An Enhanced Model for the Prediction of Cataract Using Bagging Techniques

Akazue Maureen, Ovoh Oghenefego, Abel E. Edje,
Clement O. Ogeh
Department of Computer Science, Faculty of Science, Delta State University, Abraka

Hampo JohnPaul A.C.
Computer Science Department, Federal University of Technology, Owerri

Abstract:- A clouding of the human eye lens affecting vision is called a cataract; and this is the cause to most avoidable blindness in the world especially among the aged ones. Early treatment and surgery in any patience diagnosed of cataract prevents the wrong case of blindness and total vision impairment. A wrongly classified cataract is an issue that causes wrong treatment and waste of fund. These have become a problem in the medical field, even some opticians can't swiftly detect and/or classify cataract. This research offers a solution through an enhanced model for the prediction of cataract. Bagging techniques of the ensemble algorithm of machine learning was applied in the development of this model. Bagging ensembled with KNN as the base estimator, was trained with dataset from MRL open website; and compared with some algorithms such as KNN, Navie Baye and Decision Tree. Bagging ensembled had the best accuracy as 82.66% for training set. The validation set and testing set has 82.78% and 82.88% accuracies respectively when bagging ensemble was used.

Keywords:- *Machine Learning, Ensembled Model, Human Cataract And Cataract Classification, Cataract Prediction, Bagging Technique.*

I. INTRODUCTION

Application of technology in every facet of human life correlates to technological advances in recent times; thereby, solving problems and improving lives, and simplifying the task that were mountainous in nature at a faster rate of work done and accuracy of work. The eye which is the channel for vision and source of light entrance to the body, is very essential and Hadeer et al., (2018) stated that it is considered to be the most important and sensitive of human's organ.

The leading cause of blindness and vision impairment in the world is cataract and World Health Organisation estimated the count of individuals living with vision impairment in the world to be over 2.2 billion (Nur et al., 2021; Zhang et al., 2022a; Zhang et al., 2020). 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.

Cataract is caused by the clumping up of protein behind the lens thereby clouding small area of the lenses with white colour, which in turn causes troubles in seeing. The eye lens is synonymous to the lens of a camera (Hadeer et al., 2018). The human eye lens is one of the complicated components in the eye and it regulates the focus of the eyes so that farther images can be seen clearer.

Blindness which is a socio-economical global problem reduces the productivity, capability and usefulness of the blind, leaving a negative impact on the family, community and at large the nation (Egejuru et al., 2017). Avoidance rate of blindness is between seventy five percent to eighty percent (75% - 80%) due to the fact that most blindness is caused by cataracts.

In Africa and Nigeria to be precise, the rate of blindness is relatively high. A study conducted Egejuru et al., (2017) in Nigeria indicated the estimated rate of blind individuals is one million, one hundred and thirty (1,130,000) people; of which majority are above 40 years

.

Lenses are a product of protein and water, and they are transparent in nature. Loss of crystalline lens transparency which is caused by the clump of protein together inside the lens leads to cataract (Zhang et al., 2020). Associated factors that likely results to cataract are drugs-induced change, developmental abnormalities, age, trauma, genealogy notably diabetics, behavioural pattern such as smoking and so on.

Cataract classification or grouping is either by the cause of the cataract or the position of the crystalline lens opacity. Classification by the cause of cataract yields age-related cataract (AC), pediatrics cataract (PC) and secondary cataract (SC). In classification by the lens' location, we have nuclear cataract (NC), cortical cataract (CC), and posterior subcapsular cataract (PSC) (Zhang et al. 2022a). The formation of white wedged-shaped, and radially oriented opacities, which occurrence takes place from the outside edge of the lens in the direction of the centre in a spoke-like fashion results to cortical cataract (CC). Posterior subcapsular cataract (PSC) is granular opacities and its occurrences begins as a trivial breadcrumbs or sand particles scattered beneath the lens capsule.

Cataracts though begins as a small cloudy area in the eye lens (Nur et al., 2021), it grows leading to blindness when the small cloudy area covers the eye lens in totality. The early stage of the cataract is when the cloudy area is relatively small. The advance stage is when the cloudy area is almost or fully covering the eye lens. Cataract left till the advance stage results to the individual having a high rate of inability to reverse the vision impairment; hence, cataract should be treated at the early stage. Hadeer et al (2018) stated that the cataract is treated only by surgery and that early treatment is preferred thereby decreasing the surgery risk and saving the patient from total blindness.

Machine learning is the implicit programming of systems and models. That is, machine learning is when a system learns from a given task and improves its performance with respect to its experience in the task. (Hurwitz & Kirsch, 2018; Mitchell, 1997). Machine learning; a branch of artificial intelligence, is a scientific discipline that necessitates the design and development of algorithms that permit computers to evolve behaviours based on empirical data (Thaseen & Kumar, 2017; Gamage & Samarabandu, 2020). The main objective of machine learning is to guesstimate the unknown relationship between input and target parameters using known examples. Then, the relationship thus derived can be used in forecasting the unknown target values for other values of inputs. The targets can be nominal or numerical. A problem is termed classification if the targets are nominal, else if the target is numerical, the problem is a regression one. The learning task that involves solving such a regression problem is called supervised learning.

