IDENTIFYING RISK FACTORS FOR PREDICTION OF CATARACT USING MACHINE LEARNING MODELS

Gerald Ikechukwu Onwuka[1], Babayemi Afolabi Wasiu[1] and Buhari Shehu Abubakar[1,2*]

[1] Faculty of Physical Sciences, Department of Mathematics, Abdullahi Fodio University of Science and Technology, Aliero, Nigeria
[2] Department of Statistics, Federal Polytechnic, Kaura Namoda, Zamfara State, Nigeria

*Corresponding Author: buharishehukaura@gmail.com

ABSTRACT

Cataract is a growing global health issue affecting atleast 2.2 billion people worldwide and in Nigeria, Cataract accounted for approximately 43% of visual impairment cases and 7.2% of the Nigerian population over 40 years had Cataract in one or both eyes. Previous studies on Machine Learning (ML) identified significant Demographic Risk Factors (DRFs) and Clinical Risk Factors (CRFs) for prediction of Cataract but identifying significant lifestyle risk factors (LRFs) alongside with DRFs and CRFs for prediction of Cataract in the North-West zone of Nigeria remain a challenge. The data used in this study was from on-going research by the corresponding author on eye diseases and it was collected by the corresponding author through personal interview at NEC Kaduna from 25th Nov. to 19th Dec. 2024 after approval was granted by the ethical committee of the eye centre on 21st October, 2024. The study identified twelve out of thirteen risk factors of Cataract as SRFs and too much alcohol as non-SRF. Demographic Risk Factors (DRFs) are age and sex; Clinical Risk Factors (CRFs) were Presence of Diabetes (POD), Higher Body Mass Index (HBMI), Previous Eye Injury (PEI) and Family History of Cataract (FHC); Lifestyle Risk Factors (LRFs) were Too Much Sun Exposure (TMSE), Frequent Expose to Radiation (FER), Too Much Coffee Consumption (TMCC), Frequent Use of Social Devices (FUSD), Smoking Cigarette (SMC), Medication with Steroid (MWS) and Too Much Alcohol (TMA). Support Vector Machine (SVM) predicted 96.2%, 94.8% and 90.8%, 80.0%; Decision Tree (DT) predicted 98.1%, 98.2% and 96.7%, 90%; K-Nearest Neighbour (K-NN) predicted 97.1%, 94.5% and 93.3%, 80%; Naïve Bayes (NB) predicted 95.2%, 92.7% and 86.7%, 70%, Multilayer Perceptron (MLP) predicted 98.1%, 94.5% and 93.3%, 90% Cataract and non-Cataract patients in the training and test sets respectively. The evaluation of the models performance and 5-fold cross validation indicated that the models are adequate and DT was the best model because it has higher accuracy, sensitivity and specificity. The study demonstrated that identifying significant LRFs alongside with DRFs and CRFs could ensure effective prediction of Cataract in the zone and also helps stakeholders in public health to implement intervention strategies for Cataract prevention and management but there is need to address the issue of limited sample size in future work.

Key words: Cataract, Machine learning, Risk factors, Prediction, Training and Test sets

1. INTRODUCTION

Eye is one of the five sense organs that help us to witness the happenings within the surroundings. Unfortunately, eye suffered from diseases which could lead to blindness and visual impairment, one of these diseases is Cataract.

Cataract is an eye condition in which the lens of the eye becomes cloudy. This causes vision to worsen, making it especially difficult to see things clearly. Some people's vision is only slightly affected, whereas others might lose their eyesight very quickly. It is the leading cause of blindness and vision impairment in the world and World Health Organisation (WHO) estimated the count of individuals living with vision impairment in the world to be over 2.2 billion. With respect to the aforementioned statistic of vision impaired people, Cataract was identified as the leading cause accounting for over thirty three percent (33%) and over forty five percent (45%) as the initial cause for blindness due to late detection and improper treatment (Nur *et al.,* 2021; Zhang *et al.,* 2020). In Nigeria, Cataract accounted for approximately 43% of visual impairment cases and 7.2% of the Nigerian population over 40 years had Cataract in one or both eyes (NIH, 2023).

