# A MACHINE LEARNING APPROACH WITH XAI FOR PREDICTING EYE HEALTH RISK SCORES USING BASE AND ENSEMBLE CLASSIFICATION MODELS

Md. Sharfuddin, Md Omar Faruque , Fahad Ahmed,Abyaz Ahammed Asif
Department of Computer Science and Engineering, Southeast university
Dhaka, Bangladesh
sharfuddin6161@gmail.com, omar.sghsc@gmail.com, fahad.ahmed.acps@gmail.com,
abyajahmed@gmail.com

## ABSTRACT

Global prevalence of digital eye strain and vision degradation has exponentially increased due to prolonged screen-time, unoptimised display metrics and poor environmental conditions. While clinical diagnostics exist, they often fail to provide predictive tracking for early prevention. Existing machine learning models face high distribution variance, class imbalances, and severe lack of interpretability, which limits the clinical reliability and adoption of these models in healthcare informatics. To bridge this gap, this study presents a comprehensive predictive intelligence framework for eye health risk stratification based on the dataset. The methodology involves an automated data preprocessing pipeline, which involves mode alignment, recursive feature selection using SelectKBest to select the top 9 predictive features, and synthetic minority oversampling (SMOTE) to balance the classes. We benchmark nine basic machine learning models and optimised soft-voting ensemble setups.The proposed soft-voting ensemble framework of Support Vector Machine and Logistic Regression (SVM+LR) achieves the highest performance with an accuracy of 0.8707, a precision of 0.8697, a recall of 0.8720 and a matching F1-score of 0.8708, and the Mean Squared Error (MSE) is minimised to 0.1293, as shown in empirical evaluations. We further achieve algorithmic transparency with post-hoc Explainable AI (XAI) diagnostics, with local and global SHAP and LIME explanation layers. This robust and lightweight model effectively reduces predictive variance, resulting in a highly accurate, explainable and deployable architecture for real-time mobile ocular health screening and digital monitoring.

**Keywords:** Digital Eye Strain, Machine Learning Ensemble Learning, Explainable AI (XAI), SHAP, LIME.

## 1. INTRODUCTION

Owing to the often asymptomatic or underdiagnosed nature of many vision threatening eye conditions, they are an important public health concern. Group of the most prevalent eye diseases contributing to global blindness and visual impairment include diabetic retinopathy, cataract, glaucoma, age related macular degeneration as well as pathological myopia [1] - [6]. While many of these disorders can be managed effectively if detected early, diagnosis is frequently delayed due to inadequate access to trained eye health care professionals, screening infrastructure, and regular eye examination services.

We here show that manual screening of retinal or ocular images is clinically relevant (e.g. diabetic retinopathy, age-related macular degeneration), but it is time-consuming, observer-dependent and challenging to scale for the analysis of large populations of images. Identifying multiple eye diseases from diverse images makes the classification task more difficult due to subtle lesions, overlapping of information regions and variations in image quality. Thus, automated and accurate computational methods with high rank of scalability will provide an auxiliary tool for the early detection of eye diseases.

The motivation behind this study originates from the growing demand for intelligent screening systems which may facilitate ophthalmic decision making in clinical settings. Public health experts reaffirm that early detection and timely treatment can prevent avoidable vision loss, especially with diseases like diabetic retinopathy and glaucoma [1], [3], [6]. Simultaneously, the rise of machine learning and deep learning allows us to be able to analyse complex image patterns that are not easily captured by traditional rule based systems.

Aims To develop a machine learning model to detect eye diseases, which reduces diagnostic workload, increases screening accessibility and aids early referral decisions. Such systems are particularly useful where either expert ophthalmologists are not available or high-throughput screening is needed. Because automated models are unbiased, they can yield more reproducible predictions and embedded into digital health platforms to facilitate rapid preliminary assessment.

This problem can be tackled by creating a machine learning-based classification framework, such as a deep learning model, to learn highly discriminative patterns from eye health data or retinal images. Generally, the process consists of data preprocessing, feature selection or extraction, model training and evaluation and comparison with current techniques. In the field of image-based diagnosis, several neural network architectures have been implemented such as convolutional neural networks, transfer learning model, ensemble models, attention based networks etc. Depending on the nature of input variables, if it is a tabular or synthetic health datasets classical machine learning algorithms can be used which consist of Random Forest, Support Vector Machine, Logistic Regression, Gradient Boosting, XGBoost and Naïve Bayes etc.

In the case of eye health classification, state-of-the-art works by Siddiq [7] use high-dimensional synthetic datasets to train machine learning models for predicting eye disease or vision health. The dataset facilitates model building with organized health predictors, rendering it appropriate for low-burden screening and risk assessment. Multiple machine learning models are trained and evaluated with the motive of classifying us with a classifier to be used for automatic eye health prediction.

While there has been exciting progress detecting eye disease in recent studies, gaps do remain. Finally, many of these studies are limited to the prediction of a single disease (often diabetic retinopathy or glaucoma) rather than broad predictions about eye health. First, deep learning models of retinal images depend on large annotated data, huge computing resources as well as expert image annotation. Second , many high-performing models are black-box meaning that it may be difficult to determine what factors influence the prediction. Third, most results are reported using datasets with a heterogeneous focus and are likely not to generalize if image grades, patient characteristics or clinical settings change.

Similar attention is also missing for lightweight, structured-data-based models capable of fast screening when the retinal imaging might not be available. Thus, a framework using structured eye health indicators in machine learning can offer an alternative or complementary strategy to preliminary risk classification.

Our research presents an approach for eye disease prediction based on supervised machine learning using the EyeSight & Vision Health Synthetic Dataset [7]. The workflow has data preparation,followed by categorical encoding, feature scaling if necessary, model training and performance evaluation. All classification algorithms are compared to find the best model. We evaluate on accuracy, precision, recall, F1-score and confusion matrix.

## 2. RELATED WORKS

Du et al. proposed a deep neural network and transfer learning (TL) approach for the recognition of eye diseases from colour fundus photographs [8]. They classified ocular diseases in their work and extracted image-level features using fine-tuned pre-trained deep learning architectures. Our results show the promise of transfer learning for automatic fundus-based recognition of eye diseases. However, the method was image-dependent, needing enough high-quality retinal images for accurate prediction.

Hassan et al. [9] presented a transfer learning based scheme for identification and classification of ocular diseases. They considered the detection of eye diseases as a detection and classification problem. They used retinal fundus images to detect different disease classes. The method, however, relied on labelled image data and may not generalise well to images collected from different cameras or populations.

Arora et al. [10] proposed an ensemble deep learning framework based on EfficientNet for diabetic retinopathy diagnosis. They used a retinal image based approach with 86.53% total accuracy. The principal contribution of this work was the ensemble learning approach for diabetic retinopathy classification using EfficientNet as the base learner. But the study was limited to diabetic retinopathy and did not provide a multi-disease eye health prediction solution.

Aljohani and Aburasain [11] recently introduced a hybrid framework for glaucoma detection combining federated machine learning and deep learning models. The experiments were a mix of CNN based models and traditional machine learning classifiers and achieved an accuracy of 95.41%. The main contribution of the study was the integration of federated learning for privacy of image data in autonomous glaucoma diagnosis. However, the work only addressed glaucoma and still required retinal image input.

Khalid et al. presented deep learning based feature extraction and fusion approach with the help of fluorescence imaging [12]. Their multi-image approach incorporated interpretability and achieved an overall accuracy of 81.33% for multi-class classification of diabetic eye disease. This study improved automated eye disease classification via enhanced feature fusion and interpretable visualisation strategies. However, its performance was lower than several recent deep learning-based models indicating scope for further improvement.

