training dataset. After that, to handle class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) is employed on the training set in generating synthetic samples for the minority class.

Afterwards, Support Vector Machine (SVM), Random Forest, Multi-Layer Perception (MLP), and Decision Tree are trained for Myopia prediction. Then, to evaluate these four models' traditional metrics like accuracy, precision, recall, F1-score, G-mean and AUC are calculated. Alongside pre-test and post-test measures are determined for Deep ROC Analysis. Basically, this is assessed for performance evaluation across different subgroups. Finally, to increase model explainability, LIME and SHAP are applied. The whole workflow diagram for the proposed methodology is illustrated in Figure 1. Besides, the algorithm for this experiment has been mentioned later in Algorithm 1.

2.1. Problem Formulation

To evaluate the machine learning models' performance, the widely used method is traditional Receiver Operating Characteristic (ROC) analysis, specifically in binary classification tasks. However, they fail to report variations in performance across different risk groups, whenever ROC curves provide a global assessment of model performance over all decision thresholds. Within the healthcare domain, the significance of false positives and false negatives can vary notably, subject to the patient's risk level. For example, a false positive in a high-risk patient might lead to unnecessary treatments. Meanwhile, a false negative in the same group could result in missed diagnoses. Both these situations can worsen the patient's situation. So, this study proposes Deep ROC Analysis (DRA), a more granular and clinically relevant

evaluation framework. Here, it is built on traditional ROC analysis by dividing the ROC curve into risk-based subgroups and categorizing according to the False Positive Rate (FPR) thresholds.

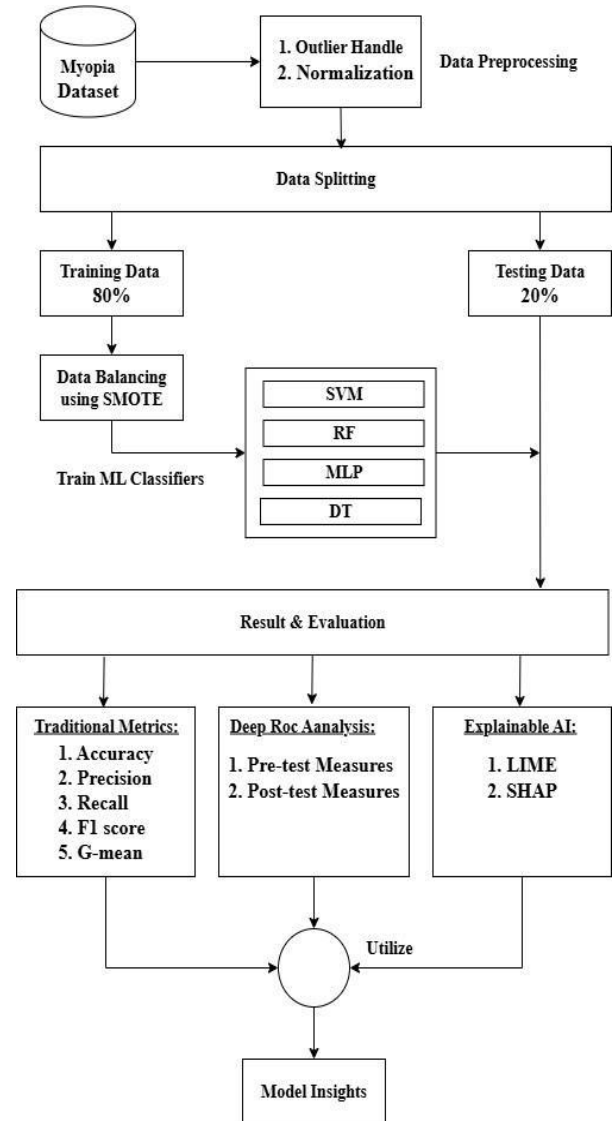


Fig. 1 Overall workflow of the proposed work

Normally, these subgroups include high-risk, medium-risk, and low-risk categories. Hence, by evaluating models at this level of granularity, DRA serves more actionable insights, particularly in clinical judgement and patient outcomes.

2.2. Illustrative Examples

This section tries to express the comparison between the traditional ROC Curve and Deep ROC analysis in an illustrative way. Mainly, the target is to represent the limitations of traditional ROC analysis. So, firstly, a small and simple dataset (toy dataset) is used. And the outcome shows.

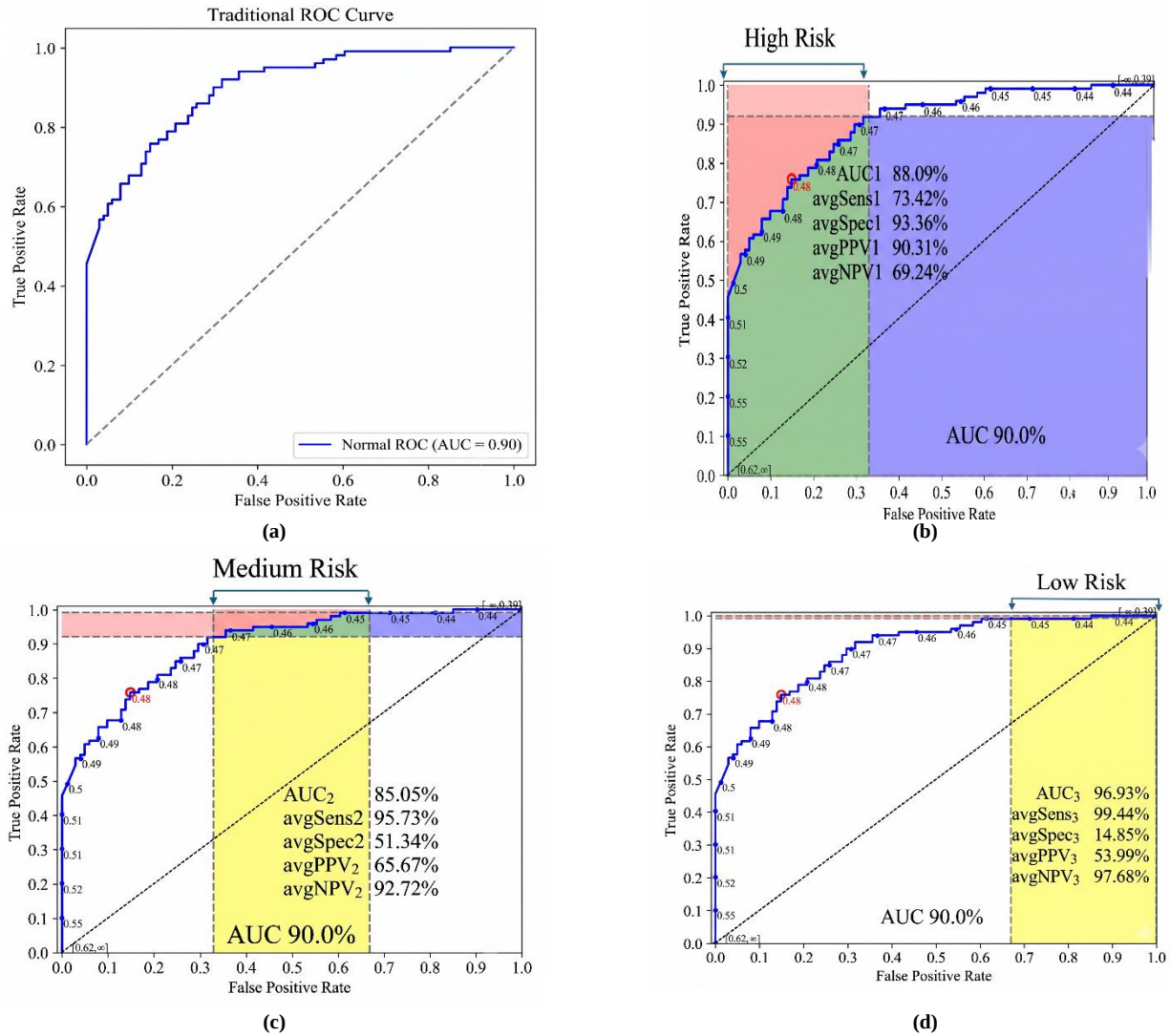


Fig. 2 Comparison of traditional ROC and deep ROC analysis across subgroups (a) Traditional ROC curve, (b) High-risk subgroup performance, (c) Medium-risk subgroup performance, and (d) Low-risk subgroup performance.

The ROC curve fails to display the performance variations for different risk groups or subgroups. Instead, it evaluates model performance over all thresholds as shown in Figure 2(a).

Basically, Area Under the ROC Curve (AUC) gives an overall measure, but it includes clinically irrelevant thresholds. Besides, performance metrics are calculated at a single threshold. In contrast, Deep ROC analysis illustrates risk-based subgroups such as high-risk, medium-risk, and low-risk. Following Figure 2(b), Figure 2(c), and Figure 2(d) show these subgroups' performance, respectively. This evaluation helps to validate model performance in the most clinically relevant groups. Hence, it can be said that Deep ROC analysis allows a more targeted classifier selection based on subgroup performance, as it does not rely on overall or single-threshold measures.

2.3. Deep ROC Analysis (DRA)

As traditional ROC analysis fails to report on performance variations for different patient risk groups, DRA is implemented. It offers a more granular and clinically relevant evaluation framework. Besides, DRA shows multiple subgroups based on predicted risk levels. This enables a clear understanding of model performance for the three groups named high-risk, medium-risk, and low-risk.

2.3.1. Stratification of Predictions

In this framework, predictions are separated into high-risk, medium-risk, and low-risk subgroups based on thresholds for False Positive Rate (FPR) or percentiles of predicted risk scores. These thresholds are defined by clinical relevance, which ensures that performance is evaluated in the context of varying risk levels. The subgroups are typically defined as follows. Stratifying the predictions in this way

permits a more inclusive assessment of model performance, especially in healthcare domains where different thresholds are critical for decision-making based on the patient's risk.