Over the past years, ophthalmologists diagnose patients for Cataract based on their experience and clinical training. But as a result of technological development, Artificial Intelligent (AI) was introduced in clinical practice to diagnose patient as either disease or non-disease. ML as a branch of AI had demonstrated significant potential in eye diseases research, especially in early diagnosis of patients using risk factors (Zhang *et al*., 2025). ML algorithms that use SRFs had shown considerable promise in predicting eye disease and non-eye disease patterns because the use of Non-SRFs could reduce the performance of the algorithms and resulted in poor prediction (Wu *et al.,* 2024). The complexity of Cataract development and progression necessitates innovative approaches for early detection and prediction. ML models offer promising solutions, but their effectiveness depends on identifying and integrating relevant risk factors.

Existing studies such as Jauro *et al.* (2024) implemented a hybrid ML approach based on CNN-SVM for detection of Glaucoma. The results for the proposed CNN-SVM offered an accuracy, precision, recall and F1-score of 100% each showing its superiority over the other existing techniques such as SVM which had accuracy, precision, recall and F1-score of 93%, 92%, 90% and 94% respectively. Zannah *et al.* (2024) used clinical risk factors and fundus images, presented a robust model Bayes SVM 500. The performance of the model was compared with seven (7) ML models namely LR, RF, DT, KNN, SVM, AdaBoost and XGBoost for prediction of Glaucoma, Cataract and Diabetic Retinopathy patients. The results obtained showed that the model performed well with 95.33% accuracy, 94.02% sensitivity, 93.18% specificity, 96.13% precision, 95.67% recall and 95.90% F-1 score. Moreover, three machine learning models [Multilayer Perceptron Neural Network (MLPNN), Decision Tree (DT) and Naïve Bayes (NB)] were developed by Egejuru *et al.,* (2017) using structured data, significant CRFs and DRFs. They found that the models achieved prediction accuracy of 100%, 87% and 84% and MLPNN had the best capability to identify the unseen pattern existing within the risk factors used to develop the predictive models for Cataract. Hassan *et al.* (2021) formulated CNN model to detect Cataract disease using fundus image dataset. The study found that the model achieved prediction accuracy of 99.5%, validation accuracy of 98.2%, sensitivity of 96.6% and specificity of 100%. The model was capable of detecting Cataract disease

The gap identified is that, none of the studies reviewed include LRFs in identifying significant SRFs for prediction of Cataract in the North-West zone of Nigeria which was addressed in this study.

2. MATERIALS AND METHODS

2.1 Instrument for Data Collection

The dataset used in this study was from on-going research which started in 2024 by the corresponding author on eye diseases. It was collected by the corresponding author through personal interview at NEC

Kaduna from 25th Nov. to 19th Dec. 2024 after approval was granted by the ethical committee of the eye centre on 21st October, 2024. This study used the dataset because patients across Nigeria visit the centre with eye problems for medication and due to insecurity in the zone, it would be the easiest way to get patients coming from the North- West zone.

The Cataract dataset had two (2) DRFs age and sex; four (4) CRFs namely FHC, HBMI, POD and PEI; seven (7) LRFs were TMSE, FER, FUSD, TMCC, SMC, MWS and TMA. The dataset contain a sample size of 200 and covered patients within the age of 10-80 years. To ensure representativeness of the sample, the study interviewed 35 patients from Kano and Kaduna states, 26 patients each from Katsina, Jigawa, Sokoto, Zamfara and Kebbi states which make up total sample size of 200 patients.