Ophthalmologists have used several ophthalmic images to manually diagnose cataracts; this is based on their experience and clinical training. A digital system will greatly aid the ophthalmologist in diagnosis cataract faster and more accurate. An ensembled machine learning model was developed for cataract prediction.

Detecting or predicting cataract manually entails a wealth of experience and expertise by an ophthalmologist (Nur et al., 2021) which is lacking in rural areas. Experience and cataract grading system is required for diagnosis of cataract patients. Ophthalmologists uses ophthalmic images for manual cataract diagnosis; however, the manual diagnosis is prone to error, consumes time, subjective and costly. The manual diagnosis is a gargantuan challenge to nations that are developing especially to the rural communities (Zhang et al., 2020). Over the years, machine learning algorithm had been used in the development of automated cataract detection system, however some are lesser in accuracy as well as other metrics or in memory management. The need to improve the existing cataract detection systems by Nur et al. (2021), with a better performance metrics (accuracy, recall and precision), which is time efficient and memory efficient, necessitate this research on an enhanced model for the prediction of cataract. Therefore this study aims to develop an enhanced model for the prediction of cataract.

II. METHODOLOGY AND SYSTEM ANALYSIS

➢ *Methodology of the Study:*

The proposed model for the prediction of cataract takes cognisance of different methods that had been employed by existing literatures in the development of cataract predicting model.

MRL opensource website serves as the source of data acquisition for the proposed model. The chosen dataset from MRL website was downloaded into the computer’s memory. The downloaded dataset was read into Jupyter notebook, therein the dataset was cleaned and transformed.

Matplotlib and seaborn was used for visualizing the trends in the engineered dataset. The dataset was split vertically into input and target variables, and later a horizontal split of both input and target variables was done for the training, validation and testing set respectively.

In a nutshell, methodology of the study consists: review of existing documents with respect to human cataract prediction, the existing system, development of the proposed model, testing of the developed model and deployment of the developed model.

➢ *Analysis of the Existing System:*

The model by Nur et al., (2021) is the existing model adopted by this research. In their work, k Nearest Neighbour was used for cataract detection using retinal fundus images. They achieved an accuracy of 80%. The model in existence is a single and stand-alone classifier which consumes processing time and computational resources due to the value of k. Also, as a single model, the accuracy of prediction is of less confidence when correlated with an ensembled model or an hybridized model.

➢ *Analysis of the Proposed System:*

Scikit-learn (also known as sklearn) was used for the development of the proposed model. Sklearn is an open-source python library used for machine learning (supervised learning and unsupervised learning). Algorithms such as Support Vector Machine (SVM), Logistic Regression, Linear Regression, Naïve Bayes, Decision Tress and others are available in scikit-learn. Other libraries that was employed in the development of the proposed model are NumPy which is used for numerical computing and Pandas which is built on NumPy but with extended features and abilities for scientific computing. NumPy and Pandas was used for dataset importation in the working (model development) environment, cleaning of the dataset and also for pre-processing of the dataset. Exploratory Data Analysis (EDA) was carried out using Pandas in Jupyter Notebook which is the working (model development) environment. The dataset during the development of the model was split into train set, test set and validation set using 60%, 20% and 20% respectively. After the split, the dataset was vectorized and fed into the models namely Naïve Baye, KNN, Decision Tree and Bagging ensemble (with KNN as the base estimator) which learns the patterns within the dataset and a trained (learned) model was dumped after the training.

- *Advantages of the Proposed System:*

The advantages of the proposed system are listed below:

- ✓ Prediction – makes better prediction than a single contributing model.
- ✓ Perform better than complex model such as deep learning.
- ✓ The spread and dispersion in prediction is reduced.