- High-risk group: Identified risk scores in the range [0, 0.33]
- Medium-risk group: Forecasted risk scores in the range [0.33, 0.67]
- Low-risk group: Identified risk scores in the range [0.67, 1]

2.3.2. Pre-Test Measures

Pre-test measures in Deep ROC Analysis mainly offer some testing criteria to assess the ability of the model to differentiate between classes within each subgroup. It has to be done before applying any clinical thresholds. The pre-test measures mentioned in the following offer an extensive understanding of model performance within the specific subgroups of the ROC curve.

Group AUC ($A \cup C_i$):

The Area Under the Curve (AUC) for each subgroup, denoted as $A \cup C_i$. It is calculated to evaluate the model's discriminatory ability within that specific group. The formula is represented by Equation (1). Where $r(x)$ is the ROC curve function. And $[x_1, x_2]$, $[y_1, y_2]$ represents the boundaries of the ROC curve for the given subgroup.

Normalized Group AUC ($A \cup C_{ni}$):

To offer easier comparison between groups, the AUC for each subgroup is normalized, which is calculated by Equation (2), where D_x and D_y represent the maximum possible areas in the horizontal and vertical directions for each subgroup.

Average Sensitivity (se_i):

Here, sensitivity measures the model's competence to correctly recognize positive instances within each subgroup, and it is calculated by Equation (3), where TP represents true positives, and FN represents false negatives.

Average Specificity (sp_i):

Here, specificity calculates the model's capability to correctly find negative instances within each subgroup and is calculated by the following Equation (4), where TN refers to true negatives, and FP refers to false positives.

$$A \cup C_i = \int_{x_1}^{x_2} r(x)dx + \int_{y_1}^{y_2} (1 - r^{-1}(y)) dy \quad (1)$$

$$A \cup C_{ni} = \frac{AUC_i}{\frac{1}{2}(D_x + D_y)} \quad (2)$$

$$se_i = \frac{TP}{TP + FN} \quad (3)$$

$$sp_i = \frac{TN}{TN + FP} \quad (4)$$

2.3.3. Post-Test Measures

Post-test measures are the assessment of model effectiveness after a decision threshold is applied. These include the average positive predictive value and average negative predictive value, as mentioned below.

Average Positive Predictive Value (PPV)

The PPV metric evaluates the percentage of true positive forecasts among all positive forecasts made by the model. It reflects how likely it is that a positive prediction is correct, and this measure is calculated by Equation (5).

$$PPV_i = \frac{TP}{TP + FP} \quad (5)$$

Average Negative Predictive Value (NPV)

NPV measures the percentage of true negative forecasts among all negative forecasts, which indicates the likelihood that a negative prediction is correct. This metric is calculated by Equation (6).

$$NPV_i = \frac{TN}{TN + FN} \quad (6)$$

Algorithm 1 Model Training and Evaluation Pipeline for Myopia Risk Prediction using Deep ROC Analysis

Input: Preprocessed dataset D , ML models $M = \{SVM, RF, MLP, DT\}$.
Output: Model performance metrics across subgroups.
1: Data Split: Divide D into train, test- 80% in D_{train} and 20% in D_{test} .
2: Class Imbalance Handling: Apply SMOTE on D_{train} only (before CV).
3: Cross-validation: Perform 10-fold cross-validation on D_{train} .
4: for each model $m \in M$ do
5: Apply GridSearchCV to find optimal hyperparameters.
6: Train the model on D_{train} .
7: Evaluate model on D_{test} by accuracy, AUC, and Deep ROC Analysis.
8: Obtain predicted scores and actual labels.
9: Stratify predictions into High-risk, Medium-risk, and Low-risk subgroups.
10: for each subgroup do
11: Compute subgroup AUC using Equation (1)
12: Compute normalized AUC using Equation (2)
13: Compute sensitivity using Equation (3)
14: Compute specificity using Equation (4)
15: end for
16: Return subgroup performance metrics.
17: end for
18: Compare: Analyze model performance across high-risk, medium-risk, and low-risk subgroups.

- 19: Select Best Model: Based on subgroup performance and interpretability.
 20: Interpretation: Apply LIME and SHAP to explain predictions of the best model across different risk groups.

2.4. Model Selection

In this research, various familiar machine learning models are selected to evaluate their performance under the framework of Deep ROC Analysis (DRA). These models are selected based on their maturity, interpretability, and proven worth in addressing classification issues in generic domains. Firstly, models are trained by data and then evaluated according to standard measures such as accuracy, AUC and as well as the new Deep ROC Analysis approach. This comparative approach facilitates a deep knowledge of how each model's performance for various risk groups provides a more detailed perception of their predictive success at various decision cut-offs. The subsequent models are selected as follows.

2.4.1. Support Vector Machine (SVM)

SVM is a robust classification algorithm with established proficiency in high-dimensional space. SVM creates a hyperplane to classify data into various classes with the largest margin between them. It is ideal for problems of binary classification and works best when the decision boundary is non-linear, especially when kernel functions are employed. SVM is used in other research in terms of myopia prediction as well [34, 35]. It is selected in this study for its ability to generalize well with limited and complex datasets, accurately discriminating between different risk clusters of myopia. The mathematical formulation of the decision boundary is given by Equation (7).

2.4.2. Random Forest (RF)

Random Forest generates numerous decision trees and then incorporates their predictions to form a single prediction. The ensemble approach prevents risk of overfitting and is stable, especially with noisy data. The final output is the result of majority voting among the trees, and it is calculated by Equation (8). Where $\hat{y}$ represents the output prediction, and $\hat{y}_t$ represents the t-th decision tree output. Besides, Random Forest is well known for its implementation regarding myopia predictions [34]. In this study, it is chosen for its robustness in handling complex datasets, ability to capture non-linear relationships, and high performance with few parameter tuning, making it appropriate for predicting myopia risk across different subgroups.

2.4.3. Multi-Layer Perceptron (MLP)

MLP is an artificial neural network that consists of many layers of neurons, allowing the model to obtain complex and nonlinear interactions in the data. It performs particularly well in feature spaces of high dimensions. The result of each layer

is calculated by Equation (9). Where σ means activation function, w_l represents a weight matrix, and b_l represents the bias vector for the l-th layer. A multi-layer perceptron is used in intraocular lens power calculation [36], which is familiar in eye diagnosis terms. For this study, it is used because of its ability to model intricate patterns and interactions between features, which is critical in accurately identifying varying levels of myopia risk.

2.4.4. Decision Tree (DT)

A decision tree builds a tree model by recursively dividing the data on feature values. It is perhaps the simplest and most interpretable model and is thus useful for understanding how decisions are reached. Here, the decision-making at every node is merely based on split criteria such as Gini impurity or information gain. The Gini impurity is calculated by Equation (10). Where p_i is the ratio of class i in the data set D , and C is the number of classes. In recent practices, decision trees are also exercised in the Myopia domain [37]. Here, DT is considered a baseline model due to its simplicity and transparency. Besides, it is easy to visualize, allowing comparison of more complex models' predictions with an intuitive and clinically understandable approach.

$$w^T x + b = 0 \quad (7)$$

$$\hat{y} = \text{mode}(\hat{y}_1, \hat{y}_2, \dots, \hat{y}_t) \quad (8)$$

$$z_l = \sigma(w_l z_{l-1} + b_l) \quad (9)$$

$$\text{Gini}(D) = 1 - \sum_{i=1}^C p_i^2 \quad (10)$$

2.4.5. Hyperparameter Tuning

To optimize the performance of the selected models (SVM, RF, MLP, and DT), GridSearchCV is employed for performing an extensive search over a predetermined set of hyperparameters. This process ensures that each model achieves the best possible performance on the training data while preventing overfitting. Table 1 represents the hyperparameters used for each model in this study. Fine-tuning these parameters is crucial for improving prediction accuracy and providing robust results across different risk groups in Deep ROC Analysis.

Table 1. Hyperparameters used in the study

Model	Hyperparameters
SVM	C: 1.0; Kernel: 'linear'; Max Iter: 1000; Tolerance: 1e-4
RF	Estimators: 100; Max Depth: None; Min Samples Split: 2; Criterion: 'gini'
MLP	Hidden Layers: (100,); Activation: 'relu'; Solver: 'adam'
DT	Max Depth: None; Min Samples Split: 2; Criterion: 'gini'

2.5. Explainable AI (XAI)

Explainable AI (XAI) is a kind of technique that makes AI models transparent and explainable, such that the reason behind the decisions of models is revealed [38]. Basically, the target is to get AI models to become interpretable and understandable for users. To achieve this, several techniques have been proposed and described below. Like LIME and SHAP are applied in this study.

2.5.1. Local Interpretable Model-Agnostic Explanations (LIME)

The strategies for explanation and interpretation used for describing the nature of complex ML models in terms of predictions are known as LIME. The core idea behind LIME is that it generates an explanation of each prediction by creating a substitute model that estimates the manner in which the black-box model works around the input point [39]. The predictions of the interpretable model are made to be very similar to those of the original model. Normally, it is close to that specific data point, which provides a better perception of the mechanism. This research utilized LIME to denote the unseen decision-making process. Besides, it provides insights in a specific manner to understand why the model predicts something.

2.5.2. Shapley Additive Explanations (SHAP)

SHAP is a cooperative game theory approach. It enables understanding the involvement of each feature in the outcome of a model [40]. SHAP values represent feature contributions to the prediction of the model for a given instance. This approach allows interpretation of individual predictions. Moreover, it is done by assigning a quantitative value to the impact of each feature in a way that is consistent with the theoretical foundation of game theory [41].

SHAP is particularly convenient for complex models. So, it is useful for cases where typical performance metrics like accuracy do not provide a good insight into the model's internal system. In this research, tree SHAP is applied to understand and analyze the predictions from the model.