2.2 Data Preparation and Preprocessing

The main steps of data preprocessing and preparation performed in this study are divided into two main categories: data cleaning and balanced sampling. Data cleaning steps are outlier detection and removal and missing value handling. For outlier detection, numerical variables are analysed using the interquartile range and according to this method, outlier was detected in numerical variables. No missing value because it was face to face interview between the patients and two ophthalmologist, researcher and a staff from research and management information system at NEC Kaduna and all the patients interviewed gave complete and detailed information.

To handle outliers, the study used One-sided Winsorization technique by replacing only the upper or lower extreme values with a specified percentile (95th percentile). The steps for winsorization were:

i. Identify the instances/features that are farthest away from the median or mean
ii. Apply Winsorization to the identified outliers. That is, choose a percentile (95th percentile to replace the upper or lower extreme values. This percentile was called the Winsorization threshold
iii. Analyze the results to ensure that the outliers have been effectively handled. Calculate the value below which 95% of the data falls.

The dataset was imbalance because it distribution between the classes is not equal. That is, 135(67.5%) of the considered patients belong to Cataract (majority class) while 65(32.5%) belong to non- Cataract (minority class). Previous study by Krawczyk (2016) showed that the classifiers trained on imbalance dataset have higher accuracy for predicting the majority class and minority class could not be trained with higher accuracy. To address this problem of imbalanced datasets, the study used Synthetic Minority Over-sampling Technique (SMOTE) and the steps applied as follows:

i. Check for class imbalance, that is, verify the datasets are imbalanced by checking the class distribution and calculate ratio (majority class size or minority class size)
ii. Split the datasets into training and testing sets and apply SMOTE only to the training set to avoid overfitting
iii. Use SMOTE library (imbalance-learn in Python) to generate synthetic samples for the minority class. Specify the oversampling ratio but in this study the library automatically determine it
iv. SMOTE generates synthetic samples by interpolating between existing minority class instances and the synthetic samples are added to the training set
v. Train SVM, DT, K-NN, NB and MLP on the oversampled training set and evaluate the models performance on the testing set

vi. Use metrics like precision, recall, F1-score and AUROC curve to evaluate the models performance.

The SMOTE parameters used are K-Neighbour (five nearest neighbour when generating synthetic samples) and sampling strategy (’auto’, ‘minority’, ‘not minority’, ‘all’).

2.3 Data Normalization

Data normalization was performed because Cataract dataset have risk factors that differ in range and unit, this would reduce the models performance and accuracy. Secondly, prevent risk factors with larger scales from dominating the learning process. There are different types of data normalization, but this study used min-max to transforms risk factors of the datasets to a specified range, usually between zero (0) and one (1) because the dataset was non-Gaussian, maintains the interpretability of the original values within the specified range. The min-max scaling formula was given by

$$X_{normalized} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{1}$$

where X is a random risk factor value that is to be normalized, $X_{\min}$ is the minimum risk factor value in the dataset and $X_{\max}$ is the maximum risk factor value. When X is minimum value, the numerator is zero ($X_{\min}$ - $X_{\min}$) and the normalized value is 0. When X is maximum value, the numerator is equal to the denominator ($X_{\max}$ - $X_{\min}$) and the normalized value is 1 (Margaret, 2023).

2.4 Pearson Correlation Ranked Based Feature Selection

Feature selection method refers to the process of reducing the number of risk factors when using ML models. This study used feature selection method called Pearson correlation ranked based to select the significant risk factors for prediction of Cataract, because of two reasons namely the method focused on identifying the most important risk factors that are essential in building models for more accurate prediction (Bustamante-Arias *et al.,* 2021) and to know the type of relationship that exists between each risk factor and the output (target variable). Pearson correlation $\left(\rho_{X,Y}\right)$ is given by

$$\rho_{X,Y} = \frac{\operatorname{cov}\left(X,Y\right)}{\sigma_X \sigma_Y} \tag{2}$$