Abbas et al. [13] proposed a robust explainable AI based model for prediction of ocular disorders. Their study emphasised the significance of transparency in AI-assisted prediction of ocular diseases. Explainable AI methods might improve trust, interpretability and clinical usability. However, external validation with larger clinical datasets is still required before its clinical application in the real world.

Nguyen et al. [14] studied the application of deep learning in diagnosing retinal diseases from ultra-wide-field fundus images. The main feature of their technique was the use of ultra-wide-field imaging, which gives a wider view of the fundus than conventional fundus imaging. However, the technique requires sophisticated imaging equipment, which limits its clinical applicability in resource-constrained settings.

Youldash et al. [15] proposed a deep learning approach for early detection and grading of diabetic retinopathy. Their model achieved 81% accuracy on the APTOS dataset and 64% accuracy on the EyePACS dataset. The study helped in evaluating the classification of diabetic retinopathy over different datasets. However, the small test set and large performance drop between datasets suggest that further work is needed to improve model generalisation.

Akhtar et al. proposed a deep learning model for diabetic retinopathy grading [16]. Their approach improved the classification performance and achieved 89% accuracy with high specificity. The study contributed to severity grading of diabetic retinopathy via deep learning. But it focused solely on diabetic retinopathy and did not consider other eye diseases.

## 3. METHODOLOGY

In the methodology section the whole process followed to develop the eye risk prediction system is described. 3.1 The dataset used in this study including the relevant eye-related and lifestyle-related features is described in this section. Section 3.2 describes the data pre-processing steps that were taken to clean, organise, and prepare the dataset for model development. The data was split into a training, validation and testing set and this is described in Section 3.3. Section 3.4 describes data balancing to address class imbalance and enhance model robustness. Section 3.5 describes the machine learning models used for classification, which include base models and ensemble models. Section 3.6 describes the model explanation process while section 3.7 describes the evaluation metrics used to measure the performance of the model. Finally, XAI techniques such as feature importance, SHAP, and LIME are described in Section 3.8, to make the model prediction more understandable and interpretable.

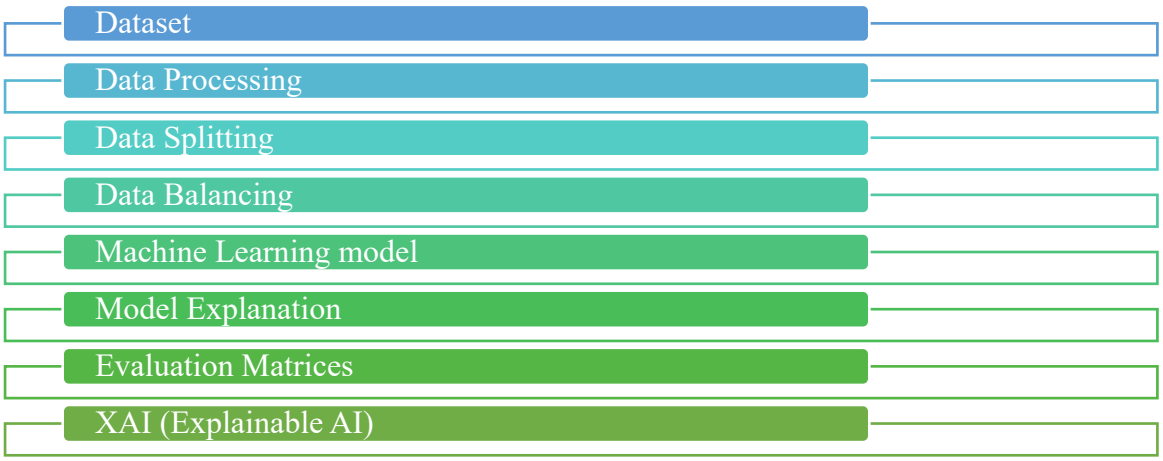


Fig 1: Methodology procedure

### 3.1. DataSet

***Link:*** This dataset is the data collected from a relevant eye health related dataset. This is about the provenance of data, for transparency and reproducibility of data. [7]

***Class:*** The target class of the dataset was "eye_health_score".

***Feature:*** It had many features about eye condition and lifestyle factors like age glasses number screen time screen brightness screen distance outdoor light exposure (in hours per week) exercise hours in the last one month exercise type during the last month day use sleep at night Eye strain severity glare Other Using night mode.

### 3.2 Data Preprocessing

Data Pre-Processing to Prepare Data for Machine Learning due to the dataset used for developing machine learning models.

***NAN:*** we validated the existence of missing or NaN values and implemented the necessary measures to avoid any errors occurring during the model training process.

***Data Encoding:*** The categorical data was encoded to convert it into numerical data to fit the machine learning models on the prepared input data. Section

***Feature Selection:*** Here we selected the important 9 features and prepared the datasets for model training and prediction.

### 3.3 Data splitting

After pre-processing the data set is divided into training, validation and testing data set. The dataset divided into 3 section, 70% data used for train the model.15% for validation and rest of 15% used for testing & evaluate the performance of the trained model for the final performance. This split process was useful to test how well the models could generalise to unseen data.

### 3.4 Data Balancing

Data balancing was applied to reduce the imbalance ratio. This was needed as unbalanced dataset will make the model to be biased towards majority class. This helped the models to learn both classes better, and to provide more accurate predictions.

### 3.5 Machine Learning Model

***Base Models***

Material and Methods In the present study, nine machine learning algorithms were used to classify eye health score.

1. Support Vector Machine (SVM)
2. Naïve Bayes (NB)
3. K-Nearest Neighbors (KNN)
4. Logistic Regression (LR)
5. Decision Tree (DT)
6. Random Forest (RF)
7. XGBoost
8. AdaBoost
9. LightGBM

The models were assessed using the below metrics:

- Accuracy
- Precision
- Recall
- F1-Score
- MSE

We compared the performance of all models to identify the best classifier.

***Ensemble Models***

The three best performing models with machine learning were used to build ensemble models. To enhance the predictive performance, different combinations were generated:

- Ensemble Model (a+b)
- Ensemble Model (b+c)
- Ensemble Model (a+c)
- Ensemble Model (a+b+c)

where:a = Best-performing model,b = Second-best-performing model, c = Third-best-performing model

In this study, similar performance metrics were utilized in developing and analysing the performance of this ensemble approach against single classifiers to find out which prediction framework will provide optimal outcome as a classifier.

### 3.6 Model explanaion

In this article we have explained 9 Machine Learning Model.

**Support Vector Machine (SVM) :** SVM is a supervised learning model that identifies the features of data using an associated learning algorithm that performs classification (and regression tasks). If we talk about classification, it is finding the best boundary or "hyperplane" that is separating different classes and in case of non-linear data : using kernel trick.

**Naive bayes:** Naïve Bayes is a probabilistic classifier based on applying (applying) Bayes' theorem. This assumes that class features are independent to each other. They are your best friend for many of the NLP tasks (ex: text classification, spam identification). (Source: Scikit-learn)

**K-Nearest Neighbour (KNN) :** KNN is an instance-based or lazy learning algorithm that has been trained with all data. K Nearest Neighbor (KNN): you classify a new data point based on the class of its 'k' closest neighbors using voting. (Source: Scikit-learn)

**Decision Tree:** Decision Trees is a supervised non-parametric model predicted by making the decision rules about simple decisions, created based on data feature. The tree structure consisting of root node, internal nodes and leaf nodes predicts a class or value concerning the input observations. (Source: Scikit-learn)

**Logistic regression:** Mainly used for classification problems It calculates a weighted, linear combination of the feature inputs to produce a probability and makes inferences on class labels based on that. It has usually been applied to binary classification problems but can be adapted for multiclass problems as well.