3. Results and Discussion

This section presents an overview of the dataset, including the pre-processed steps involved. Besides, a thorough analysis of machine learning classifiers for myopia prediction is wrapped here as well. Firstly, classical performance measurement tools are employed to find the best model.

Then, Deep ROC Analysis (DRA) is used to determine model performance across all FPR-based risk categories. It has been done by supplying a clear view of model performance across different levels of patient risk. Finally, LIME and SHAP are utilized to explain model predictions and find the most significant determinants of the risk for myopia.

3.1. Dataset Overview

This study uses Orinda Longitudinal Investigation of Myopia (OLIM) data. It is a longitudinal study that looks at how eye factors change over time and what could cause myopia (nearsightedness) in children. The data includes information from school years 1989–1990 to 2000–2001 and is publicly available on Kaggle [45]. This dataset includes 618 children who were not myopic at the beginning of the research and had at least five years of follow-up data. After that, 537 children stayed non-myopic, but 81 developed myopia. This condition creates a class imbalance in the data. So, now, imbalance conditions need to use special preprocessing methods to ensure fair and accurate model predictions. This dataset includes demographic, genetic, ocular biometric, and visual activity features, as summarized in Table 2.

Table 2. Dataset overview

Category	Details	Number
Dataset Size	Total Students	618
Class Distribution	Non-Myopic Students	537
Class Distribution	Myopic Students	81
Demographic Factors	STUDY YEAR, AGE, GENDER	3
Genetic Factors	Mother's Myopia (MOMMY), Father's Myopia (DADMY)	2
Medical Factors	Spherical Equivalent (SPHEQ), Axial Length (AL), Anterior Chamber Depth (ACD), Lens Thickness (LT), Vitreous Chamber Depth (VCD)	5
Visual Activity Factors	Sports Hours (SPORTHR), Reading Hours (READHR), Computer Hours (COMPHR), Study Hours (STUDYHR), TV Hours (TVHR), Diopter Hours (DIOPTERHR)	6
Total Features		16

3.2. Data Preprocessing and Environmental Settings

The preprocessing steps were applied to the OLIM dataset described in Section 3.1 to ensure data quality and improve model performance. The dataset was first examined for missing values. No significant missing data were observed; therefore, imputation techniques were not required. In this paper, unnecessary features like ID and STUDY YEAR are discarded to refine the dataset and focus on features directly related to myopia prediction. Exceptions, which could distort model predictions, are handled by Winsorization, where values below the 1st percentile and above the 99th percentile are replaced with more appropriate thresholds. To ensure all features contributed equally to the model, standardization is

applied to numerical features by scaling them to a comparable range. This technique prevents features with larger numeric ranges (like eye measurement) from dominating the model's learning process. The dataset exhibited class imbalance, with only 13.1% of the samples being myopic. To manage this, SMOTE is employed on the training dataset for generating

artificial samples for the smaller class (myopic children). This process increases the total sample size to 1,074. Due to this, the training dataset now contains 50% myopic and 50% non-myopic data. In this way, a balanced class distribution in the training set is achieved. Figure 3 refers to the visualization of the before and after situations while using SMOTE.

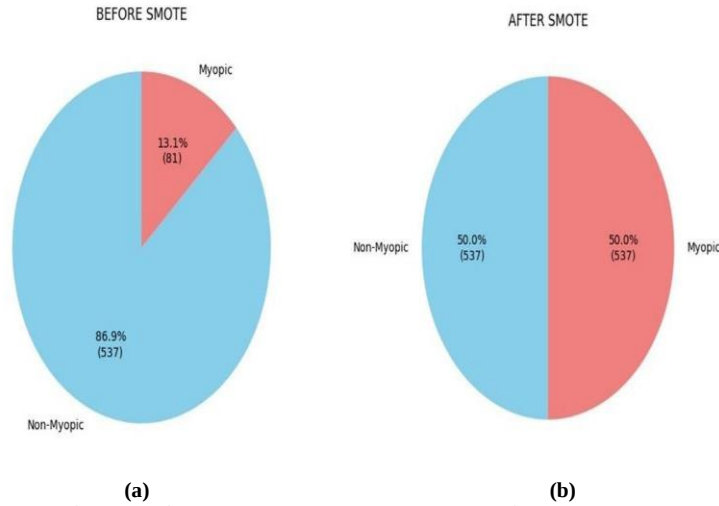


Fig. 3 Comparison of class distribution before and after SMOTE, (a) Class distribution before SMOTE, and (b) Class distribution after SMOTE.

Hence, these preprocessing steps ensure a balanced, high-quality, and well-prepared dataset for model training and evaluation. Experiments are executed on a 64-bit operating system using the Python 3.12.11 development framework on Google Colab. The system has an AMD Ryzen™ 5 3500U processor with a frequency of 2.10 GHz, Radeon™ Vega 8 Graphics with 2 GB memory, and 8 GB of RAM. Python.

3.3. Experimental Results

The experiment was carried out in four aspects:

- Experiment A: Performance analysis of ML Models.
- Experiment B: Deep ROC Analysis (DRA) of Model Performance Across Risk Groups.
- Experiment C: Clinical Validation of Deep ROC Risk Groups
- Experiment D: Calibration of Predicted Probabilities.
- Experiment E: Feature Analysis and Interpretability

3.3.1. Performance Analysis of ML Models

According to experiment A, the initial evaluation of the myopia dataset is performed by using conventional machine learning metrics. These include Accuracy, Precision, Recall, Specificity, G-Mean, AUC, F1 Score, and FPR.

Models named SVM, DT, RF, and MLP are compared, and the findings are summarized in Table 3. From this table, it can be observed that Random Forest outperformed, achieving an accuracy of 95.00%, a precision of 98.16%, a specificity of 98.15%, and an AUC of 98.94%. This result

indicates the strong ability of RF in identifying both myopic and non-myopic patients. Meanwhile, Decision Tree and MLP exhibited moderate performance. Whereas SVM shows relatively high recall (89.31%) and produces the highest FPR (14.07%). Moreover, this suggests a tendency to misclassify non-myopic patients as myopic.

3.3.2. Deep ROC Analysis (DRA) of Model Performance across Risk Groups

In experiment B, the same model's performance is compared to predict the likelihood of developing myopia. The models are evaluated by their outcomes across three predicted risk groups using False Positive Rate (FPR): high risk ($FPR \in [0, 0.33]$), medium risk ($FPR \in [0.33, 0.67]$), and low risk ($FPR \in [0.67, 1]$). This is based on metrics such as AUC, sensitivity, specificity, PPV, and NPV. Now, Table 4 outlines the 10-fold cross-validation results. It illustrates the group-wise performance of all models, while Figure 4 shows the deep ROC curve analysis and overall AUC-based comparison across different risk groups. AUC is examined as the primary measure of model discrimination. Here, Random Forest achieves the highest AUC at 99%. Besides, with a narrower confidence interval, RF outperforms the other models by a 3-6% margin. Next, MLP followed with an AUC of 97%, which is 3% lower than Random Forest. Besides, SVM achieved an AUC of 95%, showing a 1% improvement over MLP, while Decision Tree had the lowest AUC at 93%, which is 6% lower than Random Forest. These outcomes indicate that Random Forest consistently performed the best across all models in terms of overall AUC.

Table 3. Performance comparison using conventional evaluation metrics

Model	Accuracy	Precision	Recall	Specificity	G-Mean	AUC	F1 Score	FPR
SVM	87.68%	87.23%	89.31%	85.93%	87.60%	92.87%	88.25%	14.07%
Random Forest	95.00%	98.16%	92.07%	98.15%	95.05%	98.94%	95.00%	1.85%
MLP	89.73%	90.78%	89.83%	89.63%	89.63%	94.73%	90.17%	10.37%
Decision Tree	92.32%	92.64%	92.59%	92.04%	92.30%	92.31%	92.60%	7.96%