where $\operatorname{cov}\left(X,Y\right)$ is the covariance between X and Y and σ is the standard deviations (SDs) on X and Y. $\rho_{X,Y}$ value lies between -1 and +1, where -1 means a negative correlation between X and Y, 0 indicates no correlation between X and Y and +1 shows a positive correlation between X and Y. The closer the $\rho_{X,Y}$ value is to 1, the higher the correlation between X and Y. For the selection of SRFs, Vidhya (2024) and Mc Elduff *et al*. (2002) decision was implemented, that is, risk factors with P-value less than 0.05 ($P < 0.05$) were chosen as SRFs and those with P-value greater than 0.05 ($P > 0.05$) are not SRFs.

2.5 Random Forest for Feature Importance Scores

Random forest was used to identify the importance scores or contribution of risk factors to the predicted risk of the eye diseases.

Random Forest steps as follows:

i. Data preparation. That is, collect and pre-process the data including the risk factors (input) and the outcome variable (output). Split the data into training and testing sets
ii. Train a Random Forest Model on the training data to predict the risk of the eye diseases. Tune hyperparameters to optimize model performance.
iii. Calculate feature importance: Use Random forest model to calculate the feature importance scores for each risk factor. Feature importance scores represent the contribution of each risk factor to the model predictions.
iv. Interpret the feature importance score to determine the contribution of each risk factor to the predicted risk of the eye diseases. High score indicate a greater contribution.
v. Identify the risk factors with the highest feature importance scores, indicating the greatest contribution to the predicted risk of the eye diseases.

2.6 Evaluation of the Models Performance

The performance of the models was evaluated using measurement performance indices namely accuracy, sensitivity and specificity. Accuracy measure the proportion of cases correctly classified, sensitivity measure the fraction of positive cases that are classified as positive and specificity measure the fraction of negative cases that are classified as negative.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (3)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (4)$$

$$Specificity = \frac{TN}{TN + FP} \quad (5)$$

where TP is the true positive, TN is the true negative, FP is the false positive and FN is the false negative. Model with higher accuracy, sensitivity and specificity is the best model for prediction of Cataract (Fogarty & Bamber, 2005)

To ensure higher performance of the models in the light of limited sample (n = 200), the study make sure that relevant risk factors were selected using Pearson correlation ranked based and used min-max data normalization to prevent risk factors with larger scales from dominating the learning process. In order to avoid overfitting during training and improve generalization, the study used L1 regularization techniques called LASSO (Least Absolute Shrinkage and Selection Operator), hyperparameter tuning was carried out during training using Grid search because this search defines set of parameters values to search over and the algorithms tries all possible combination. Similarly, the study used model-centric approach because this approach searches for the optimal combination of hyperparameters within a predefined set of possible values.

2.7 Cross -Validation of the Models

This study employed cross-validation to validate the performance of the models used for prediction of Cataract eye disease. There are several types of cross-validation techniques such as k-fold cross-validation, leave-one-out cross-validation, holdout cross-validation, stratified cross-validation but this study used k–fold cross- validation because it maximizes the use of limited data, provides a more robust and reliable performance estimate and minimizes the risk of overfitting to a particular data split (Mc Elduff *et al.*, 2002).

To use k-fold cross-validation, the sampled data of Cataract was randomly partitioned into five equal sized sub-sample (i.e., k = 5). One sub-sample was used for testing and the remaining four (4) equal sub-samples for training. Then, Erickson & Kitamura (2021) decision was applied to interpret cross-validation result. That is, 70%-80% good cross-validation and above 80% perfect cross-validation.

3. RESULTS AND DISCUSSIONS

3.1 Identification of Significant Risk Factors of Cataract

Pearson correlation ranked-based feature selection method results in Table 1 revealed that out of thirteen (13) risk factors for Cataract twelve (12) were significant because their p-values was less than 0.05 (i.e. $P < 0.05$), one risk factor was not significant because it p-value was greater than 0.05 (i.e. $P > 0.05$) and also Figure 1(a) showed the plot of Pearson correlation coefficients for Cataract risk factors and Figure 1(b) illustrated risk factors importance scores. The SRFs were two (2) DRFs: age and sex, five (4) CRFs: namely FHC, HBMI, POD, PEI and five (6) LRFs namely TMSE, FER, FUSD, TMCC, MWS and SMC while TMA was the non-SRF.