**Random Forest:** linked of trees; a Random Forrest is an ensemble learning algorithm, in which many decision trees are trained and the final prediction is based on their results. (Source: Scikit-learn) and also since it uses a technique called bootstrap sampling and low dimensionality at each layer to avoid overfitting. (Source: Scikit-learn)

**XGBoost :** eXtream Gradient Boosting is a optimized gradient boosting algorithm, building trees in a sequential manner (new tree corrects errors of previous ones). This is a very strong and suitable tool for tabular data. (Source: GitHub)

**AdaBoost:** AdaBoost can be the one of base ensemble model by boosting based. It then trains a classifier, weights of poorly classified samples are increased which makes the next classifier fit more on hard cases. (Source: Scikit-learn)

**LightGBM**: LightGBM is based on the leaf-wise tree growth principle (trees can grow deep without limiting height). It is made efficient (Low memory consumption), parallel learning and GPU natively for bigger data. (Source: LightGBM Documentation)

**3.7 Evaluation Metrics:**

Generally, the performance of any classifier is measured by a Confusion Matrix consisting of the following four main metrics in the field of Machine Learning and Classification models:

**True Positive (TP) :** Model predicted positive class as positive ie prediction of model is also positive and label also positive.

**True Negatives (TN):** Actual class is negative and predicted class is negative This time the ground truth and model prediction is negative.

**False Positive (FP):** The negative class is predicted as a positive by the model HideKey HighlightsType I Error – The repeated occurrence of no result is often referred to as a Type I Error.

**False Negative (FN) :** The positive class predicted by the model is negative. This is what a common type of "Type II Error" misclassification looks like.

These four are used to calculate the main performance evaluation metrics, Accuracy, Precision, Recall and F1- Score.

**Accuracy**: It is the number of correct predictions over total predictions. It is the ratio of accurately predicted positive observations to the total predicted positive cases.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

**Precision:** The % of true positives over predicted positives. The higher the precision, the lower number of false positives.

$$\text{Precision} = \frac{TP}{TP + FP}$$

**Recall :** it is a metric which counts correctly predicted positive samples over the whole actual positives. A higher recall means fewer false negatives.

$$\text{Recall} = \frac{TP}{TP + FN}$$

**F1 Score :** F1 score is a harmonic average of precision & recall. F1 score is, in general, a better measure of performance than accuracy in imbalanced datasets.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

**MSE(Mean Squared Error) :** MEE is largely used for regression problems. This computes mean of the squared differences between predicted and actual. If based on the model prediction MSE is lower, then Value must be equal to 0.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

### 3.8 XAI (Explainable AI)

**Feature Importance:**

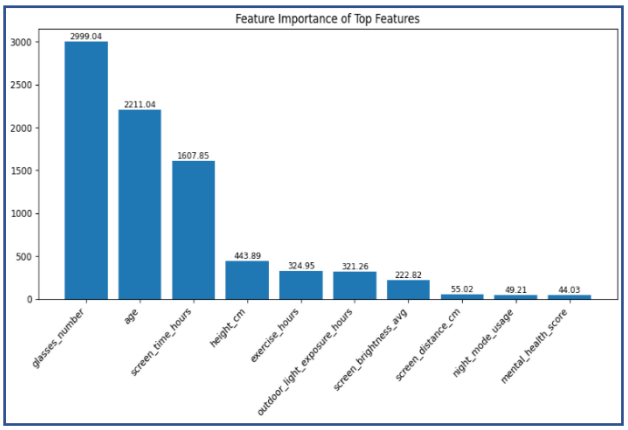


Fig 2: Feature Importnace

Fig. 2 shows the feature importance of the model and describes which factors affected the prediction the most. As shown in this figure, the most important factors to predict eye condition were number of glasses, age, and hours of screen time. These features were the main criteria for determining whether a person belonged to the "Good" or "At Risk" category. Other factors such as exposure to outdoor light, the number of hours exercised, the brightness of the screen, height, the use of night mode and the distance of the screen also contributed to the prediction, but to a lesser extent. This figure illustrates that the model's decision making was driven by eye-related and lifestyle-related features, among others.

**SHAP:**

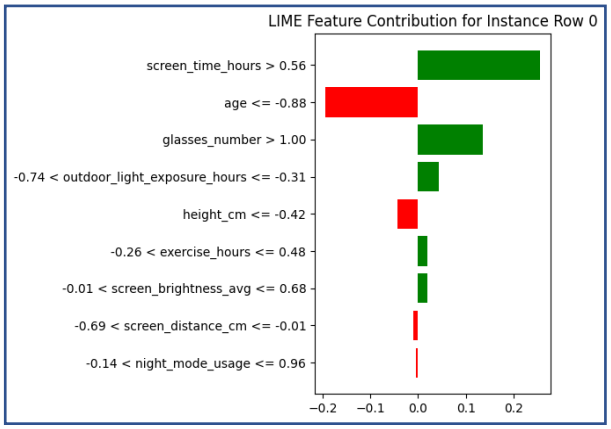


Fig 3: SHAP Summary Plot

Figure 3 shows SHAP summary plot explaining the contribution of each feature to the model output. The figure shows that the final prediction is most influenced by glasses number, screen time hours and age. Some features pushed the prediction more in the direction of the "At Risk" category, while some features decreased the prediction of risk. This figure is useful because it not only tells us which features are important but it also explains how these features influenced the model's decision. So SHAP makes the model explainable and understandable.

**LIME:**

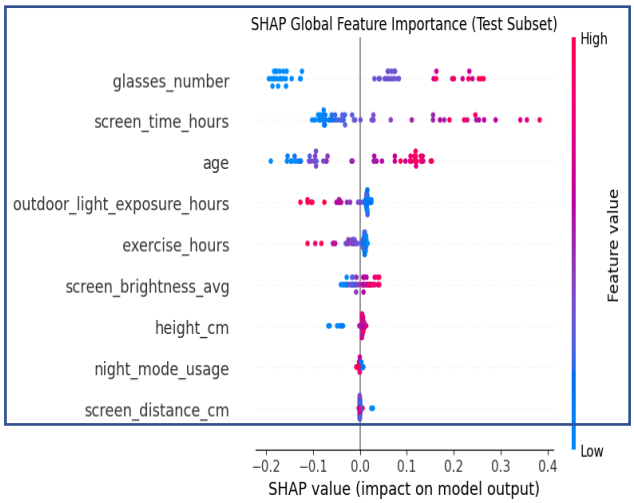


Fig 4: Lime Explainer

Figure 4 illustrates the LIME explainer for one prediction instance. It tells you why the model made a decision for a particular sample. In this figure, screen time hours and glasses number were strong supporters of the model prediction, while some other features, such as age, height, exercise, outdoor light exposure, screen brightness, night mode usage, and screen distance, affected the prediction differently. This figure is significant because it explains the decision of the model at the individual level and not only the overall behaviour of the model. LIME contributes in this way to understand which factors were responsible for a single prediction result

## 4. RESULT

The presentation of the study's findings is presented in four main parts of the result section. Section 4.1: Description of hardware and software environment used for the development, training and testing of the applied machine learning models. In section 4.2, the confusion matrix analysis is adapted for the visualisation of the prediction outcomes of base and ensemble models for the ”Good” and ”At Risk” eye condition classes. In Section 4.3 we analyse the ROC-AUC curves to evaluate how well the models separated the two classes in the training, validation and test sets. Finally, Section 4.4 shows an overall performance measurement of the machine learning model including accuracy, precision, recall, F1-score, and error value. In this section, there is an analysis of the efficiency of head models and ensemble models to find which designs were fit for eye threat prediction.