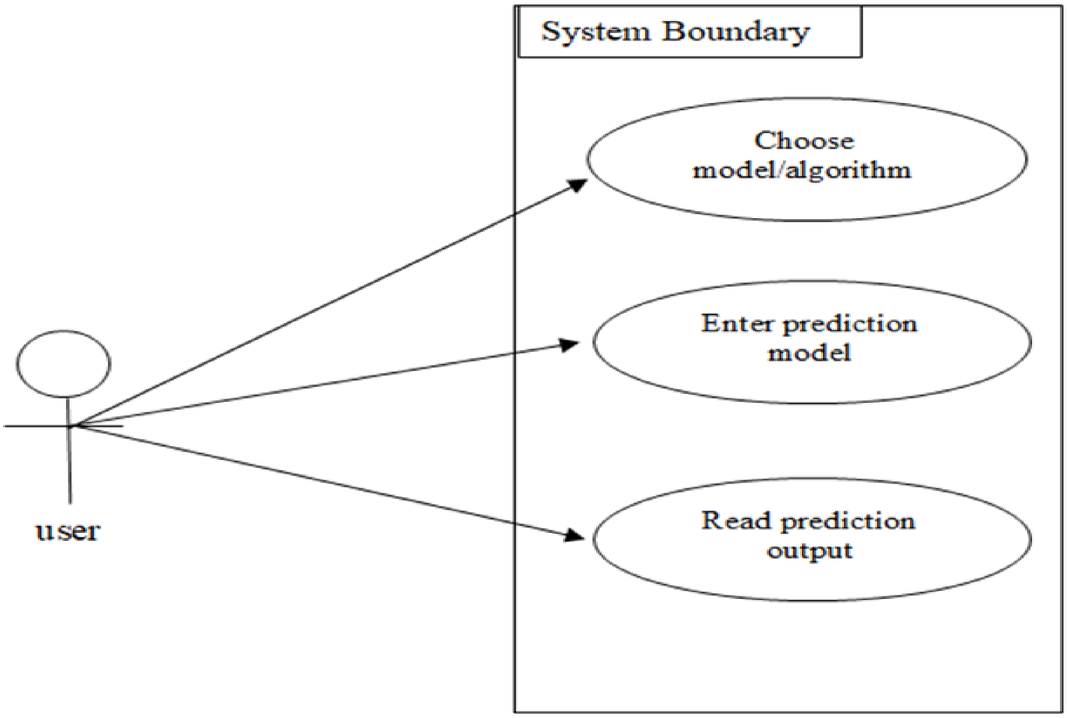


Fig 1 Use Case Diagram of the Model.

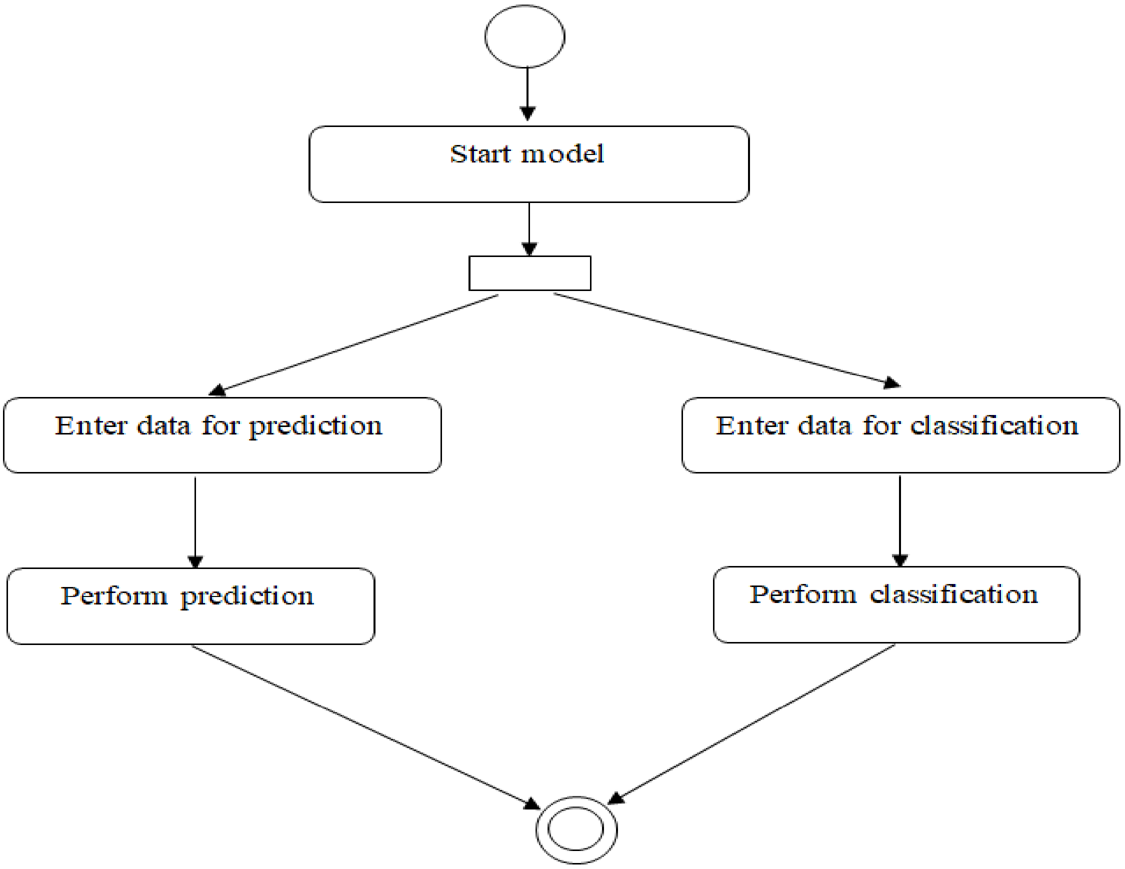


Fig 2 Activity Diagram of Model.