Table 1: Significant Risk Factors of Cataract

Risk Factors	Correlation Values	P-Value	Risk factors Importance Scores in (%)
Age	0.2432	5.2136e-04	22.0
Sex	0.2179	1.9307e-03	8.0
Family history of Cataract	0.5866	7.0416e-20	20.0
High body mass index	0.7059	1.7678e-31	36.21
Presence of diabetes	0.6805	1.4928e-28	43.01
Previous eye injury	0.6417	1.3401e-24	32.0
Too much sun exposure	0.6836	6.8481e-29	62.97
Frequent exposure to radiation	0.6537	9.2429e-26	39.05
Frequent use of social devices	0.5599	6.7350e-18	22.0
Too much coffee consumption	0.5970	1.0595e-20	28.33
Smoking cigarette	0.2816	5.3418e-05	3.1
Medication with Steroid	0.2087	3.0209e-03	10.0
Too much alcohol consumption	-0.0256	7.1845e-01	1.0

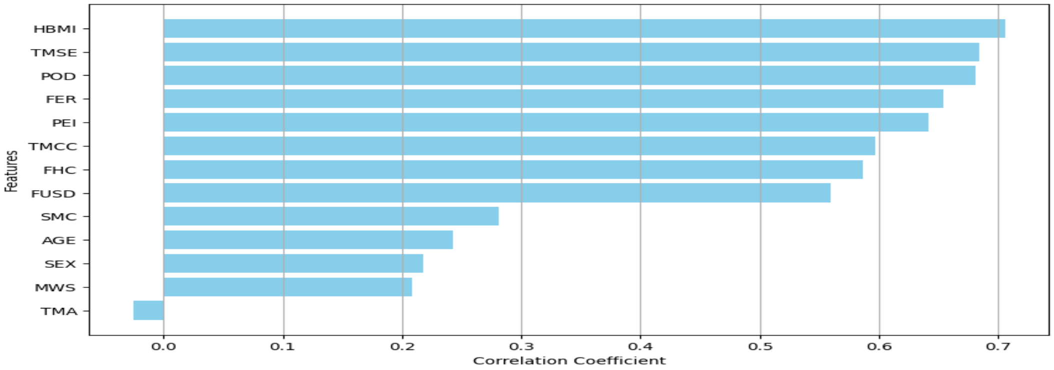

Figure 1(a): Plot of Pearson correlation coefficients for Cataract risk factors

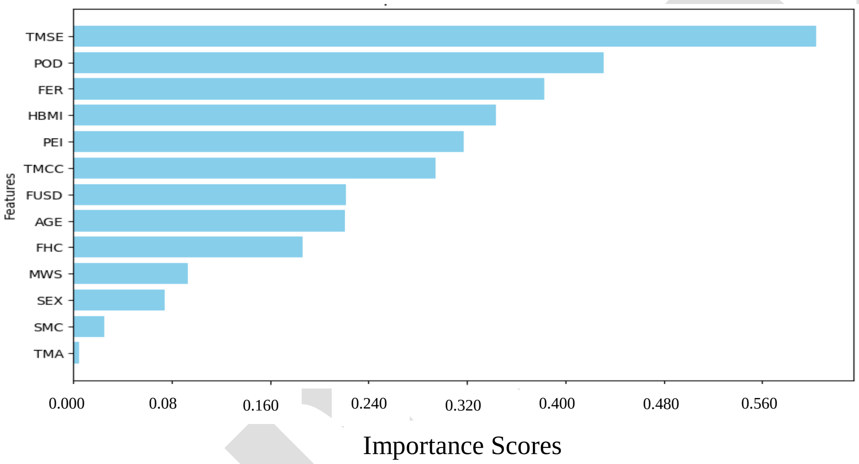

Figure 1(b): Plot of risk factors importance Scores

The Random forest approach showed that the importance scores of the thirteen (13) risk factors to the predicted risk of Cataract was presented in Table 1 and Figure1 (b). The importance scores of the SRF which could be interpreted as the contribution of the risk factors to the predicted risk of Cataract as follow: DRFs age and sex contributed 22% and 8%, CRFs namely POD, HBMI, PEI, FHC and MWS contributed 43.01%, 36.21%, 32%, 20%, and 10%; LRFs namely TMSE, FER, TMCC, FUSD, SMC and TMA contributed 62.97%, 39.05%, 28.33%, 22%, 3.10% and 1% to the predicted risk of Cataract respectively. TMA had the lowest importance score and was eliminated.

3.2 Training of the Models Using the Significant Risk Factors

Before training of the models using the SRFs, the dataset was splitted into training and test sets for model training and evaluation as done in the work of Alegre (2018) and Fiza *et al.* (2022). The study tried different data splits such as 60:40, 70:30 and found that the ratio 80:20 consistently provided the best result in terms of model stability and accuracy. The training set contained 80% (160) of the data which was used to train the model and the remaining 20% (40) was held out as the testing set to assess the model performance. The results of splitting Cataract dataset was presented in Table 2.