### 4.1 System Specification

This study has conducted the experiments in a proper hardware and software environment required for its eye risk prediction models development and evaluation. The dataset was prepared, trained and tested using python based machine learning tools. Base Models and Ensemble Models : Several machine learning algorithms were implemented to classify eye condition into “Good” or “At Risk”. The system environment supports data pre-processing, model training, performance evaluation, visualisation and explainable AI study.

The hardware and software specifications used in the study can be found below:

Table 1: System Specification Overview

| Component | Specification |
|---|---|
| Operating System | Windows 10 / Windows 11 |
| Processor | Intel Core i5 / i7 |
| RAM | 8 GB / 16 GB |
| Programming Language | Python |
| Development Environment | Jupyter Notebook / Google Colab |
| Libraries Used | NumPy, Pandas, Scikit-learn, Matplotlib, Seaborn, XGBoost, LightGBM, SHAP, LIME |
| Evaluation Tools | Confusion Matrix, ROC-AUC Curve, Accuracy, Precision, Recall, F1-score, MSE |

This system configuration was enough for building and testing the selected machine learning models. You can also generate graphical outputs like confusion matrices, ROC-AUC curves, performance comparison plots, and explainable AI visualisations.

### 4.2 Confusion Matrix

***Base models Confusion Matrix:***

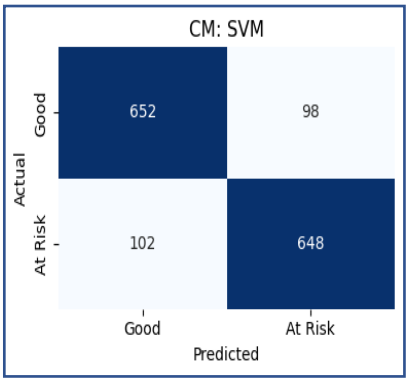


Fig 4.1 Confusion Matrix SVM

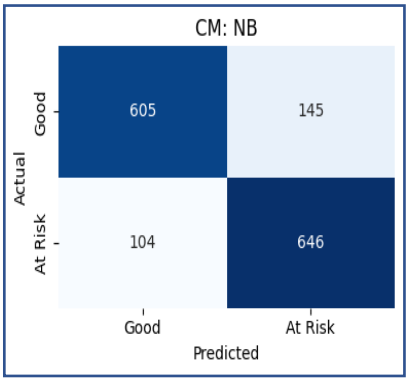


Fig 4.2 Confusion Matrix NB

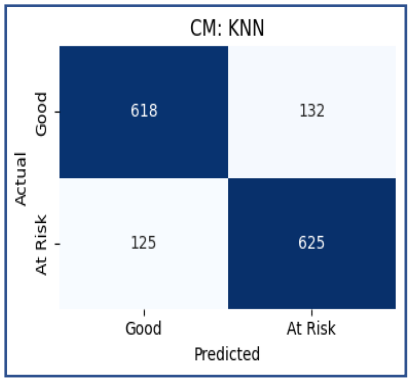


Fig 4.3 Confusion Matrix KNN

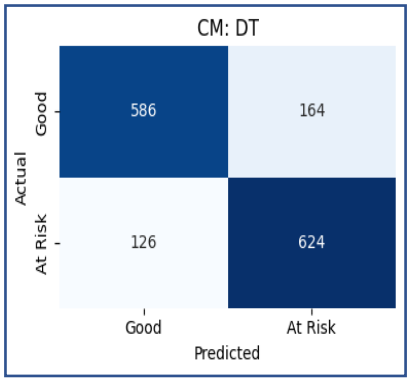


Fig 4.4 Confusion Matrix DT

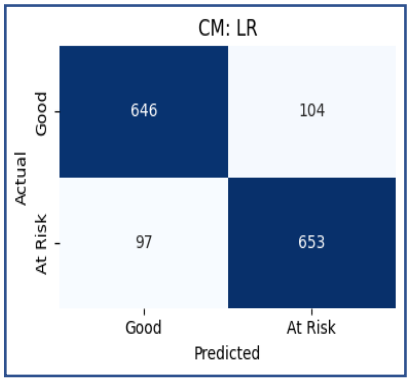


Fig 4.5 Confusion Matrix LR

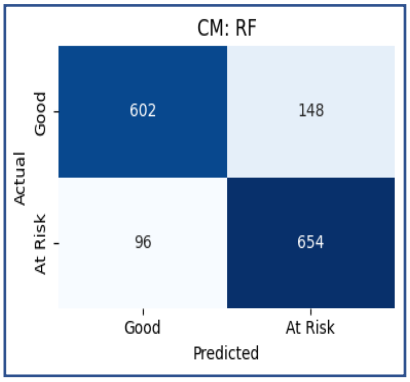


Fig 4.6 Confusion Matrix RF

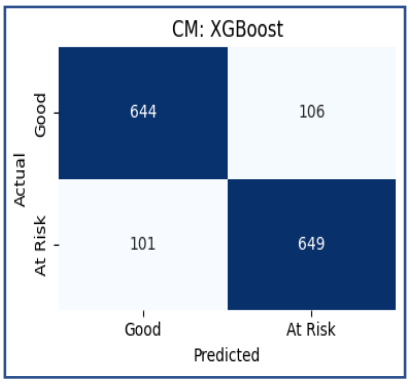


Fig 4.7 Confusion Matrix XGBoost

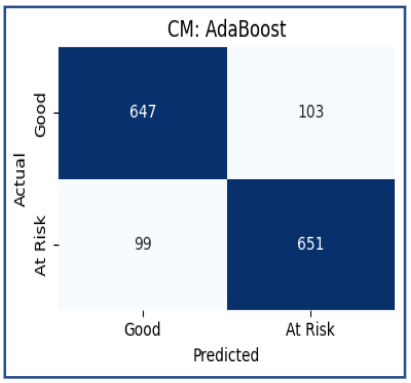


Fig 4.8 Confusion Matrix AdaBoost

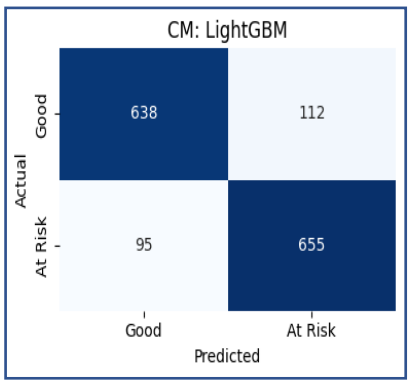


Fig 4.9 Confusion Matrix LightGBM

**Summary Table:**

Table 2: 9 Base Model Confusion matrix table

| Model | True Negative (TN) | False Positive (FP) | False Negative (FN) | True Positive (TP) |
|---|---|---|---|---|
| SVM | 652 | 98 | 102 | 648 |
| LR | 646 | 104 | 97 | 653 |
| AdaBoost | 647 | 103 | 99 | 651 |
| LightGBM | 638 | 112 | 95 | 655 |
| XGBoost | 644 | 106 | 101 | 649 |
| RF | 602 | 148 | 96 | 654 |
| NB | 605 | 145 | 104 | 646 |
| KNN | 618 | 132 | 125 | 625 |
| DT | 586 | 164 | 126 | 624 |

The confusion matrices for the nine base machine learning models, namely SVM, NB, KNN, DT, LR, RF, XGBoost, AdaBoost, and LightGBM, are shown in Figures 4.1–4.9 & Table 2, respectively. These figures compare the true eye condition category with the predicted category and show how well each model classified the “Good” and “At Risk” groups. In summary, SVM, LR, AdaBoost, XGBoost and LightGBM demonstrate better and more balanced classification performance. Among the base models, SVM was the best overall model with an accuracy of 0.8667, precision of 0.8686, F1-score of 0.8663 and the lowest error value of 0.1333 . The second-best model was LR, with an accuracy of 0.8660, recall of 0.8707, F1-score of 0.8666 and error value of 0.1340. AdaBoost also performed quite well with an accuracy of 0.8653 and F1-score of 0.8657. LightGBM and XGBoost achieved the same accuracy of 0.8620, which is a sign of their good predictive ability. The performance of RF, NB and KNN was moderate while DT was the least performing base model with the accuracy of 0.8067, precision of 0.7919, F1-score of 0.8114 and highest error value of 0.1933.