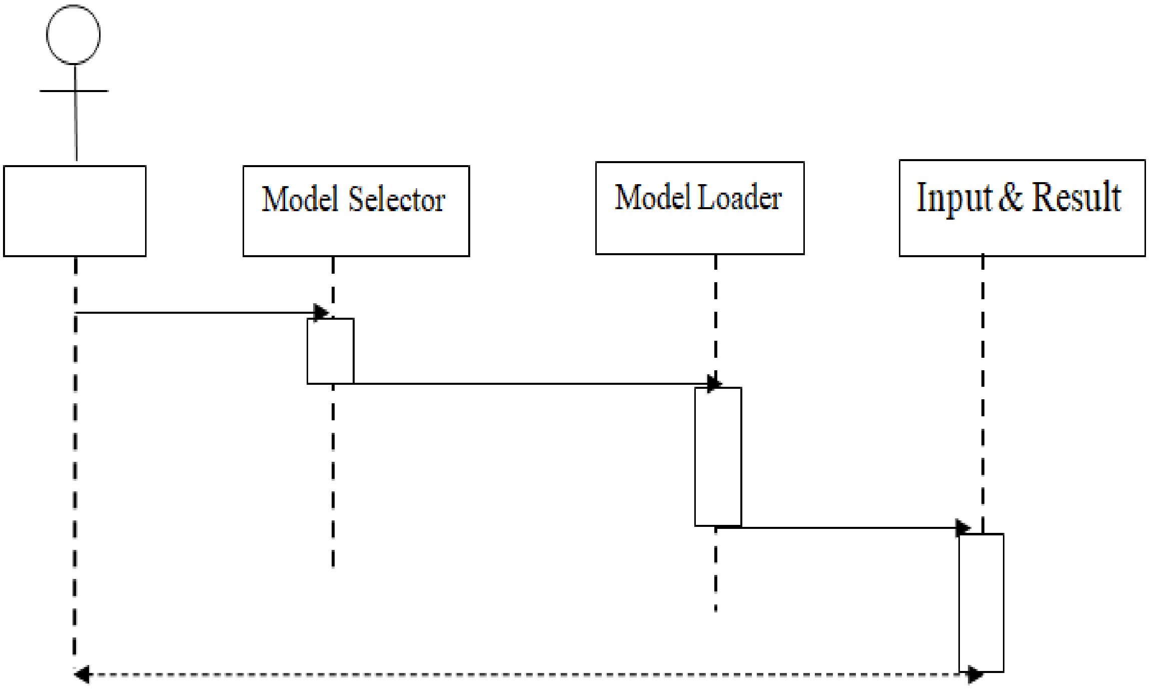


Fig 3 Sequence Diagram of the Model.

- *Model Architecture:*

A bagging ensembled machine learning model is presented in this study for the prediction of cataract. The dataset used for this research is the MRL Eye Dataset gotten from MRL website (http://mrl.cs.vsb. cz/data/ eyedataset /pupil.txt).

It is a publicly available dataset. The annotation of the dataset was used with the inclusion of the visual acuity based on the condition of the glass state of the eye and the reflection of the eye. The detail of the annotation is depicted in Table 1.