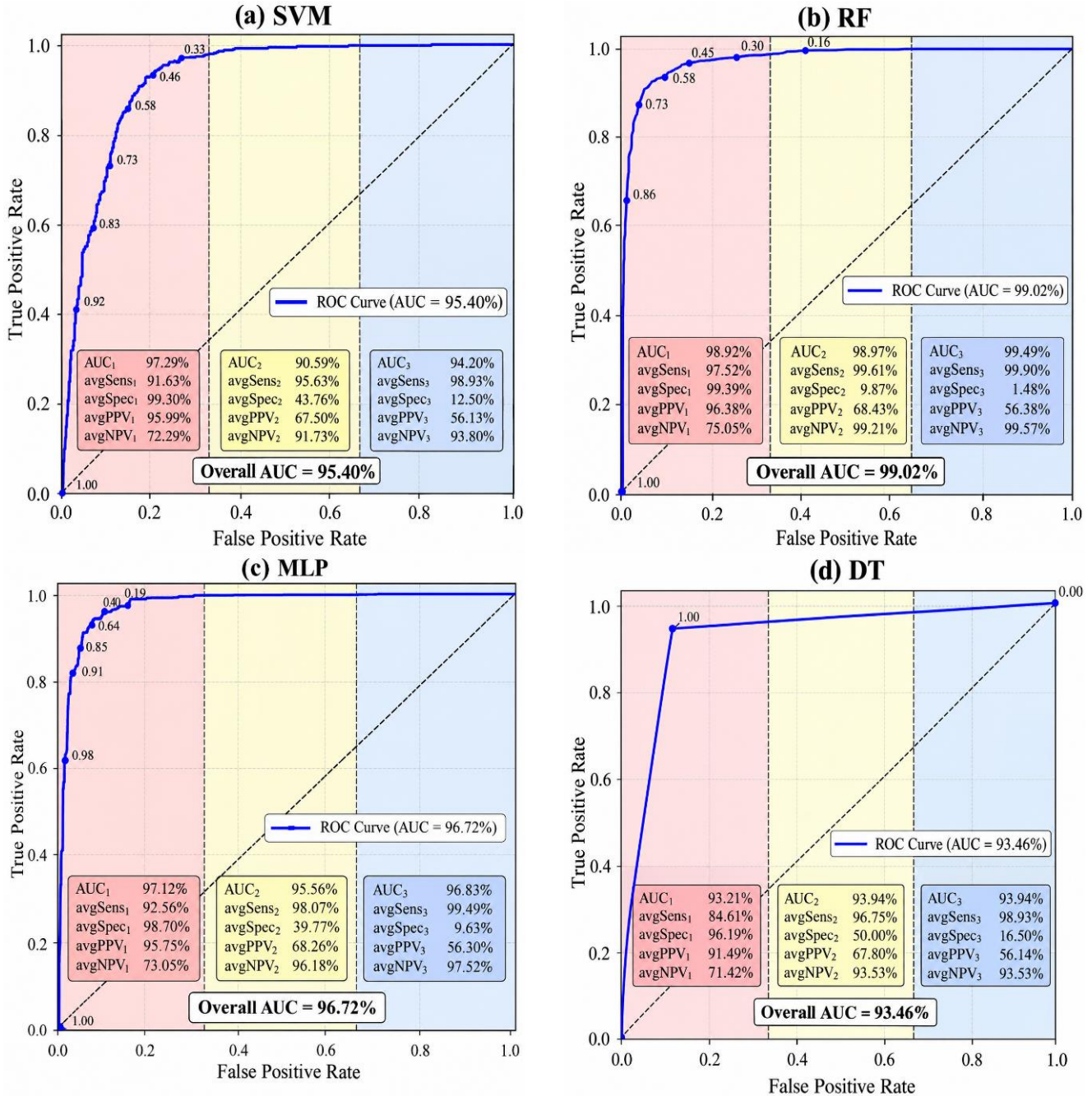


Fig. 4 Deep ROC curve comparison of models across different risk groups

When evaluating performance in the high-risk group ($FPR \leq 0.33$), RF shows the outstanding performance with an AUC_{ni} value of 98.92%, which is better than MLP (97.12%),

SVM (97.29%), and Decision Tree (93.21%). These results confirm Random Forest's superiority in detecting high-risk cases, followed by SVM and MLP. When talking about

average sensitivity, Random Forest again led with 97.52%, which is higher than MLP (92.56%), SVM (91.63%), and Decision Tree (84.61%). It emphasizes the ability to identify high-risk patients correctly.

In terms of low-risk group $FPR \in [0.67, 1]$, Random Forest achieved an AUC of 99.49%. This is stronger than both MLP (96.83%) and SVM (94.20%). In contrast, the Decision Tree has the least performance in this group. Specificity is highest for Random Forest (99.90%), showing a strong ability to correctly classify low-risk individuals, while MLP (9.63%)

shows a much lower specificity, indicating a tendency to overpredict positive cases. The PPV in the high-risk group is highest for Random Forest at 96.38%, which is greater than MLP (95.75%) and SVM (95.99%), and 5% higher than Decision Tree (91.49%).

This indicates that Random Forest reliably identifies true high-risk cases. Negative Predictive Value (NPV) is also highest for Random Forest, with values of 99.21% in the medium-risk level and 99.57% in the low-risk level. These highlight its capacity to perfectly classify low-risk individuals.

Table 4. Group-wise performance of each model

FPR	[0,1]	[0,0.33]	[0.33,0.67]	[0.67,1]
Pred. Risk	All	High	Medium	Low
$A \cup C$	95.40%±1.09		SVM	
$A \cup C_{ni}$	95.40%±1.09	97.29%±0.53	90.59%±2.86	94.20%±3.13
se_i	95.40%±1.09	91.63%±1.75	95.63%±1.23	98.93%±0.73
sp_i	95.40%±1.09	99.30%±0.26	43.76%±11.63	12.50%±8.00
PPV_i	83.54%±0.42	95.99%±0.18	67.50%±0.31	56.13%±0.26
NPV_i	79.76%±1.38	72.29%±0.80	91.73%±2.12	93.80%±4.83
$A \cup C$	99.02%±0.63		Random Forest	
$A \cup C_{ni}$	99.02%±0.63	98.92%±0.56	98.97%±1.29	99.49%±1.16
se_i	99.02%±0.63	97.52%±1.41	99.61%±0.47	99.90%±0.21
sp_i	99.02%±0.63	99.39%±0.29	9.87%±15.10	1.48%±3.35
PPV_i	84.92%±0.26	96.38%±0.22	68.43%±0.13	56.38%±0.06
NPV_i	83.28%±0.53	75.05%±0.55	99.21%±0.97	99.57%±0.97
$A \cup C$	96.72%±1.09		MLP	
$A \cup C_{ni}$	96.72%±1.09	97.12%±0.46	95.56%±2.74	96.83%±2.99
se_i	96.72%±1.09	92.56%±1.75	98.07%±1.39	99.49%±0.53
sp_i	96.72%±1.09	98.70%±0.28	39.77%±20.01	9.63%±7.69
PPV_i	84.00%±0.41	95.75%±0.15	68.26%±0.36	56.30%±0.13
NPV_i	81.41%±1.17	73.05%±0.87	96.18%±2.69	97.52%±2.87
$A \cup C$	93.46%±1.90		Decision Tree	
$A \cup C_{ni}$	93.46%±1.90	93.21%±2.65	93.94%±1.59	93.94%±1.59
se_i	93.46%±1.90	84.61%±5.28	96.75%±0.89	98.93%±0.29
sp_i	93.46%±1.90	96.19%±1.75	50.00%±0.00	16.50%±0.00
PPV_i	80.99%±1.59	91.49%±2.41	67.80%±0.20	56.14%±0.07
NPV_i	79.30%±1.12	71.42%±1.25	93.53%±1.69	93.53%±1.69

According to the above discussion, RT consistently outperformed the other models across all performance measures, including AUC, sensitivity, specificity, PPV, and NPV, particularly in the high-risk group. While SVM and MLP performed well in detecting high-risk individuals, they underperformed in the lower-risk groups. In the meantime, DT showed inferior predictive ability across all risk categories, especially in detecting low-risk individuals. Additionally, Deep ROC evaluation strengthens the necessity of assessing model performance across predicted risk groups. After all, relying only on AUC may lead to overestimating a model's effectiveness, whereas analyzing performance across subgroups provides more meaningful insights for clinical decision-making.

3.3.3. Clinical Validation of Deep ROC Risk Groups

The FPR-based risk groups reflect the clinical cost asymmetry between missed and false-positive cases, not just an arbitrary statistical split. The high-risk group ($FPR \leq 0.33$) prioritizes sensitivity, since missing a child progressing toward myopia delays needed intervention, while the low-risk band ($FPR \in [0.67, 1]$) prioritizes specificity, since false alarms lead to unnecessary follow-up and parental anxiety;

The medium-risk band sits in between, where both error types are important. To establish the clinical relevance of these groups, the Kruskal-Wallis test was used for continuous data and the chi-square test was used for categorical features, shown in Tables 5 and 6, respectively.

Table 5. Kruskal-Wallis test for continuous variables

Variable	High Risk Mean (SD)	Medium Risk Mean (SD)	Low Risk Mean (SD)	Statistic	p_value
AGE	constant, skipped	constant, skipped	constant, skipped		
SPHEQ	0.14 (0.30)	0.35 (0.31)	1.07 (0.35)	720.8	<0.001 3.02E-157
AL	22.63 (0.50)	22.55 (0.62)	22.42(0.63)	23.29	8.77E-06
ACD	3.67 (0.18)	3.62 (0.20)	3.53 (0.21)	88.43	6.27E-20
LT	3.48 (0.10)	3.54 (0.13)	3.55 (0.15)	65.47	6.08E-15
VCD	15.48 (0.51)	15.39 (0.62)	15.33 (0.63)	14.84	0.000600156
SPORTHR	10.05 (6.07)	10.80 (6.77)	12.45 (7.45)	17.63	0.00014822
READHR	2.79 (2.40)	2.68 (2.71)	2.35 (2.63)	13.84	0.000986393
COMPHR	1.41 (1.79)	1.63 (1.75)	1.50 (1.63)	6.95	0.030911138
STUDYHR	1.02 (1.23)	1.22 (1.51)	1.08 (1.49)	2.9	0.234057821
TVHR	7.50 (4.12)	8.81 (5.22)	8.75 (5.43)	11.23	0.003643012
DIOPTERHR	25.98(14.22)	26.18 (14.11)	23.61(13.33)	8.38	0.015171154

Table 6. Chi-square test for categorical variables

Variable	High	Medium	Low	χ^2	p_value
GENDER (Male/Female)	57.3/ 42.7	57.9/ 42.1	56.8/ 43.2	0.09	0.95428148
MOMMY myopia (No/Yes)	56.7/ 43.3	52.9/ 47.1	59.5/ 40.5	3.36	0.186395666
DADMY myopia (No/Yes)	53.8/ 46.2	54.2/ 45.8	58.7/ 41.3	2.23	0.327672747
MYOPIC (No/Yes)	0.8 / 99.2	45.8/ 54.2	98.6/ 1.4	710.54	5.10E-155

Table 5 displays that the suggested high-risk, medium-risk, and low-risk groups represent clinically meaningful patient populations rather than arbitrary statistical partitions. AGE is marked "constant, skipped" because all children in this dataset are the same age, so there is no variation to test. Children in the high-risk group have a lower SPHEQ (0.14 D) than the medium-risk cluster (0.35 D) and the low-risk cluster (1.07 D), so the value increases clearly as risk goes down. The eye measurements follow the same pattern.