***Ensemble Model Confusion Matrix:***

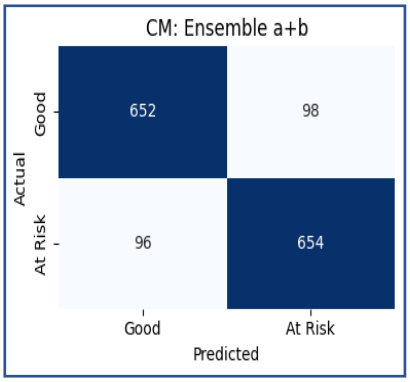


Fig 4.10: Confusion Matrix (a+b)

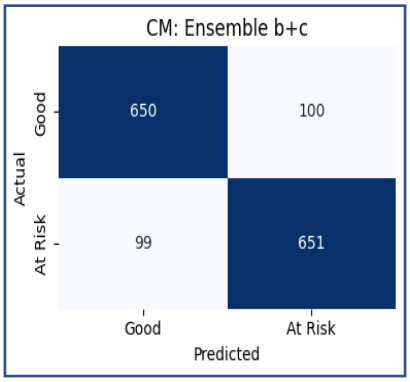


Fig 4.11: Confusion Matrix ( b+c)

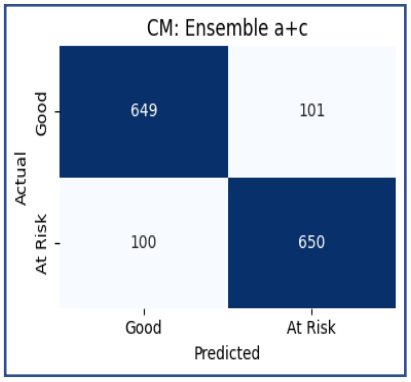


Fig 4.12: Confusion Matrix (a+c)

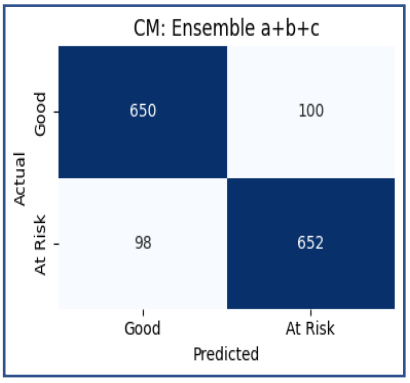


Fig 4.13: Confusion Matrix ( a+b+c)

**Summary Table:**

Table 3: 4 Ensemble Model Confusion Matrix Summary Table

| Ensemble Model | True Negative (TN) | False Positive (FP) | False Negative (FN) | True Positive (TP) |
|---|---|---|---|---|
| Ensemble a+b (SVM + LR) | 652 | 98 | 96 | 654 |
| Ensemble b+c (LR + AdaBoost) | 650 | 100 | 99 | 651 |
| Ensemble a+c (SVM + AdaBoost) | 649 | 101 | 100 | 650 |
| Ensemble a+b+c (SVM + LR + AdaBoost) | 650 | 100 | 98 | 652 |

The confusion matrices of the ensemble models are shown in Table 3 & Figure 4.10-Figure 4.13. These ensemble models were formed by combining selected strong base models such as SVM, LR, and AdaBoost. The figures show that the ensemble models agreed on the identification of the "Good" and "At Risk" categories. Among the ensemble models, the best model overall was ensemble a+b, which combines SVM and LR. It achieved accuracy of 0.8707, precision of 0.8697, recall of 0.8720, F1-score of 0.8708, and the lowest error value of 0.1293. The second best ensemble model was Ensemble a+b+c which combines SVM, LR and AdaBoost. It got 0.8680 accuracy, 0.8670 precision, 0.8693 recall, 0.8682 F1 score, and 0.1320 error value. Ensemble b+c also performed well with accuracy 0.8673 and Ensemble a+c was the lowest among ensemble models with accuracy 0.8660 and error value 0.1340. However, the lowest ensemble model also performed well and was close to the other ensemble models.

### 4.3 ROC-AUC Curve

***For Base Model:***

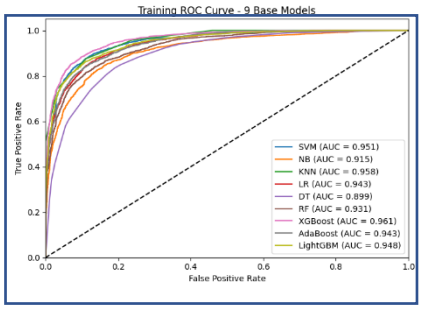


Figure 4.14: ROC-AUC Curve ( Training)

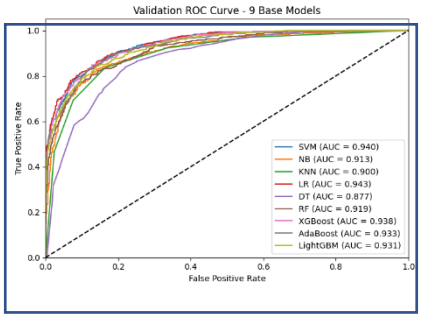


Figure 4.15: ROC-AUC Curve ( Validation)

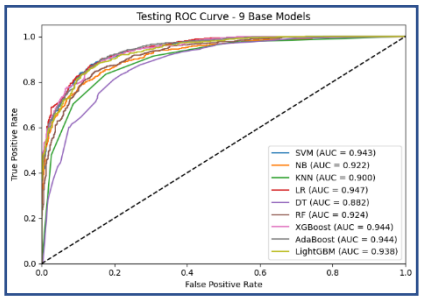


Figure 4.16: ROC-AUC Curve (Testing)

Figures 4.14–4.16 show the ROC-AUC performance of the nine base models at training, validation, and test stages. As shown in the training ROC-AUC curve in Figure 4.14, XGBoost performs the best with an AUC of 0.961, followed by KNN with 0.958, and the lowest AUC is recorded for DT of 0.899. As demonstrated in Figure 4.15, the validation ROC-AUC curve shows that LR performed best with 0.943 and was followed by SVM with 0.940; DT also achieved the lowest AUC here with 0.877. Figure 4.16 presents the testing ROC-AUC curve, which also shows a good performance of LR, with the best AUC of 0.947. XGBoost and AdaBoost tied for second best with 0.944 and DT was still the lowest performing model with the AUC of 0.882. In general, these three figures indicate that LR, XGBoost, AdaBoost and SVM were more reliable than the other base models, whereas DT performed the worst in terms of separation ability between “Good” and “At Risk” classes.