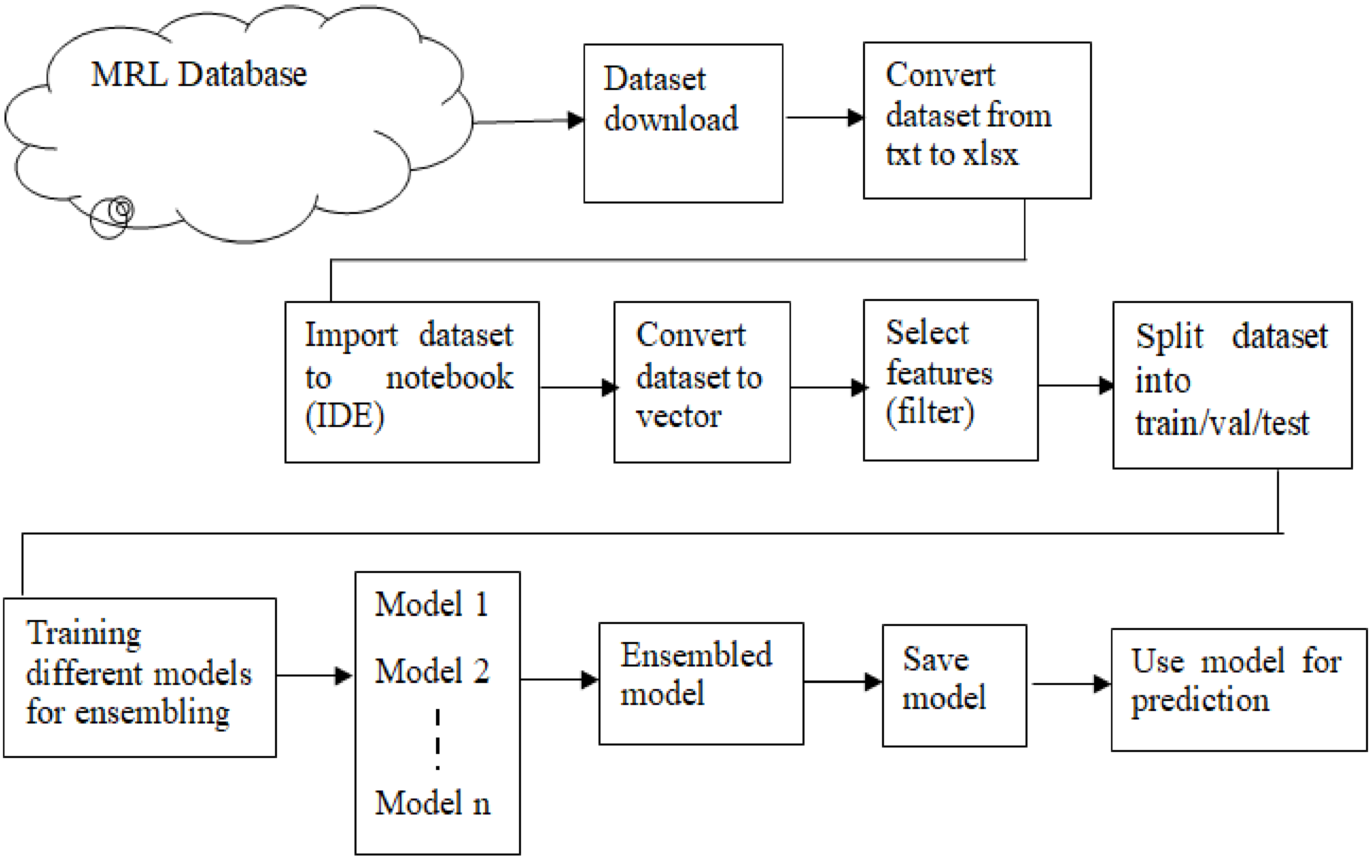

MRL Database
Dataset download
Convert dataset from txt to xlsx
Import dataset to notebook (IDE)
Convert dataset to vector
Select features (filter)
Split dataset into train/val/test
Training different models for ensembling
Model 1
Model 2
Model n
Ensembled model
Save model
Use model for prediction

Fig 4 High-level Architecture of the Proposed Model

Table 1 Modified MRL Dataset Description

Index	Title	Datatype	Description
0	subject_ID	object	in the dataset, we collected the data of 37 different persons (33 men and 4 women)
1	image_ID	int64	the dataset consists of 84,898 images
2	gender	int64	the dataset contains the information about gender for each image (0 for man, 1 for woman)
3	Age	int64	age of the patients
4	glasses	int64	[0 - no, 1 - yes]; the information if the eye image contains glasses is also provided for each image (with and without the glasses)
5	eye_state	int64	[0 - closed, 1 - open]; this property contains the information about two eye states (open, close)
6	reflections	int64	[0 - none, 1 - small, 2 - big]; we annotated three reflection states based on the size of reflections (none, small, and big reflections)
7	lighting_conditions	int64	[0 - bad, 1 - good]; each image has two states (bad, good) based on the amount of light during capturing the videos
8	sensor_ID	int64	[01 - RealSense, 02 - IDS, 03 - Aptina]; at this moment, the dataset contains the images captured by three different sensors (Intel RealSense RS 300 sensor with 640 x 480 resolution, IDS Imaging sensor with 1280 x 1024 resolution, and Aptina sensor with 752 x 480 resolution)
9	image	object	The file name of the annotated image
10	visual_acuity	object	The visual acuity based on the image reflection and state of glasses

To enable the model to learn without overfitting or under fitting and neither the model imitating, the MRL dataset was engineered on by increasing the features and also the observations by simulation.

Exploratory data analysis was performed to discover some trends in the dataset using visualizations. It was discovered as shown in Figure 5, Figure 6 and Table 2 that the dataset is relatively balanced especially in the gender distribution. Hence it has to be resampled either using overfitting or under fitting techniques and algorithm.