Table 2: Splitting of Cataract Dataset into Training and Test Sets

	TRAINING SET			TEST SET		
	CATARACT STATUS			CATARACT STATUS		
	Cataract	Non-Cataract	Total	Cataract	Non-Cataract	Total
Count	105	55	160	30	10	40
Percentage	65.6	34.4	100.0	75	25	100.0

3.3 Prediction of Cataract Status using the Trained Models

The trained ML models are used to predict Cataract and Non-Cataract patients in the training and test sets, the results of the five models are presented in Tables 3 to 7. Table 3 showed that the SVM model predicted 96.2% and 90.8% Cataract patients, 94.5% and 80.0% Non-Cataract patients in the training and test sets.

Table 3: Prediction of Cataract Status using SVM Model

Training set	Cataract	Non-Cataract	Predicted cases (%)
Cataract	101	4	96.2
Non-Cataract	3	52	94.5

Test set	Cataract	Non-Cataract	Predicted cases (%)
Cataract	27	3	90.8
Non-Cataract	2	8	80.0

Table 4 showed that DT model predicted 98.1% and 96.7% Cataract patients, 98.2% and 90% Non-Cataract patients in the training and test sets respectively.

Table 4: Prediction of Cataract Status using DT Model

Training set	Cataract	Non-Cataract	Predicted cases (%)
Cataract	103	2	98.1
Non-Cataract	1	54	98.2

Test set	Cataract	Non-Cataract	Predicted cases (%)
Cataract	29	1	96.7
Non-Cataract	1	9	90.0

Table 5 revealed that K-NN model predicted 97.1% and 93.3% Cataract patients, 94.5% and 80% Non-Cataract patients in the training and test sets.

Table 5: Prediction of Cataract Status using K-NN Model

Training set	Cataract	Non-Cataract	Predicted cases (%)
Cataract	102	3	97.1
Non-Cataract	3	52	94.5

Test set	Cataract	Non-Cataract	Predicted cases (%)
Cataract	28	2	93.3
Non-Cataract	2	8	80.0

Table 6 indicate that NB model predicted 95.2% and 86.7% Cataract patients in the training and test sets, 92.7% and 70% Non-Cataract patients in the training and test sets.