AL drops from 22.63 mm in the high-risk level to 22.55 mm in the medium-risk level and 22.42 mm in the low-risk level. Similarly, ACD drops from 3.67 mm to 3.62 mm in the medium band and to 3.53 mm in the low band. Conversely, LT rises from 3.48 mm to 3.54 mm and 3.55 mm in high, medium and low bands, respectively. All of these changes move in the exact direction expected as eyes shift toward myopia. Meanwhile, behavioral habits show the same trend. Sports hours rise steadily from the high-risk group to the low-risk group, while reading hours fall across the groups. From Table 6, it is observed that the percentage of myopic children falls sharply across the three groups, from 99.2% in the high-risk level to 54.2% in the medium-risk level to just 1.4% in the low-risk level. This confirms that the risk tiers reflect real clinical outcomes at every level, not just at the two extremes. Gender and parental myopia did not differ across groups, suggesting the grouping comes mainly from eye health and lifestyle factors rather than family history. Overall, the findings of these analyses confirm the clinical validity and relevance associated with the three subcategories defined in the proposed Deep ROC framework.

3.3.4. Calibration of Predicted Probabilities

A calibration curve shows how well predicted probabilities match actual outcomes, with the diagonal line representing perfect calibration. Figure 5 demonstrates the curves for Random Forest, Decision Tree, SVM, and MLP, with Brier scores. Where for Random Forest it is 0.037, for Decision Tree it is 0.065, and for SVM, MLP it is 0.089 and 0.066, respectively. Lower Brier scores indicate more reliable predictions. From Figure 4, Random Forest has the lowest score, which shows the most accurate probabilities. Meanwhile, SVM has the highest, which shows the least reliable estimates. On the other side, MLP and Decision Tree are moderately reliable. In short, the calibration curves in this outcome complement Random Forest's superior performance in traditional evaluation metrics as well as in the Deep ROC subgroup analysis.

3.3.5. Feature Analysis and Interpretability

This research examines the significance of features in the myopia prediction model. It focuses on enhancing the interpretability of the ML model. The experiment identifies Random Forest as the best performer. Next, the step is to explore RT's clarity to ensure that the predictions are transparent and reliable in a clinical context. In medical diagnostics, especially in eye-care applications, it is crucial that ML models provide high predictive accuracy as well as explanations behind the results. By this, the clinician can understand and rely on it for informed decision-making. Hence, to achieve this, two widely used Explainable AI techniques named LIME and SHAP are applied to explain the model's reasoning process. LIME delivers local clarity by approximating the opaque model with simpler, understandable

models in the area of each data point. Meanwhile, SHAP computes the participation of each feature in the model's prediction. This provides more granular information on how individual attributes impact the final prediction.

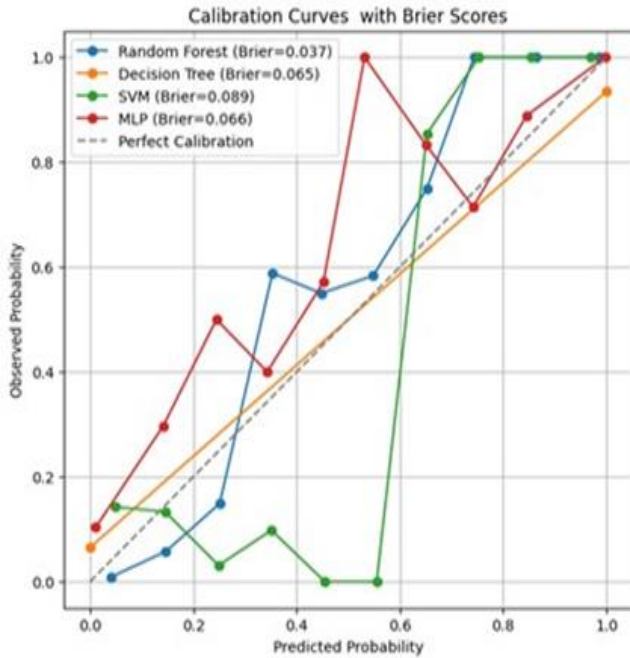


Fig. 5 Calibration curves of the classifiers

LIME Analysis for “Has Myopia” Prediction

LIME explanation for a patient predicted to have myopia (95% confidence) is represented in Figure 6(a). The orange bars of the figure indicate contributions toward “Has Myopia,” while blue bars show efforts toward “No Myopia.” Here, SPHEQ (spherical equivalent) and SPORTHR (sports hours) positively influenced the “No Myopia” prediction. Specifically, the values of SPHEQ greater than or equal to 0.56 strongly predicted the absence of myopia. In addition, higher SPORTHR represents more outdoor activity and reduces myopia risk.

Meanwhile, Diopter Hours (DIOPTERHR), Study Hours (STUDYHR), Vitreous Chamber Depth (VCD), and AGE increased the myopia likelihood. So, the extended duration of study time and near-vision tasks increased the risk notably. This indicates that it is necessary to follow preventive strategies. For example, taking regular breaks from long-term near work.

LIME Analysis for “No Myopia” Prediction

Figure 6(b) illustrates the LIME explanation for a patient predicted to have no myopia (93% confidence). Here, the blue bars denote “No Myopia,” and the orange bars denote features increasing myopia risk. In this analysis, SPORTHR is the strongest protective factor, whereas DIOPTERHR and VCD support this outcome as well. On the other side, SPHEQ,

STUDYYEAR, and STUDYHR have minor negative effects. This suggests that lengthy study time or higher refractive error can increase myopia risk among non-myopic individuals. These results reflect the complex interactions between lifestyle and visual physiology.

SHAP Analysis

SHAP analysis is applied to realize the influence of individual features on predictions. Here, the SHAP values represent the average involvement of each feature for all probable combined features. Positive SHAP values show the increasing likelihood of a specific prediction. Moreover, the negative values suggest a reduction of likelihood. SHAP analysis presents some key features that influence the model's predictive ability. For example, Spherical Equivalent (SPHEQ) and Sports Hours (SPORTHR) are for non-myopia and assume a lower risk of developing myopia. In contrast, Study Hours (STUDYHR) and Diopter Hours (DIOPTERHR) refer to the likelihood of myopia and refer to a higher risk. Figure 8 demonstrates the SHAP values for these key features.

The positive SHAP values for SPHEQ and SPORTHR mention their strong predictive power for non-myopia. Moreover, the negative values for STUDYHR and DIOPTERHR show that higher values increase the likelihood of myopia. So, there is a complex relationship between these features and myopia risk.

In addition, feature attribution analysis using SHAP has been conducted separately for the high-risk, medium-risk, and low-risk subgroups to evaluate the significance of features across different groups. The resulting feature rankings are shown in Figure 7, and Table 7 demonstrates that there is a considerable variation in the relative impact of the predictors for the three groups.

Although SPHEQ is the most significant predictor in the overall, high-risk, and medium-risk analyses, its importance has reduced significantly in the low-risk subgroup, where SPORTHR is the most influential feature. Moreover, ocular biometric parameters remained consistently important across all risk categories, whereas behavioral factors are more influential in the low-risk subgroup.

These findings show that the evaluation of features using the concept of subgroups contributes to more insights than evaluating the features with the concept of global feature importance.

The combination of LIME and SHAP enhances model transparency. They provide individual-level details and insights into the global feature importance. These interpretability tools validate the Random Forest model's predictive decisions. Besides, it provides guidance for clinical decision-making and preventive strategies in myopia management.

3.4. Observation

The key observations regarding the proposed approach for myopia prediction are as follows:

- Deep ROC Analysis (DRA) evaluates model performance over different risk groups (high, medium, low), and Random Forest outperformed other models while detecting high-risk cases. This consistent result shows its clinical usefulness to identify the would-be myopic patients correctly.
- The calibration analysis and Explainable AI (XAI) techniques confirm that Random Forest model is reliable with the lowest Brier score. And it enhances the transparency of predictions as well. Besides, SPHEQ and SPORTHR are identified for providing actionable insights for clinicians. This supports the model's practical applications in myopia management.
- Lastly, it can be noted that the combination of Random Forest with Deep ROC and Explainable AI techniques is a progressive approach in myopia prediction. It enhances the prediction accuracy and reliability for myopia control.