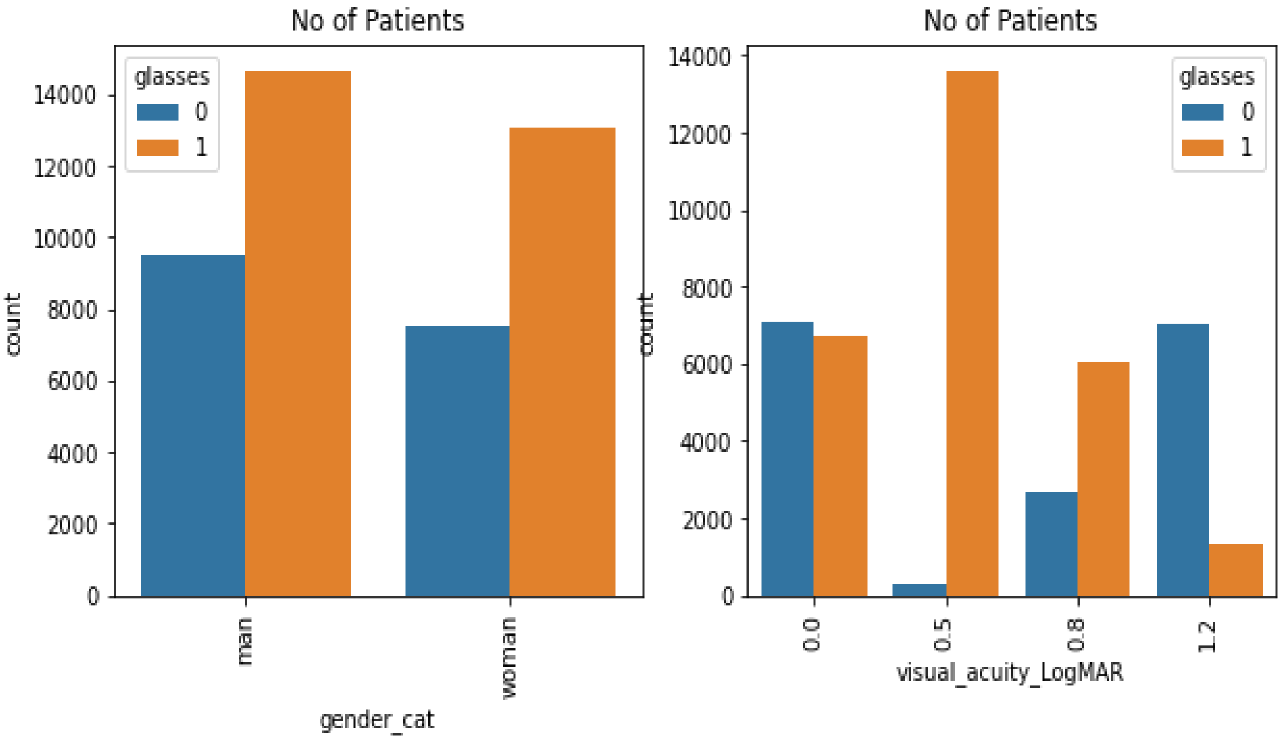


Fig 5 Exploratory Data Analysis

Table 2 Gender Value Count

Value	Count	Count (Normalized)
Man	24179	0.540156
Woman	20584	0. 459844

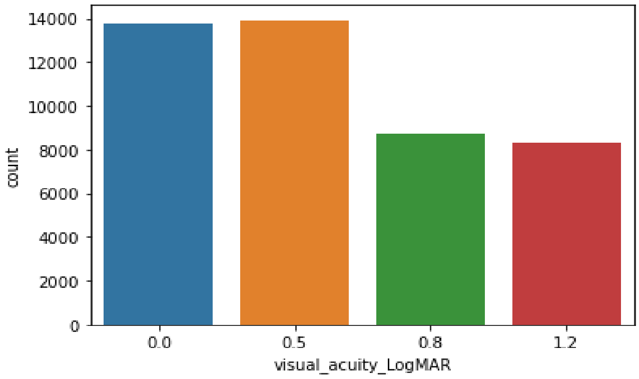


Fig 6 Class Balance Visualization

➢ *Features Selection:*

The filter method of supervised feature selection was implemented in this research. In this method, features are dropped based on their relationship with the output or target variable. That is, if any feature is correlated with the target, the feature is dropped and features not correlated with the target are retained.

III. SYSTEM DESIGN AND IMPLEMENTATION

➢ *Algorithm:*

Bagging ensemble also known as bootstrap aggregation forms a build several instances of a black-box estimator on random subsets of the original training set and then aggregate their individual predictions to form a final prediction. The base estimator used in this research is the k Nearest Neighbour algorithm. The general bagging algorithm is given below:

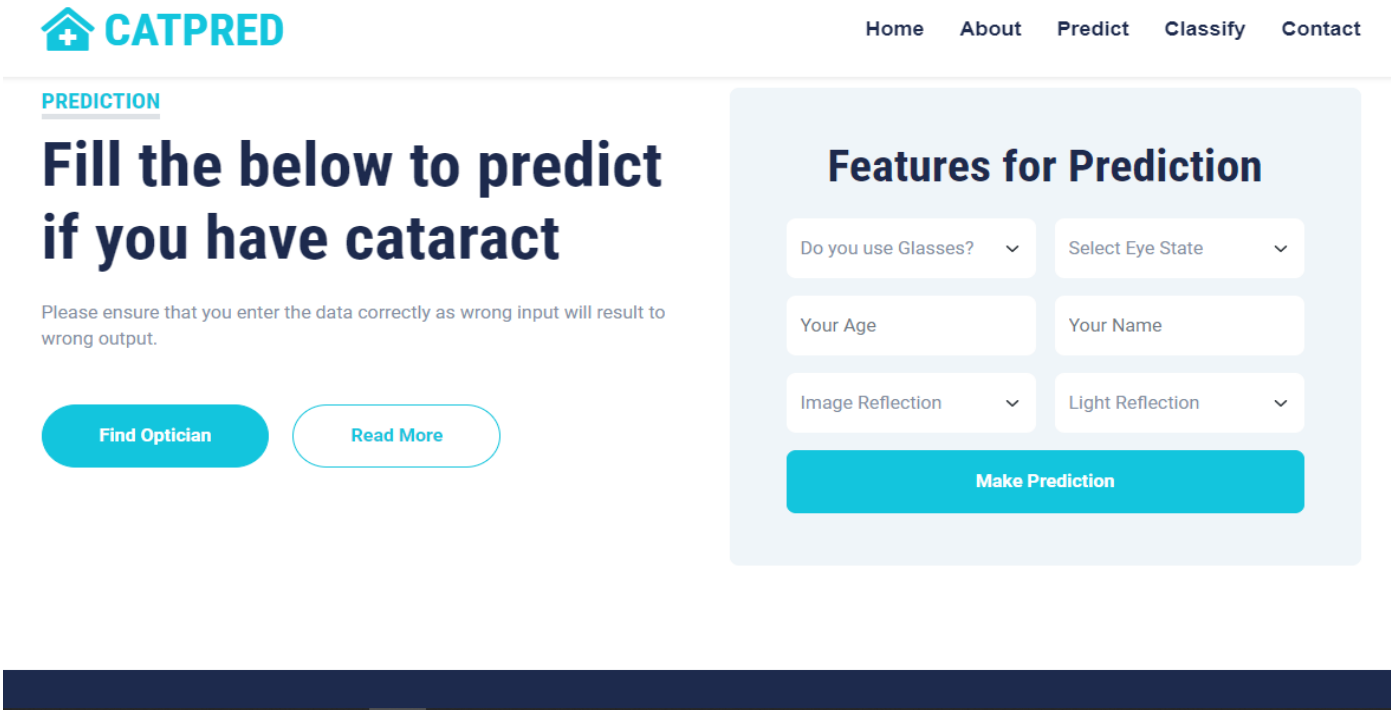
- Step 1. Generate bootstrap sample of same size repeatedly from train dataset with each record having same probability of selected
- Step 2. Train a model on each sample size in step 1 and record the prediction for each sample size
- Step 3. If model = classification then select the class with the most vote in step 2, else if model = regression then calculate the average of the step 2