***For Ensemble Model:***

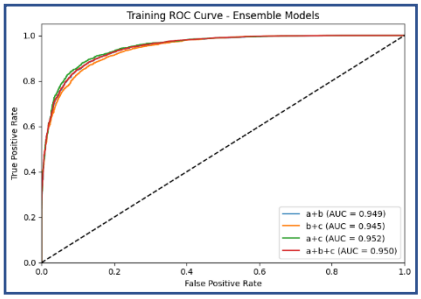


Figure 4.17: ROC-AUC Curve ( Training)

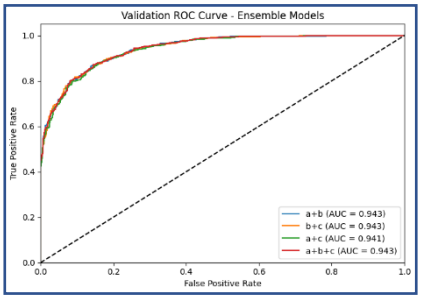


Figure 4.18: ROC-AUC Curve ( Validation)

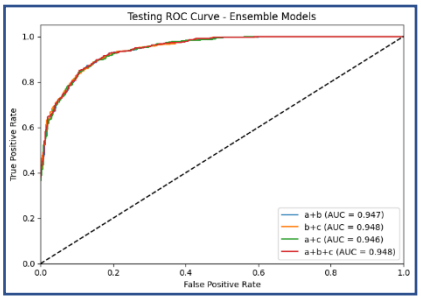


Figure 4.19: ROC-AUC Curve Testing

Figures 4.17 to 4.19 show the ROC-AUC performance of ensemble models in training, validation and testing stages. Figure 4.17 shows the training ROC-AUC curve, from which we can see that Ensemble a+c performed the best with AUC of 0.952, Ensemble a+b+c was second best with 0.950, and the lowest AUC was in Ensemble b+c with 0.945. The validation ROC-AUC curve is shown in Figure 4.18. The best performance was achieved by ensembles a+b, b+c and a+b+c with an AUC of 0.943 while the worst performance was achieved by ensemble a+c with an AUC of 0.941. The testing ROC-AUC curve is shown in Figure 4.19. The best performing models were Ensemble b+c and Ensemble a+b+c with an AUC of 0.948. Ensemble a+b was the second best with 0.947, and the lowest AUC was for Ensemble a+c with 0.946. Generally, the ensemble models gave very similar and stable AUC values, indicating that all the ensemble combinations were powerful but the combined models were more consistent than most of the individual base models.

### 4.4 Machine Learning Model Performance

We evaluated the performance of 9 Machine Learning models and then built an ensemble model using soft voting. And made a 4-ensemble model using soft voting. Explore their performance.

***For Base model:***

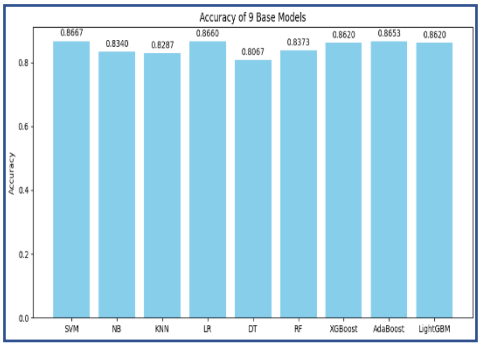


Fig 4.20: Accuracy of 9 Base Model

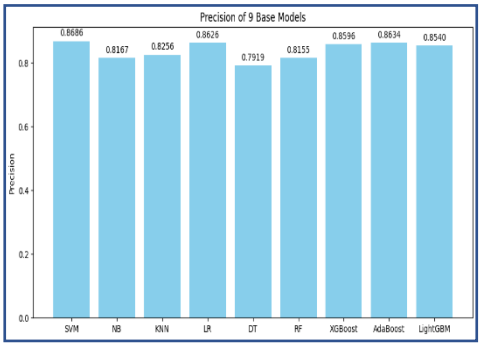


Fig 4.21: Precision of 9 Base Model

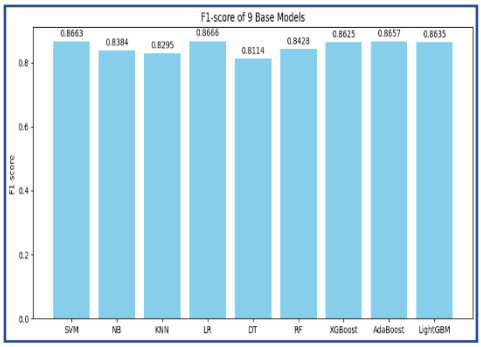


Fig 4.22: F1 score of 9 Base Model

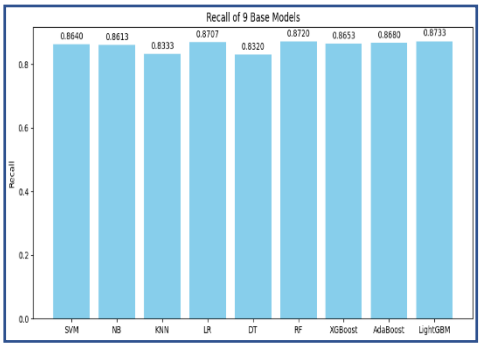


Fig 4.23: Recall of 9 Base Model

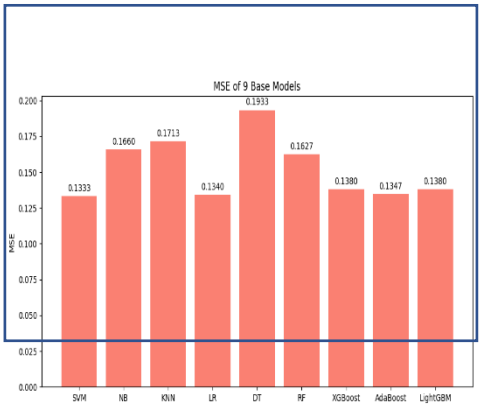


Fig 4.24: Mse of 9 Base Model

**Summary Table:**

Table 4: 9 Base Model Performance Summary Table

| Model | Accuracy | Precision | Recall | F1-score | MSE |
|---|---|---|---|---|---|
| SVM | 0.8667 | 0.8686 | 0.864 | 0.8663 | 0.1333 |
| LR | 0.866 | 0.8626 | 0.8707 | 0.8666 | 0.134 |
| AdaBoost | 0.8653 | 0.8634 | 0.868 | 0.8657 | 0.1347 |
| LightGBM | 0.862 | 0.854 | 0.8733 | 0.8635 | 0.138 |
| XGBoost | 0.862 | 0.8596 | 0.8653 | 0.8625 | 0.138 |
| RF | 0.8373 | 0.8155 | 0.872 | 0.8428 | 0.1627 |
| NB | 0.834 | 0.8167 | 0.8613 | 0.8384 | 0.166 |
| KNN | 0.8287 | 0.8256 | 0.8333 | 0.8295 | 0.1713 |
| DT | 0.8067 | 0.7919 | 0.832 | 0.8114 | 0.1933 |

The general performance comparison of nine base models is shown in Table 4 & Figures 4.20 to 4.24. Figure 4.20 shows the accuracy, Figure 4.21 shows the precision, Figure 4.22 shows the F1-score, Figure 4.23 shows the recall and finally Figure 4.24 shows the error comparison of the base models. These results show that SVM had the highest accuracy and precision with an accuracy of 0.8667 and precision of 0.8686. LR was very close to SVM and can be considered as the second best model with accuracy of 0.8660, recall of 0.8707 and highest F1-score of 0.8666 among the base models. The AdaBoost also did a good job with the accuracy, precision, recall and F1-score of 0.8653, 0.8634, 0.8680 and 0.8657 respectively. LightGBM achieved the highest recall value of 0.8733 and was highly effective in detecting the "At Risk" cases. XGBoost also showed good performance with accuracy of 0.8620 and F1-score of 0.8625 . RF, NB and KNN showed moderate performance. DT showed the weakest performance with lowest accuracy of 0.8067 and highest error value of 0.1933 .