Table 6: Prediction of Cataract Status using NB Model

Training set	Cataract	Non-Cataract	Predicted cases (%)
Cataract	100	5	95.2
Non-Cataract	4	51	92.7

Test set	Cataract	Non-Cataract	Predicted cases (%)
Cataract	26	4	86.7
Non-Cataract	3	7	70.0

Table 7 showed that MLP model predicted 98.1% and 93.3% Cataract patients, 94.5% and 90% Non-Cataract patients in the training and test sets.

Table 7: Prediction of Cataract Status using MLP Model

Training set	Cataract	Non-Cataract	Predicted cases (%)
Cataract	103	2	98.1
Non-Cataract	3	52	94.5

Test set	Cataract	Non-Cataract	Predicted cases (%)
Cataract	28	2	93.3
Non-Cataract	1	9	90.0

3.4 Result of Model Evaluation

The result of models evaluation indicated that the SVM model achieved 98.5% accuracy, 94.8% sensitivity and 92.3% specificity. DT model had 99.5% accuracy, 97.8% sensitivity and 96.9% specificity. The accuracy, sensitivity and specificity of K-NN model were 97.8%, 96.3% and 92.3. NB model had 96.3% accuracy, 93.3% sensitivity and 89.2% specificity. MLP model achieved 99.2% accuracy, 97.0% sensitivity and 93.8% specificity..

3.5 Result of 5-fold Cross-Validation

The 5-fold cross validation results showed that the mean across the folds for SVM were 98.4% accuracy, 94.6% sensitivity and 92.1% specificity. DT had 99.4% accuracy, 97.7% sensitivity 96.8% specificity. K-NN had 97.6% accuracy, 96.2% sensitivity and 92.1% specificity. NB achieved 96.2% accuracy, 96.2% sensitivity and 89.1% specificity. MLP had 99.1% accuracy, 96.7% sensitivity and 93.6%

specificity respectively. Thus, DT was the best model for prediction of Cataract because it has higher accuracy, sensitivity and specificity

4. CONCLUSION

This study aimed at identifying risk factors for prediction of Cataract using ML models. It was found that twelve out of thirteen Cataract risk factors were significant. Five ML models are trained using the SRFs to predict patients' Cataract status and all performed well both in the training and test sets despite limited sample size. DT was the best model because it has higher accuracy, sensitivity and specificity both in the evaluation of models performance and 5-folds cross validation. The study demonstrated that identifying significant LRFs alongside with DRFs and CRFs could help for effective prediction of Cataract in the zone and also help stakeholders in health sector to implement appropriate intervention strategies for Cataract prevention and management at early stage.

LIMITATION OF THE STUDY

Small sample size n = 200 is the major limitation of the study. There is need to check the performance metrics using larger and independent datasets, this would justify the performance of the models using limited sample size (n = 200). Also, there is need to cross-validate the models using larger, independent datasets.

REFERENCES

Alegre, P. (2018). Development and Comparison of Machine Learning Methods for Subjective Refraction Prediction. Work Presented in Partial Fulfilment of the Requirement for the Bachelor Degree in Computer Science.

Bustamante – Arias, A., Cheddad, A., Jimenez – Prezez, J.C., & Rodriguez – Garcia, A. (2021). Digital Image Processing and Development of Machine Learning Models for the Discrimination of Corneal Pathology: An Experimental Model. *Phonics Journal*, 8(4): 118 – 124. https://doi. org / 10. 3390 / phonics 8040118.

Erickson, J.B., & Kitamura, F. (2021). Nine Performance Metrics for Machine Learning Models. *Journal of Radiol Artif intell*, 3(3). Retrieved from https://doi. Org/10.1148/ryai.2021200126.

Egejuru, N.C., Balogun, J.A., Mhambe, P.D., Sahiah, F.O., & Idowu, P.A., (2017). Model for Prediction of Cataract using Supervised Machine Learning Algorithms. Journal of *Computing Information Systems*, *Development Informatics and Allied Research*, 8(3), 47 – 62. Doi: 10.1136/jcisdiar – 2023 – 100815.