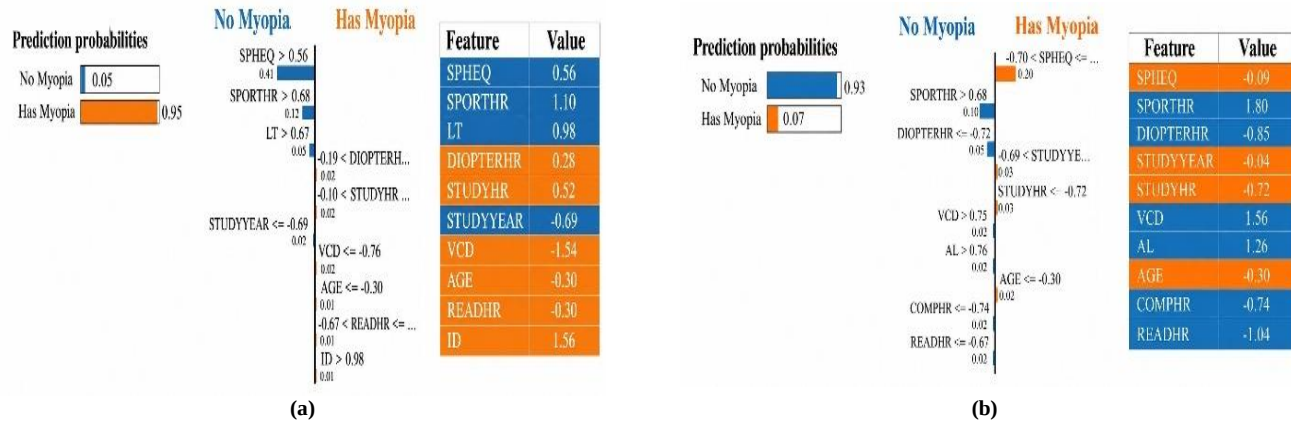
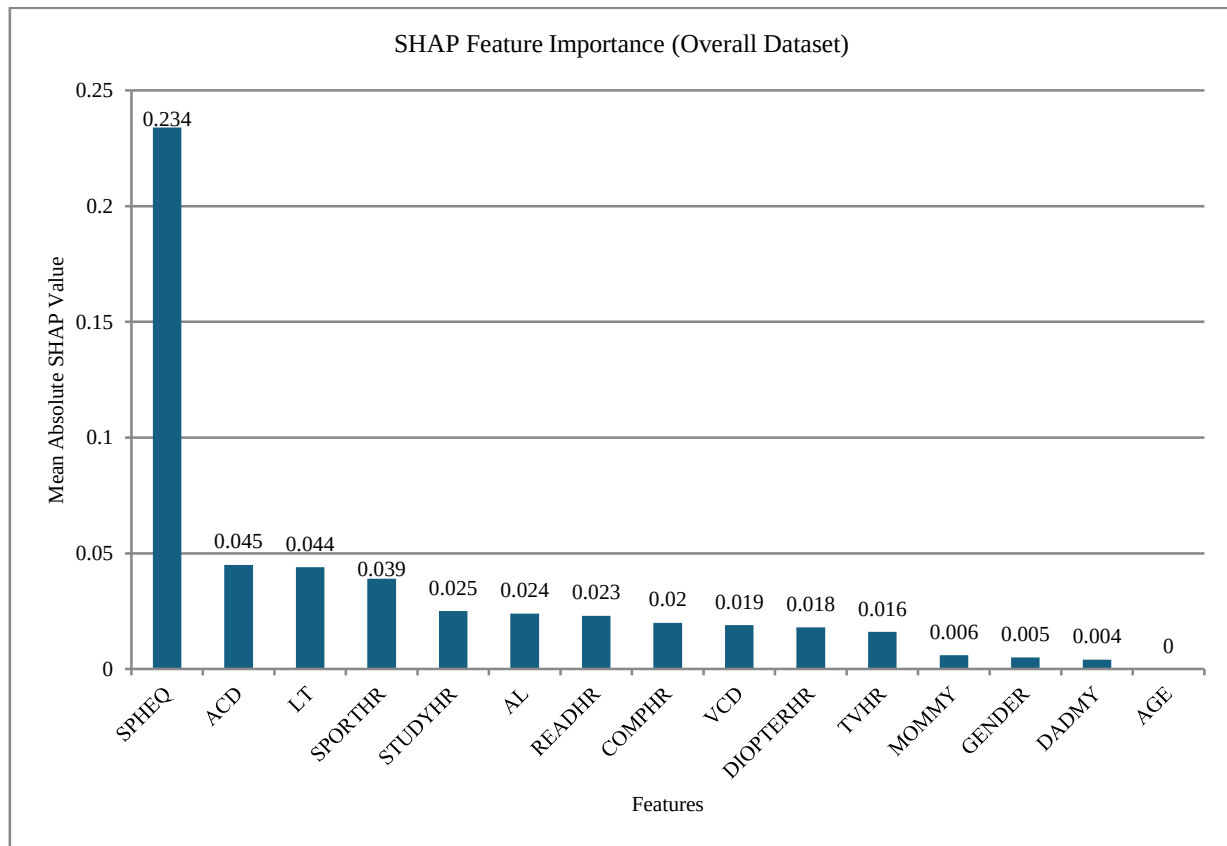
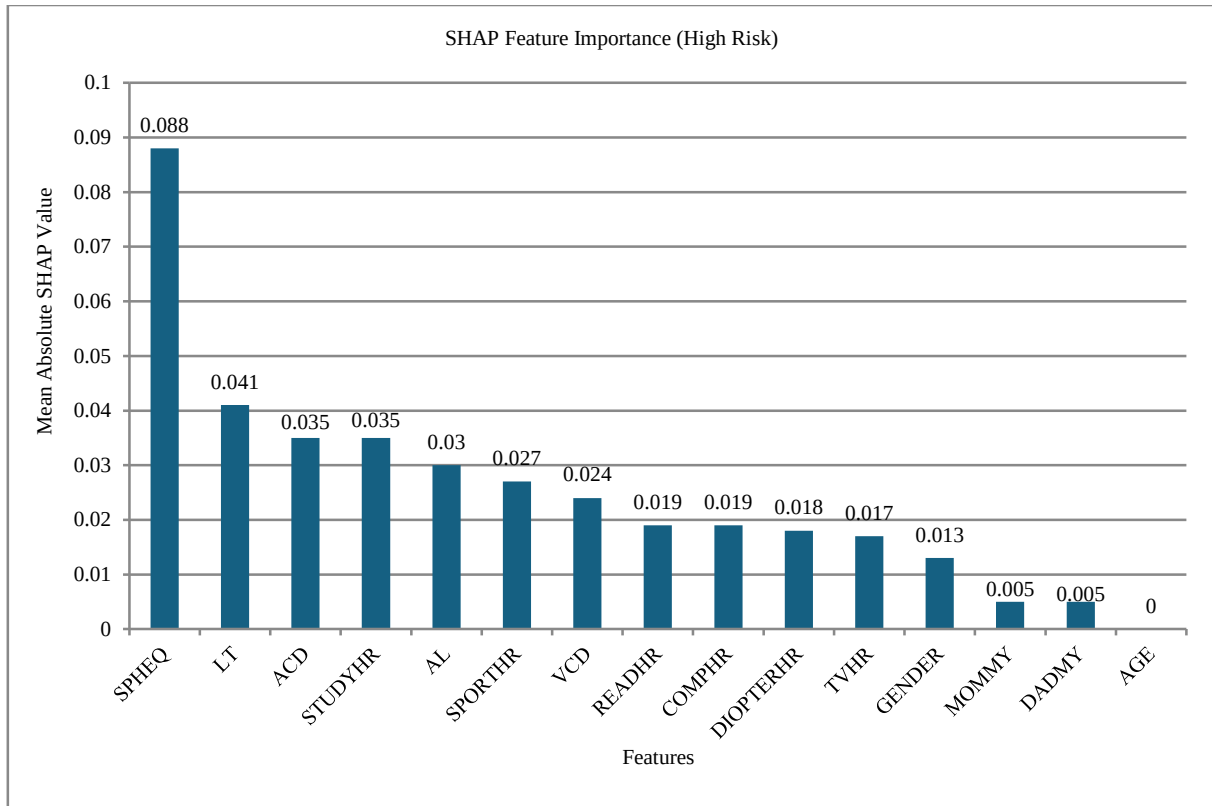


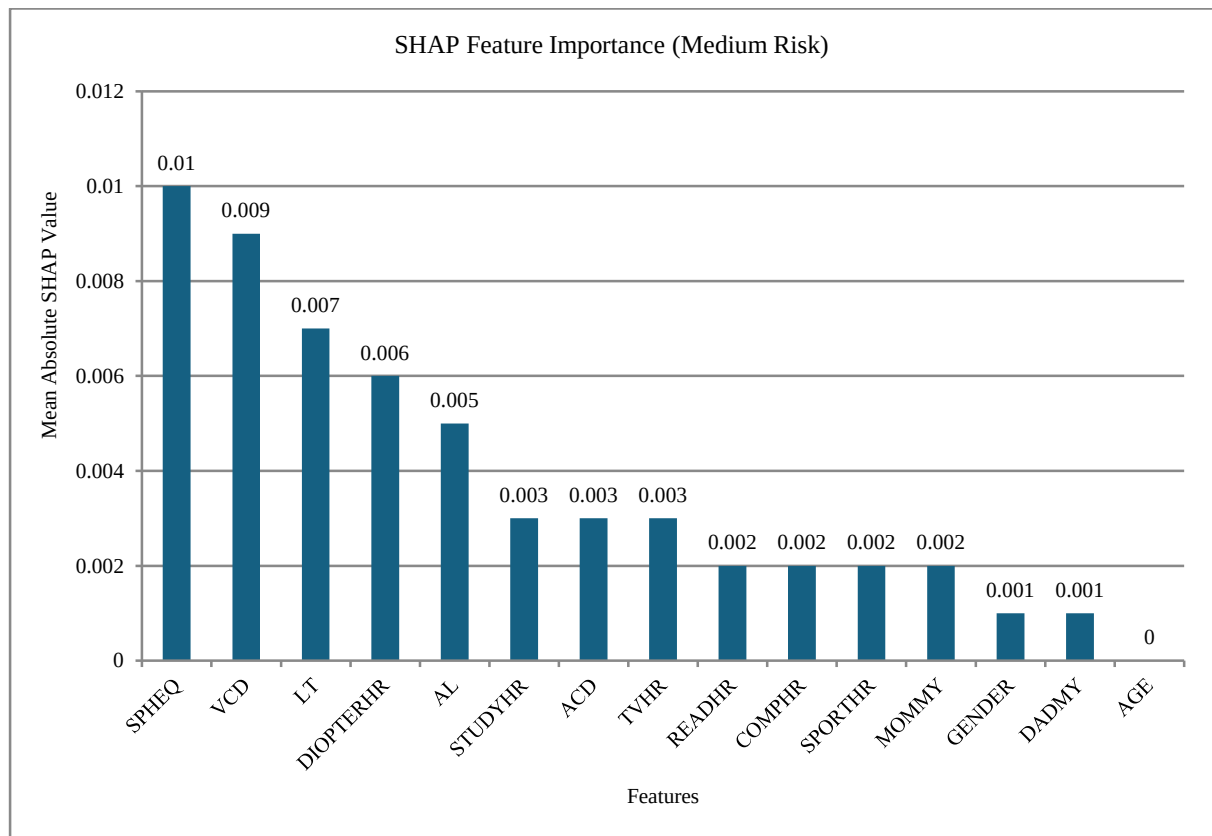
Fig. 6 Interpretation instance using LIME, (a) Explanation of positive case instances, and (b) Explanation of negative case instances.