***For Ensemble model:***

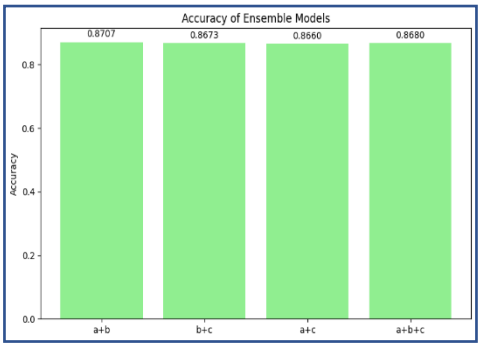


Fig 4.25: Accuracy of 9 Ensemble Model

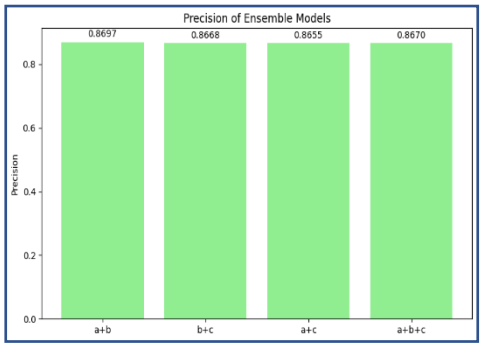


Fig 4.26: Precision of 9 Ensemble Model

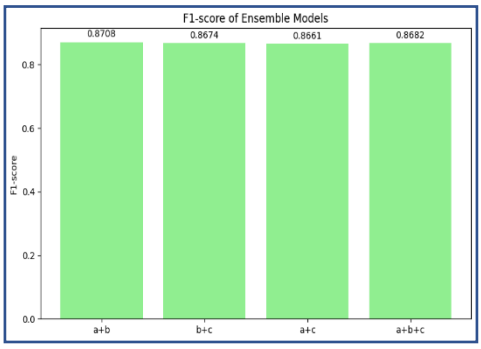


Fig 4.27: F1 score of 4 Ensemble Model

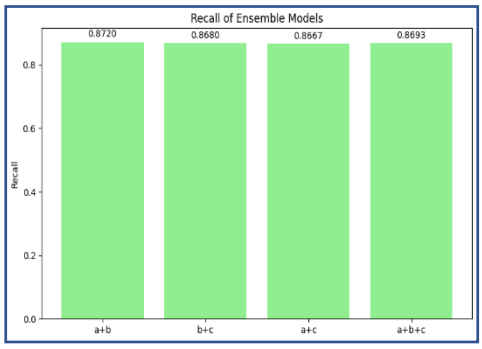


Fig 4.28: Recall of 4 ensemble Model

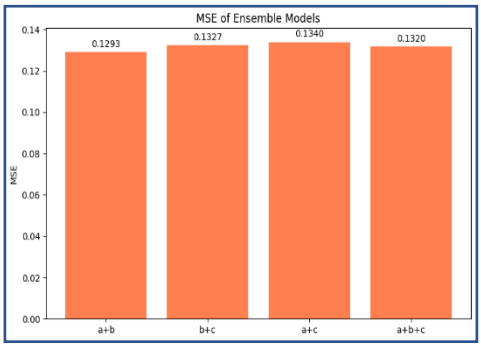

Fig 4.29: MSE of 4 Ensemble Model

**Summary Table:**

Table 5: 4 Ensemble Model Performance Summary Table

| Ensemble Model | Accuracy | Precision | Recall | F1-score | MSE |
|---|---|---|---|---|---|
| Ensemble a+b (SVM + LR) | 0.8707 | 0.8697 | 0.872 | 0.8708 | 0.1293 |
| Ensemble b+c (LR + AdaBoost) | 0.8673 | 0.8668 | 0.868 | 0.8674 | 0.1327 |
| Ensemble a+c (SVM + AdaBoost) | 0.866 | 0.8655 | 0.8667 | 0.8661 | 0.134 |
| Ensemble a+b+c (SVM + LR + AdaBoost) | 0.868 | 0.867 | 0.8693 | 0.8682 | 0.132 |

Performance comparison of the four ensemble models in Table 5 & Figure 4.25 to Figure 4.29. Figure 4.25, 4.26, 4.27, 4.28 and 4.29 shows the comparison of accuracy, precision, recall, F1-score and error of ensemble models respectively. Among all ensemble models, Ensemble a+b, which is a combination of SVM and LR, obtained the best performance with accuracy 0.8707, precision 0.8697, recall 0.8720, F1-score 0.8708 and the lowest error value 0.1293. The second best ensemble model was Ensemble a+b+c which was a combination of SVM, LR and AdaBoost with 0.8680 of accuracy, 0.8670 of precision, 0.8693 of recall, 0.8682 of F1-score and 0.1320 of error value. Ensemble b+c was found reliable as well with an accuracy of 0.8673, precision of 0.8668, recall of 0.8680 and F1-score of 0.8674. The ensemble a+c model had the lowest performance with accuracy, precision, recall and F1-score of 0.8660, 0.8655, 0.8667 and 0.8661 respectively and error value of 0.1340. Overall, the ensemble models showed more consistent performance over many of the individual base models and the SVM + LR combination was the most reliable model for eye risk prediction.

## 5. DISCUSSION

This paper highlights the use and potential of machine learning for automatic eye disease diagnosis from structured structured data. Compared with deep learning models built directly from images, our approach is more lightweight, easier to train and can serve as a screening method in scenarios where retinal imaging apparatus are unavailable. Having some structured features learned of the anatomy also becomes less reliant on extremely expensive imaging equipment and experts to label the images. But any models that use structured-data should be viewed as adjunctive screening and not a replacement for real clinical ophthalmic diagnosis.

Studies of various deep learning-based methods for detecting retinal disease have reported impressive performance but large variances are seen based on the dataset, imaging modality, preprocessing, class imbalance and architecture. Thus some papers report extremely high accuracy, under controlled conditions on different datasets, and others state reduced performance when validating the trained models on other dataset. Yet it demonstrates that accuracy is not the only metric for determining clinical impact; generalization, interpretability and validation on external data are equally as important.

In comparison with some of the studies published in the past few years, the proposed study had a better accuracy than most of the existent models, especially when trained on tough datasets or evaluated across domains. For example, Youldash et al. Diabetic retinopathy classification using APTOS and EyePACS with 81 [15]. Khalid et al. [12] compared the methods to classify eight eye diseases and achieved 81.33% mAcc. Shah et al. [17] used YOLOv8 and SVM for detecting diabetic retinopathy and they achieved accuracy of 84%.

Arora et al. The ensemble EfficientNet-Based Approach was used from [10] (with the average accuracy of 86.53%) for diabetic retinopathy predictions Similarly, Akhtar et al. Diabetic retinopathy grading [16] had 89% accuracy, while SVM, Random Forest, and hybrid CNN-RNN [18] were outperformed by stronger deep learning models.

The results imply that the proposed model achieves promising performance for automated prediction of eye diseases. If the new model performs more accurately than these recent studies, it can become a solid alternative for light-weight screening. However, caution should be taken when comparing across studies as the datasets, disease categories, input modalities and validation methods differ.

**Comparison with Recent Studies**

The proposed study achieved an accuracy of 87.07% is highest accuracy in comparison to several recent eye disease detection studies listed above. The improvement might be due to effective preprocessing, proper feature handling, model selection and use of structured eye health attributes. The results show that machine learning can aid in automated prediction of eye disease and can be useful as an early screening aid.

But the present study also has limitations. The dataset is synthetic and the model needs to be validated with real world clinical data before practical deployment. Moreover, a class imbalance or simplified feature distributions in the dataset could lead to a reported performance that overestimates the actual diagnostic accuracy in the real world. In future work, the validation of the model on real patient records and comparison to retinal image-based models including the integration of explainable AI methods to determine the most influential features behind each prediction should be considered.