➢ *The Existing System's Algorithm is:*

- Step 1. Select the number K of the neighbors
- Step 2. Calculate the Euclidean distance of K number of neighbors
- Step 3. Take the K nearest neighbors as per the calculated Euclidean distance.
- Step 4. Among these k neighbors, count the number of the data points in each category.
- Step 5. Assign the new data points to that category for which the number of the neighbor is maximum.
- Step 6. Our model is ready.

➢ *The Enhanced Algorithm for this Model is Below:*

- Step 1. Generate bootstrap repeatedly from train dataset
- Step 2. Initialize the estimator
- Step 3. Set the base estimator as KNN
- Step 4. Set number of estimators as 100
- Step 5. Set maximum number of samples as 0.8
- Step 6. Set out of bag score as true
- Step 7. Enable random state
- Step 8. Fit sample into the estimator
- Step 9. Find the most voted class

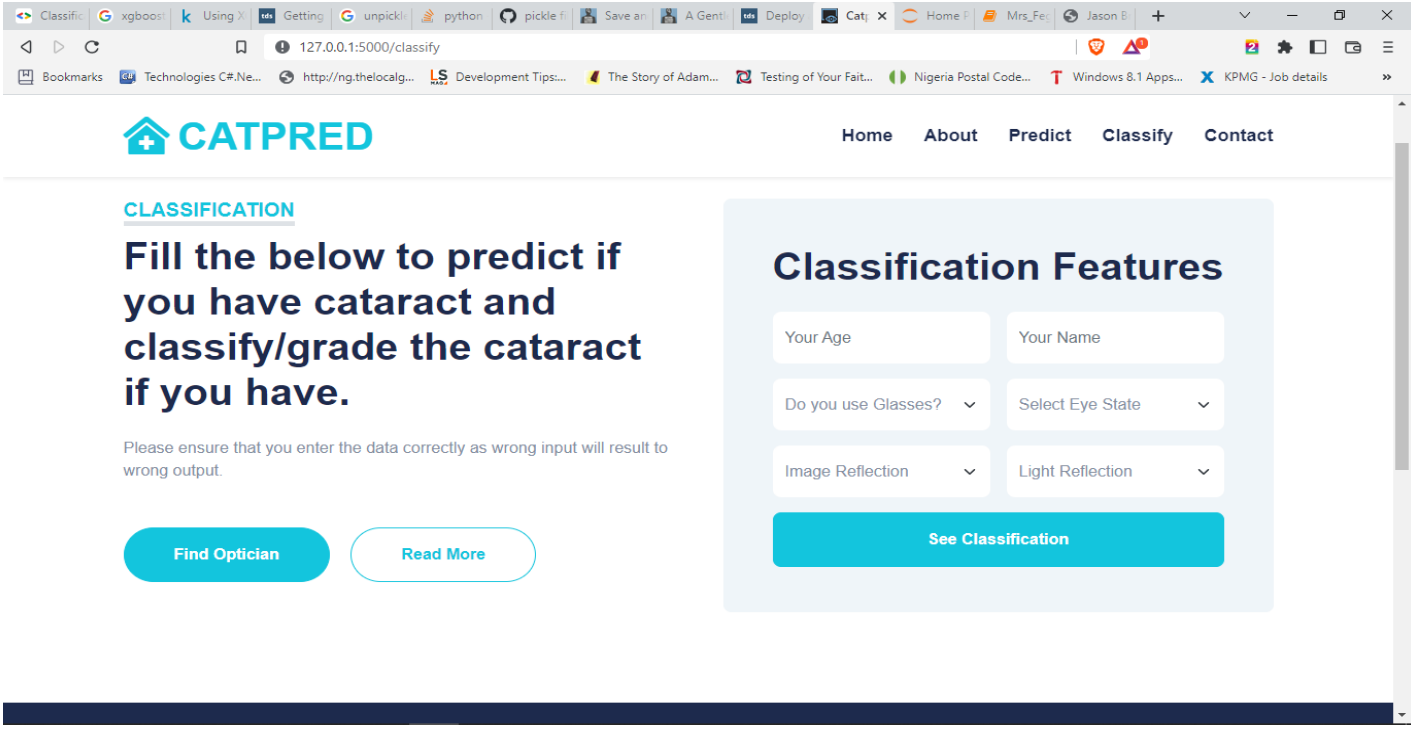


Fill the below to predict if you have cataract

Please ensure that you enter the data correctly as wrong input will result to wrong output.

Features for Prediction

Fig 7 Input Interface of the Deployed Model (Prediction)



CLASSIFICATION

Fill the below to predict if you have cataract and classify/grade the cataract if you have.

Please ensure that you enter the data correctly as wrong input will result to wrong output.

Classification Features

Your Age

Your Name

Do you use Glasses?

Select Eye State

Image Reflection

Light Reflection

Fig 8 Input Interface of the Deployed Model (Classification)

- *System Implementation:*

The implementation of the system was enabled using flask framework. Flask is a light web framework for the development of web applications using python.