Fiza, S., Pratiksha, M., Pooja, B., Sidahali, N. (2022). Eye Diseases Detection using Machine Learning. International Journal of Advance Research in Science, Communication and Technology (IJARSCT), 2(1), 4659-4665. https:// doi.org/10.48175/IJARSCT-4659.

Fogarty, E., & Bamber, D. (2005). Area Above the Ordinal Dominance Graph and Area Below the Receiver Operating Characteristic Curve. *Journal of Maths Psychology*, 12, 387-415. https://doi. org/14.1512/matps.2005/12387-0015.

Hassan, K., Tanha, M.D., Amin, T., Faruk, R.M.D., Khan, O., Aljahdaili, M.M., & Masud, S. (2021). Cataract Diseases Detection by using Transfer Learning – Based Intelligent Methods. *Handawi*

Journal of Computational and Mathematical Methods in Medicine, 202(14), 121- 132. Doi: 10. 1155/2021/7666365.

Jauro, S.S., Ali, S.Y., & Ahmed, M.K. (2024). Glaucoma Detection using Hybrid Machine Learning Techniques. *Dutse Journal of Pure and Applied Science (DUJOPAS)*, 10(3), 2476-8316. https://dx.doi.org/10.4314dujopas v10i3c. ISSN (Print).

Krawczyk, B. (2016). Learning from Imbalanced Data: Open Challenges and Future Directions. *Journal of Progress in Artificial Intelligence*, 5(4), 221-32. https://doi.org /10.1310/jpai.54221-32

Margaret, R. (2023). Machine Learning. Retrieved from https://www. technopedia.com

Mc Elduff, P., Attia, J., Ewald, B., Cockburn, J., & Heller, R. (2002). Estimating the Contribution of Individual Risk Factors to Disease in a Person with More Than One Risk Factor. *Journal of Clinical Epidemiology*, 55(6), 588-592. https://doi.org/10.1016/SO895-4356(02) 00388-8.

National Institute of Health (2023). Eye Diseases. Retrieved from https://www.ncbi.nlm.nih.gov www.ncbi.nlm.gov

Nur, N., Cokrowibowo, S., & Konde, R. (2021). Cataract Detection in Retinal Fundus Image using Gray Level Co-Occurrance Matrix and K-Nearest Neighbour. *Advances in Engineering Research Journal*, 20(4): 362-371.

Vidhya, A. (2024). Feature Selection in Machine Learning. Retrieved from https://www.analytic vidhya.com

Wu, J.H., Moghimi, S., Nishida, T., Mahmoudinezad, G., Zangwill, L.M., & Weinreb, R.N. (2024). Detection and Agreement of Event-Based OCT and OCTA Analysis for Glaucoma Progression. *Journal of Front. Med*., 12, 2567-2584. https:// doi. org/10.3389/fmed.2025.1573329.

Zannah, T.B., Hil-Kafi, A., Sheak, A., Hasan, Z., Shuva, F., Bhuiyan, T., Rahman, T., Khan, R.T., Kaiser, M.S., & Whaiduzzaman, A. (2024). Bayesian Optimized Machine Learning Model for Automated Eye Disease Classification from Fundus Images. *Journal of Computation*, 12(9), 190-197. https:// doi.org/10.3390/computation 12090190.

Zhang, X., Lv, J., Zheng, H., & Sang, Y. (2020). Attention-Based Multi-Model Ensemble for Automatic Cataract Detection in B-Scan Eye Ultrasound Images. University of New South Wales.

Zhang, XQ., Hu, Y., Xiao, ZJ., Fang, JS., Higashita, R., & Liu, J. (2025). Machine Learning for Cataract Classification/ Grading on Ophthalmic Imaging Modalities: A Survey. *Journal of Machine Intelligence Research*, 19(3): 184-208. https://doi.org/10.1007/s11633-.022-.1329-0.