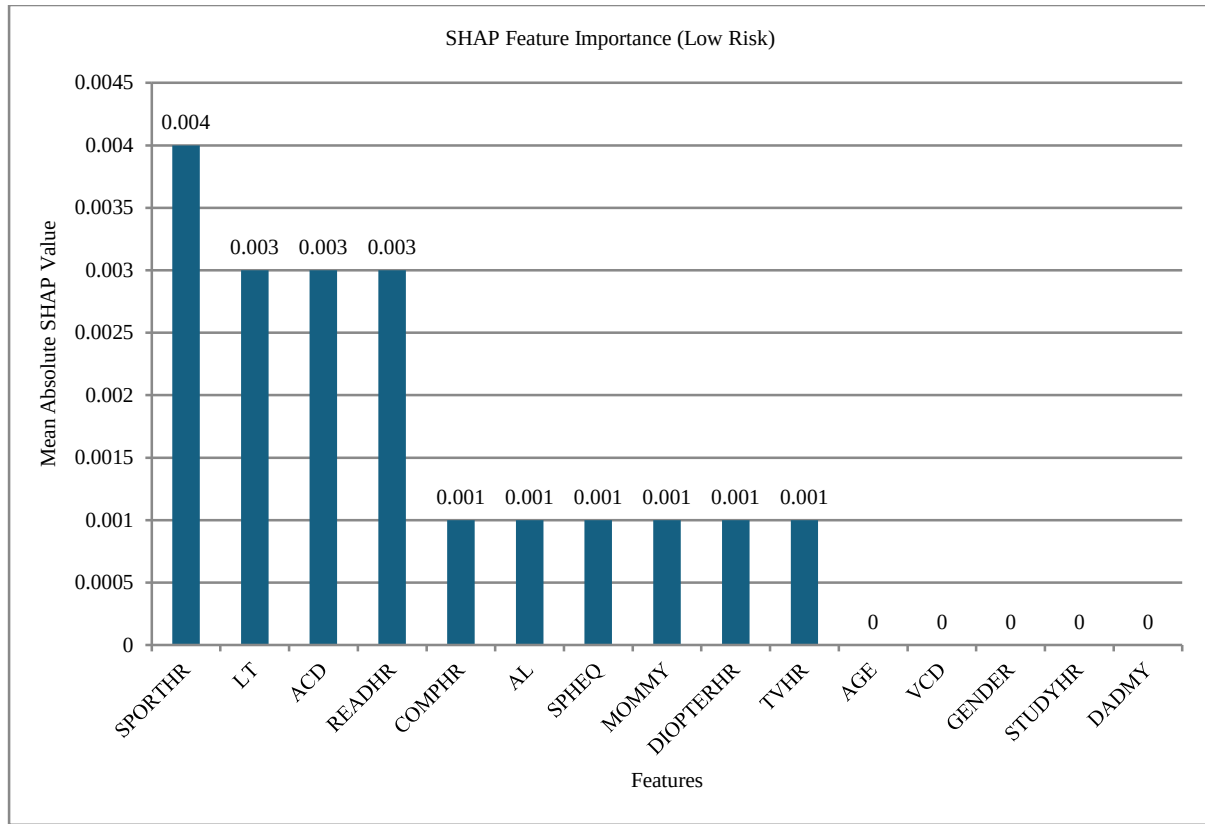
(a)



(b)



(c)



(d)

Fig. 7 SHAP feature importance across the Deep ROC risk categories: (a) Overall dataset, (b) High-risk subgroup, (c) Medium-risk subgroup, and (d) Low-risk subgroup.

Table 7. Subgroup-specific SHAP feature attribution rankings for the Deep ROC risk categories

Rank	Overall Feature	Mean ISHAPI	High Risk	Mean ISHAPI	Medium Risk	Mean ISHAPI	Low Risk	Mean ISHAPI
1	SPHEQ	0.234	SPHEQ	0.088	SPHEQ	0.010	SPORTH	0.004
2	ACD	0.045	LT	0.041	VCD	0.009	LT	0.003
3	LT	0.044	ACD	0.035	LT	0.007	ACD	0.003
4	SPORTH	0.039	STUDYHR	0.035	DIOPTERHR	0.006	READHR	0.003
5	STUDYHR	0.025	AL	0.030	AL	0.005	COMPHR	0.001
6	AL	0.024	SPORTH	0.027	STUDYHR	0.003	AL	0.001
7	READHR	0.023	VCD	0.024	ACD	0.003	SPHEQ	0.001
8	COMPHR	0.020	READHR	0.019	TVHR	0.003	MOMMY	0.001
9	VCD	0.019	COMPHR	0.019	READHR	0.002	DIOPTERHR	0.001
10	DIOPTERHR	0.018	DIOPTERHR	0.018	COMPHR	0.002	TVHR	0.001
11	TVHR	0.016	TVHR	0.017	SPORTH	0.002	AGE	0.000
12	MOMMY	0.006	GENDER	0.013	MOMMY	0.002	VCD	0.000
13	GENDER	0.005	MOMMY	0.005	GENDER	0.001	GENDER	0.000
14	DADMY	0.004	DADMY	0.005	DADMY	0.001	STUDYHR	0.000
15	AGE	0.000	AGE	0.000	AGE	0.000	DADMY	0.000

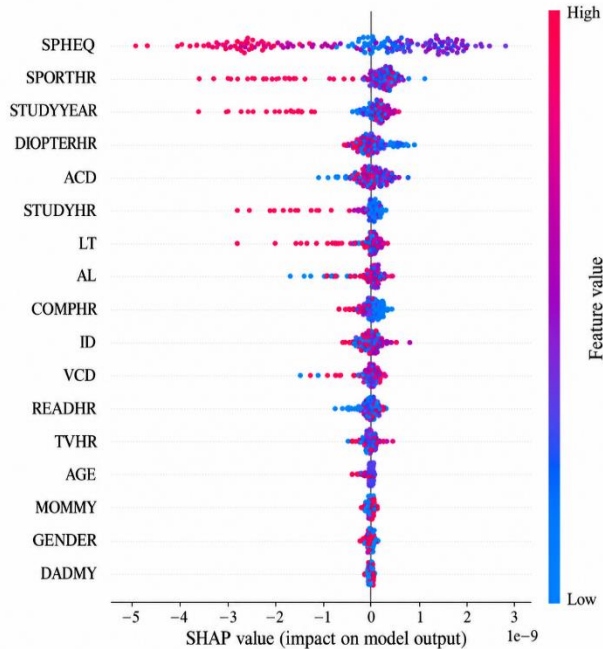


Fig. 8 Myopia detection model explainability using SHAP

The proposed framework provides an overall assessment of the performance of myopia prediction models compared to existing studies, which are based primarily on overall prediction accuracy. Deep ROC analysis allows the assessment of the different subgroups, and statistical significance testing can be used to ensure that the different risk groups identified are clinically meaningful. Furthermore, SHAP analyses conducted at the subgroup level show that the importance of the factors in the predictions differs among subgroups. This is not possible using only global performance or overall feature importance analysis, which further demonstrates the clinically relevant added value of the proposed framework.

4. Conclusion

This study explains an evaluation framework, which combines Deep ROC analysis with Explainable AI (XAI) approaches in terms of predicting myopia. An experiment has been done to evaluate ML classifiers named SVM, Random Forest, MLP, and Decision Tree. Among these, Random Forest shows optimal results for all the risk groups. When applying Deep ROC analysis, it complements the limitations of traditional measures like ROC and AUC. Then, the inclusion of XAI techniques (LIME and SHAP) increases the transparency of model predictions significantly.

These elements comprise a more comprehensive and clinically relevant evaluation framework than existing ones. The interpretability outputs generated by LIME and SHAP do not work only as a visualization tool; rather, they provide clinically actionable insights. By identifying which factors have the strongest influence on the prediction, these

explanations help clinicians understand why a child has been classified into a high-risk or low-risk category. For example, in the higher-risk groups, the ocular biometric variables were more prevalent, while the behavioral factors were relatively more prevalent in the lower-risk group.

Explanations that are specific to the subgroups can be useful in understanding what is driving the different risks and the interpretation of model predictions. Although no external validation data were found, the strength of the proposed framework was assessed through stratified 10-fold CV, Deep ROC analysis for subgroups and Brier Score-based calibration. The uniform performance across folds and across risk groups, together with the good calibration performance, indicates that the model learned stable and reliable predictive patterns in the studied population. However, further external validation with independent cohorts is important for future studies. Instead of the promising results, several gaps remain. First, limited data is a limitation of this study.

Second, the task was addressed with orderly data and without utilizing retinal images or time-series data. Third, although SHAP and LIME make the model more transparent, the explanations are not causal relationships, but rather associations that were found by the model. Future work will involve testing the framework with independent external cohorts, enhancing the data set to include more diverse populations and combining multimodal data, including retinal pictures and behavioral data. Furthermore, future prospective longitudinal studies could contribute to a thorough understanding of the mechanisms of myopia development and help to improve the application of the proposed framework.

Data and Code Availability

The dataset deployed in this study is publicly available (Kaggle OLIM dataset). The implementation code will be shared by the corresponding author upon a reasonable request via email.

Conflicts of Interest

The author of this research has no financial or non-financial conflicts of interest.

Acknowledgments

The author gratefully acknowledges the support provided by the MMU-TMR&D co-creation research project, under grant no MMUE/240094.

Contribution

SA designs the whole research and supervision. RS implements the research. KFT writes the original draft. SNW completes necessary edits. SBH does writing, reviewing, and editing. AM acquires funds. NHAA does review and fund acquisition.