- *System Testing:*

The trained model was first tested using the validation set and this set was used for hyper parameter tuning till a satisfied performance level was obtained. Then, the testing set was used on the trained and validated model to see its performance. The model did not suffer from overfitting nor underfitting, hence it was deployed for testing with live data.

The trained model which is an ensembled model was pickled after it was validated and tested with the validation set and the testing set. The pickled model was deployed using Flask framework in a Visual Studio Code environment.

The model was tested with 40% of the dataset which was randomly chosen through the train-test-split algorithm in scikit learn. The 40% was further split into validation set and testing set New and live data that was transformed into a two-dimension array (a data frame) will also be used for testing the model.

- *Performance Evaluation:*

The models were evaluated using their accuracies. Bagging ensemble techniques and the random forest algorithm did well in training above all other algorithms. However, their performance in validation and testing was low. Hence, Bagging ensembled which had an accuracy of 82.66% during training did well in validation and testing than all other model with accuracies of 82.78% and 82.88% respectively. The performance of the models in their accuracies is shown in table 3 while figure 9 visualized the accuracies of the models.

- *Accuracy* – the accuracy of the classification is given as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} * 100$$

Where TP is True Positive (number of correct classified class as positive), TN is True Negative (number of correct classified class as negative), FP is False Positive (number of wrong classified class as positive) and FN is False Negative (number of wrong classified class as negative).
The accuracy of the testing set gave 82.88%.

Table 3 Models Performance Evaluation

	Naive Bayes	Decision Tree	KNN	Bagging Ensemble
Training	79.70	80.04	80.93	82.66
Validation	81.60	81.60	81.44	82.78
Testing	80.59	80.59	81.31	82.88

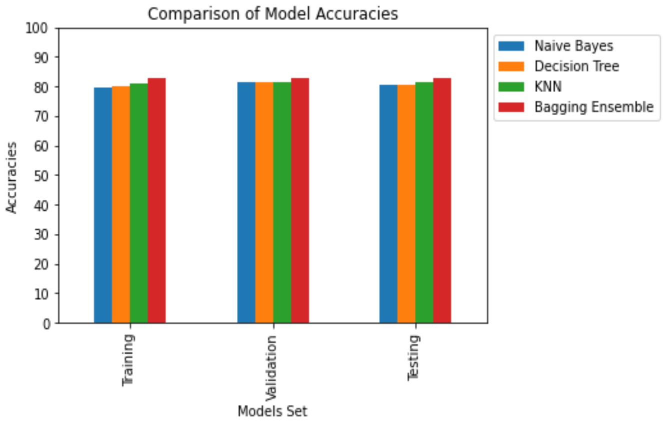


Fig 9 Performance Evaluation Bar chart

➢ *Limitations of the System:*

This model does not read image data and also cannot predict other eye-related issues. Also, there tends to be a loss in the interpretability of the model.

IV. CONCLUSION

This study has shown that human cataract prediction can be improved from the existing system that had an accuracy of 80% to this developed model which is based on the bagging techniques of ensemble learning using KNN as the base estimator. The accuracy of the developed model (CATPRED) is 82.88%.

❖ *Suggestions for Further Studies*

➢ *The suggestions for further studies are as follows:*

- Further research can be carried out using the hybridization of ensemble learning and deep learning.
- Further research could seek to deploy the model into a mobile device.

REFERENCES

[1]. Egejuru, N.C., Balogun, J.A., Mhambe, P.D., Asahiah, F.O. and Idowu, P.A (2017). Model for Prediction of Cataracts Using Supervised Machine Learning Algorithms. Computing, Information Systems, Development Informatics & Allied Research Journal Vol. 8 No. 3

[2]. Gamage, S. and Samarabandu, J. (2020). Deep learning methods in network intrusion detection: A survey and an objective comparison. Journal of Network and Computer Applications. Volume 169, 1

[3]. Hadeer R.M.T, Rania A. K.B., and Amani A.S. (2018). Early Recognition and Grading of Cataract Using a Combined Log Gabor/Discrete Wavelet Transform with ANN and SVM. World Academy of Science, Engineering and Technology International Journal of Computer and Information Engineering Vol:12, No:12

[4]. Hurwitz, J. and Kirsch, D. (2018). Machine Learning For Dummies®, IBM Limited Edition. John Wiley & Sons, Inc

[5]. Nur, N., Cokrowibowo, S. and Konde, R. (2021). Cataract Detection in Retinal Fundus Image Using Gray Level Co-occurrence Matrix and K-Nearest Neighbor. Advances in Engineering Research

[6]. Thaseen, I.S. and Kumar, C.A. (2017). Intrusion detection model using fusion of chi-square feature selection and multi class SVM. Journal of King Saud University - Computer and Information Sciences. Volume 29, Issue 4, Pages 462-472,

[7]. Tom M. Mitchell (1997). Machine Learning. McGraw-Hill Science/Engineering/Math

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

[9]. Zhang, X., Xiao, Z., Fu, H., Hu, Y., Yuan, J., Xu, Y., Higashita, R., and Liu, J. (2022a). Attention to region: Region-based integration-and-recalibration networks for nuclear cataract classification using AS-OCT images. Medical Image Analysis. s 80, 102499