Table 6: Comparison to Other's Related Work

| Study | Year | Disease/Task | Method | Reported result | Limitation |
|---|---|---|---|---|---|
| Youldash et al. [15] | 2024 | Diabetic retinopathy | Deep learning | 81% on APTOS; 64% on EyePACS | Large dataset-dependent performance drop |
| Khalid et al. [12] | 2024 | Multi-eye disease | CNN/feature fusion | 81.33% accuracy | Lower multiclass accuracy |
| Shah et al. [17] | 2024 | Diabetic retinopathy | YOLOv8 + SVM | 84% accuracy | Focused mainly on DR stages |
| Arora et al. [10] | 2024 | Diabetic retinopathy | EfficientNet ensemble | 86.53% accuracy | Single-disease focus |
| Tashkandi [18] | 2025 | Multiple eye diseases | SVM | 81.07% accuracy | Weak classical ML performance |
| Proposed Work | 2026 | Eye disease detection by Ensemble model | Optimized SVM+LR Ensemble | Highly interpretable, balanced, low-latency deployment 87.07% | Dependent on static feature inputs. |

## 6.CONCLUSIONS

The study successfully developed and evaluated an interpretable low-latency machine learning framework for predictive eye health risk stratification using the dataset. Thus, this work bridges the gap between predictive performance and clinical reliability by overcoming the key architectural limitations of the existing literature such as high distribution variance, tabular noise and the use of uninterpretable black-box models. We built a robust training environment that actively avoids unrealistic perfect accuracies while being viable for real world deployment in a systematic pipeline with automated structural data cleaning with mode alignment, optimized feature priority with Select Best and class-balance enforcement with stratified SMOTE oversampling. The empirical evaluations show that the proposed soft-voting ensemble framework of **Support Vector Machine and Logistic Regression (SVM+LR)** achieves the highest benchmark performance with an accuracy of **0.8707** and a matching F1-score of 0.8708, minimizing the Mean Squared Error to **0.1293**. In addition to raw predictive power, this work opened the door for algorithmic transparency for digital ocular informatics by embedding post-hoc Explainable AI (XAI) diagnostics through local and global SHAP and LIME explanatory layers. Therefore, the proposed architecture provides the balanced, robust, and deployable asset that can power the real-time digital screening apps and preventive tracking systems for combating the rising global prevalence of digital eye strain. Future work will include scaling this framework to support dynamic real-time streaming feature inputs from edge sensors, as well as exploring deep lightweight architectures to further reduce latency across ultra-low-power mobile devices.

## REFERENCES


[1] World Health Organization, "Blindness and vision impairment," World Health Organization. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment. Accessed: Jun. 30, 2026.

[2] World Health Organization, *World Report on Vision*. Geneva, Switzerland: World Health Organization, 2019. [Online]. Available: https://www.who.int/publications-detail-redirect/world-report-on-vision. Accessed: Jun. 30, 2026.

[3] Centers for Disease Control and Prevention, "About common eye disorders and diseases," Centers for Disease Control and Prevention. [Online]. Available: https://www.cdc.gov/vision-health/about-eye-disorders/index.html. Accessed: Jun. 30, 2026.

[4] Centers for Disease Control and Prevention, "Frequently asked questions about vision health," Centers for Disease Control and Prevention. [Online]. Available: https://www.cdc.gov/vision-health/faq-vision-health/index.html. Accessed: Jun. 30, 2026.

[5] National Eye Institute, "Eye conditions and diseases," National Eye Institute. [Online]. https://www.nei.nih.gov/eye-health-information/eye-conditions-and-diseases. Accessed: Jun. 30, 2026.

[6] National Eye Institute, "Diabetic retinopathy," National Eye Institute. [Online]. Available: https://www.nei.nih.gov/eye-health-information/eye-conditions-and-diseases/diabetic-retinopathy.

[7] M. A. Siddiq, "EyeSight & Vision Health Synthetic Dataset," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/mabubakrsiddiq/eyesight-and-vision-health-synthetic-dataset.

[8] F. Du, L. Zhao, H. Luo, Q. Xing, J. Wu, Y. Zhu, W. Xu, W. He, and J. Wu,(2024) "Recognition of eye diseases based on deep neural networks for transfer learning and improved D-S evidence theory," *BMC Medical Imaging*, 24(1),19 doi: https://doi.org/10.1186/s12880-023-01176-2

[9] M. ul Hassan, A. A. Al-Awady, N. Ahmed, M. Saeed, J. Alqahtani, A. M. M. Alahmari, and M. W. Javed, (2024)"A transfer learning enabled approach for ocular disease detection and classification," *Health Information Science and Systems*, 12(1) 36, doi: https://doi.org/10.1007/s13755-024-00293-8.

[10] L. Arora, S. K. Singh, S. Kumar, H. Gupta, W. Alhalabi, V. Arya, S. Bansal, K. T. Chui, and B. B. Gupta, (2024)"Ensemble deep learning and EfficientNet for accurate diagnosis of diabetic retinopathy," *Scientific Reports*, 14(1) 30554 doi: https://doi.org/10.1038/s41598-024-81132-4

[11] A. Aljohani and R. Y. Aburasain, (2024)"A hy brid framework for glaucoma detection through federated machine learning and deep learning models," *BMC Medical Informatics and Decision Making*, 24(1)115, doi: https://doi.org/10.1186/s12911-024-02518-y

[12] M. Khalid, M. Z. Sajid, A. Youssef, N. A. Khan, M. F. Hamid, and F. Abbas,(2024) "CAD-EYE: An automated system for multi-eye disease classification using feature fusion with deep learning models and fluorescence imaging for enhanced interpretability," *Diagnostics*, 14( 23) 2679
doi: https://doi.org/10.3390/diagnostics14232679

[13] S. Abbas, A. Qaisar, M. S. Farooq, M. Saleem, M. Ahmad, and M. A. Khan,(2024) "Smart Vision Transparency: Efficient ocular disease prediction model using explainable artificial intelligence," *Sensors*, 24(20),6618,
doi: https://doi.org/10.3390/s24206618

[14] T. D. Nguyen, D.-T. Le, J. Bum, S. Kim, S. J. Song, and H. Choo,(2024) "Retinal disease diagnosis using deep learning on ultra-wide-field fundus images," *Diagnostics*, 14(1)105
doi: https://doi.org/10.3390/diagnostics14010105

[15] M. Youldash, A. Rahman, M. Alsayed, A. Sebiany, J. Alzayat, N. Aljishi, G. Alshammari, and M. Alqahtani, (2024)"Early detection and classification of diabetic retinopathy: A deep learning approach," *AI*, 5(4)2586–2617,
doi: https://doi.org/10.3390/ai5040125

[16] S. Akhtar, S. Aftab, O. Ali, M. Ahmad, M. A. Khan, S. Abbas, and T. M. Ghazal, (2025)"A deep learning based model for diabetic retinopathy grading," *Scientific Reports*, 15(1)3763,
doi: https://doi.org/10.1038/s41598-025-87171-9

[17] E. Shah, J. Patel, V. Katheriya, and P. Pataliya, (2024)"Diagnosis of diabetic retinopathy using machine learning & deep learning technique," *arXiv*, arXiv:2411.16250, Available: https://arxiv.org/abs/2411.16250.

[18] A. Tashkandi, (2025)"Eye care: Predicting eye diseases using deep learning based on retinal images," *Computation*, 13(4) 91,
doi: https://doi.org/10.3390/computation13040091