References

- [1] Shahab S. Band et al., "Application of Explainable Artificial Intelligence in Medical Health: A Systematic Review of Interpretability Methods," *Informatics in Medicine Unlocked*, vol. 40, pp. 1-20, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [2] Amir Masoud Rahmani et al., "Machine Learning (ML) in Medicine: Review, Applications, and Challenges," *Mathematics*, vol. 9, no. 22, pp. 1-52, 2021. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [3] André M. Carrington et al., "A new Concordant Partial AUC and Partial c Statistic for Imbalanced Data in the Evaluation of Machine Learning Algorithms," *BMC Medical Informatics and Decision Making*, vol. 20, no. 4, pp. 1-12, 2020. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [4] Lior Rabkin, Ilan Cohen, and Gonen Singer, "Resource Allocation in Ordinal Classification Problems: A Prescriptive Framework Utilizing Machine Learning and Mathematical Programming," *Engineering Applications of Artificial Intelligence*, vol. 132, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [5] Elie Dolgin, "The Myopia Boom," *Nature*, vol. 519, no. 7543, pp. 276-278, 2015. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [6] Brien A. Holden et al., "Global Prevalence of Myopia and High Myopia and Temporal Trends from 2000 through 2050," *Ophthalmology*, vol. 123, no. 5, pp. 1036-1042, 2016. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [7] Lik Thai Lim et al., "Impact of Parental History of Myopia on the Development of Myopia in Mainland China School-Aged Children," *Therapeutic Advances in Ophthalmology*, vol. 6, pp. 31-35, 2014. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [8] L.L.K. Lin et al., "Prevalence of Myopia in Taiwanese Schoolchildren: 1983 to 2000," *Annals-Academy of Medicine Singapore*, vol. 33, no. 1, pp. 27-33, 2004. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [9] Mingguang He et al., "Effect of Time Spent Outdoors at School on the Development of Myopia Among Children in China: A Randomized Clinical Trial," *Jama Network*, vol. 314, no. 11, pp. 1142-1148, 2015. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [10] Ian G Morgan, Kyoko Ohno-Matsui and Seang-Mei Saw, "Myopia," *The Lancet*, vol. 379, no. 9827, pp. 1739-1748, 2012. [[Google Scholar](#)] [[Publisher Link](#)]
- [11] Chen-Wei Pan, Dharani Ramamurthy, and Seang-Mei Saw, "Worldwide Prevalence and Risk Factors for Myopia," *Ophthalmic and Physiological Optics*, vol. 32, no. 1, pp. 3-16, 2012. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [12] Richard F. Spaide, Kyoko Ohno-Matsui, and Lawrence A. Yannuzzi, *Pathologic Myopia*, 1st ed., Springer New York, NY, 2014. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [13] Ian G. Morgan et al., "The Epidemics of Myopia: Aetiology and Prevention," *Progress in Retinal and Eye Research*, vol. 62, pp. 134-149, 2018. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [14] Annechien E.G. Haarman et al., "The Complications of Myopia: A Review and Meta-Analysis," *Investigative Ophthalmology and Visual Science*, vol. 61, no. 4, pp. 1-19, 2020. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [15] Pavan Kumar Verkicharla, Priyanka Kammari, and Anthony Vipin Das, "Myopia Progression Varies with Age and Severity of Myopia," *PLoS One*, vol. 15, no. 11, pp. 1-9, 2020. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [16] Namra Rauf, Syed Omer Gilani, and Asim Waris, "Automatic Detection of Pathological Myopia using Machine Learning," *Scientific Reports*, vol. 11, no. 1, pp. 1-9, 2021. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [17] Shi-Ming Li et al., "Machine Learning to Determine Risk Factors for Myopia Progression in Primary School Children: The Anyang Childhood Eye Study," *Ophthalmology and Therapy*, vol. 11, no. 2, pp. 573-585, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [18] Jianxia Fang et al., "Machine Learning for Predicting the Treatment Effect of Orthokeratology in Children," *Frontiers in Pediatrics*, vol. 10, pp. 1-11, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [19] Yuzhuo Fan et al., "Machine Learning Algorithm Improves Accuracy of Ortho-K Lens Fitting in Vision Shaping Treatment," *Contact Lens and Anterior Eye*, vol. 45, no. 3, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [20] Tien-En Tan et al., "Retinal Photograph-based Deep Learning Algorithms for Myopia and a Blockchain Platform to Facilitate Artificial Intelligence Medical Research: A Retrospective Multicohort Study," *The Lancet Digit Health*, vol. 3, no. 5, pp. 317-329, 2021. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [21] Ruben Hemelings et al., "Pathological Myopia Classification with Simultaneous Lesion Segmentation using Deep Learning," *Computer Methods and Programs in Biomedicine*, vol. 199, pp. 1-18, 2021. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [22] Worood Esam Noori Noori, and A.S. Albahri, "Towards Trustworthy Myopia Detection: Integration Methodology of Deep Learning Approach, XAI Visualization, and User Interface System," *Applied Data Science and Analysis*, vol. 2023, pp. 1-15, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [23] Jai Prashar, and Nicole Tay, "Performance of Artificial Intelligence for the Detection of Pathological Myopia from Colour Fundus Images: A Systematic Review and Meta-Analysis," *Eye*, vol. 38, no. 2, pp. 303-314, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [24] Andrzej Grzybowski et al., "A Review on the Epidemiology of Myopia in School Children Worldwide," *BMC Ophthalmology*, vol. 20, no. 1, pp. 1-11, 2020. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [25] Mengru Ba, and Zhijie Li, "The Impact of Lifestyle Factors on Myopia Development: Insights and Recommendations," *AJO International*, vol. 1, no. 1, pp. 1-21, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]

- [26] Li Lu et al., "AI-Model for Identifying Pathologic Myopia based on Deep Learning Algorithms of Myopic Maculopathy Classification and "Plus" Lesion Detection in Fundus Images," *Frontiers in Cell and Developmental Biology*, vol. 9, pp. 1-10, 2021. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [27] Li Lu et al., "Development of Deep Learning-based Detecting Systems for Pathologic Myopia using Retinal Fundus Images," *Communications Biology*, vol. 4, no. 1, pp. 1-8, 2021. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [28] Chen Guo, Minzhong Yu, and Jing Li., "Prediction of Different Eye Diseases based on Fundus Photography via Deep Transfer Learning," *Journal of Clinical Medicine*, vol. 10, no. 23, pp. 1-11, 2021. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [29] Ling-Ping Cen et al., "Automatic Detection of 39 Fundus Diseases and Conditions in Retinal Photographs using Deep Neural Networks," *Nature Communications*, vol. 12, no. 1, pp. 1-13, 2021. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [30] Xiaomeng Li et al., "Rotation-Oriented Collaborative Self-Supervised Learning for Retinal Disease Diagnosis," *IEEE Transactions on Medical Imaging*, vol. 40, no. 9, pp. 2284-2294, 2021. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [31] Junjia Huang et al., "Myopia prediction for Children and Adolescents via Time-Aware Deep Learning," *Scientific Reports*, vol. 13, no. 1, pp. 1-10, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [32] Yifan Wang et al., "Enhancing AI Transparency in IoT Intrusion Detection using Explainable AI Techniques," *Internet of Things*, vol. 33, pp. 1-26, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [33] André M. Carrington et al., "Deep ROC Analysis and AUC as Balanced Average Accuracy, for Improved Classifier Selection, Audit and Explanation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 329-341, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [34] Jingfeng Mu, Haoxi Zhong, and Mingjie Jiang, "Machine-Learning Models to Predict Myopia in Children and Adolescents," *Frontiers in Medicine*, vol. 11, pp. 1-10, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [35] Xu Yang et al., "Prediction of Myopia in Adolescents through Machine Learning Methods," *International Journal of Environmental Research and Public Health*, vol. 17, no. 2, pp. 1-17, 2020. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [36] José Carlos Fernández-Álvarez et al., "Using a Multilayer Perceptron in Intraocular Lens Power Calculation," *Journal of Cataract and Refractive Surgery*, vol. 45, no. 12, pp. 1753-1761, 2019. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [37] Jun-Wei Chen et al., "Decision Tree Modeling to Predict Myopia Progression in Children Treated with Atropine: Toward Precision Ophthalmology," *Diagnostics*, vol. 15, no. 16, pp. 1-13, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [38] Sajid Ali et al., "Explainable Artificial Intelligence (XAI): What we know and what is left to Attain Trustworthy Artificial Intelligence," *Information Fusion*, vol. 99, pp. 1-52, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [39] Turker Berk Donmez et al., "Explainable AI in Action: A Comparative Analysis of Hypertension Risk Factors using SHAP and LIME," *Neural Computing and Applications*, vol. 37, no. 5, pp. 4053-4074, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [40] Iasonas Bakas, and Karolos J. Kontoleon, "ANNs and Game Theory SHAP Algorithms Quantifying the Contribution of Geometric and Temperature-Dependent Material Properties to the Thermal Response of Concrete Walls Exposed to Elevated Temperatures," *Construction and Building Materials*, vol. 493, pp. 1-14, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [41] Ying Liu et al., "Clinical Decision Support Tool for Breast Cancer Recurrence Prediction using SHAP Value in Cooperative Game Theory," *Heliyon*, vol. 10, no. 2, pp. 1-11, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [42] Haohan Zou et al., "Retinal Features as Predictive Indicators for High Myopia: Insights from Explainable Multi-Machine Learning Models," *Frontiers in Bioengineering and Biotechnology*, vol. 13, pp. 1-9, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [43] Qi Feng et al., "Interpretable Machine Learning Models for Predicting Childhood Myopia from School-based Screening Data," *Scientific Reports*, vol. 15, no. 1, pp. 1-11, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [44] Isteaq Kabir Sifat et al., "Prediction of Myopia Among Undergraduate Students using Ensemble Machine Learning Techniques," *Health Science Reports*, vol. 8, no. 5, pp. 1-12, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [45] George Papachristou, Myopia Study, Kaggle, 2025. [Online]. Available: <https://www.kaggle.com/datasets/mscgeorges/myopia-study>