International Journal of Engineering and Science Invention (IJESI)
ISSN (Online): 2319-6734, ISSN (Print): 2319-6726
www.ijesi.org ||Volume 13 Issue 9 September 2024 || PP 127-136

Designing Predictive Model Using DLMNN and Optimization Algorithm

HR Ravikumar[1], Prasadu Peddi[2]
[1]*Research Scholar, Dep of CSE, Shri Jagdishprasad Jhabarmal Tibrewala University Jhunjhunu, Rajasthan, India*
[2]*Professor, Dep of CSE & IT, Shri Jagdishprasad Jhabarmal Tibrewala University Jhunjhunu, Rajasthan, India*

ABSTRACT
This research employs a traditional machine learning and deep learning methodology, breaking it down into several phases. The first phase involves dividing the dataset into testing and training subsets. Oversampling is performed on the training dataset to balance data for both classes using SMOTE analysis. Min-max scaling is used to standardize the range of features, ensuring uniform distribution of attributes. A chi-square feature selection technique is used to find the best features, optimizing the model's performance and reducing computational complexity. The training data is split into a train and validation set, and supervised machine learning algorithms are applied to the preprocessed dataset. Each algorithm is chosen based on its specific characteristics and potential to model relationships between diabetes risk factors. Performance is evaluated on metrics such as accuracy, precision, recall, and F1-score. To improve model performance, hyper-tuning is performed on the validation dataset using Research's best hyperparameters. Manual hyper-tuning is performed by increasing and decreasing parameter values according to the hyperparameters. Evaluation metrics such as accuracy, precession, recall, f1score, and area under the receiver operating characteristic (ROC) curve are used.
KEYWORDS: *Diabetic disease, Deep learning, Machine Learning and ANN.*

I. INTRODUCTION

This paper presents two methods for grading diabetic retinopathy: SURF and spatial local binary pattern methods. The SLBP method generates spatial descriptors, while the SURF technique yields local descriptors. The optimized feature set is evaluated using various classifiers, such as k-Nearest Neighbours, Extreme Learning Machines, and Artificial Neural Networks. A comparative study is conducted to assess the three classifiers using various performance measures.

The second model employs a cutting-edge micro-macro textural feature extraction approach to enhance the diabetic retinopathy grading system. The optimal random forest classifier achieved 0.70 accuracy, 0.79 recall, 0.78 precision, and 0.78 F1-score in the diabetes-free patient group. Diagnostic evaluation results suggested potential pre-diabetes with an F1-score of 0.59, recall of 0.60, and accuracy of 0.57. The diabetes model's recall rate is 0.63, overall accuracy is 0.71, and F1-score is 0.67. Deep learning, also known as Artificial Neural Networks (ANNs) and multilayer perceptrons (ML), enables the automated creation of models from massive datasets. Neural networks, also known as topologies, are the foundation of deep learning and are structured directed visual representations constructed from neurones. The main emphasis of this thesis is to classify and evaluate diabetic retinopathy photographs using convolutional neural networks (CNN). Optimizing loss functions, which are parametric and include learnable parameters, is crucial for improving network performance in regression and modeling tasks. Batch normalization and regularizers are crucial tools for improving stability, decreasing covariate shift, and successfully standardizing inputs. ReLU networks often surpass linear models in performance due to their distinct zero-level non-differentiability and linearity. Machine learning models, such as neural networks and supervised learning, play a crucial role in predicting real-world situations. To ensure accurate forecasts, it is essential to employ models that minimize both variance and bias. Techniques like bagging, boosting, and model stacking can help minimize variance and bias, while data-driven model construction is essential in machine learning system design. In this study, a Deep Convolutional Neural Network (DCNN) is used to automate the process of identifying diabetic retinopathy by analyzing fundus photographs. DCNNs eliminate the need for feature analysis and picture preparation, functioning as an automated classification model with supervised learning. To avoid overfitting, large datasets are required, and data augmentation approaches can improve data gathering from limited datasets. The log-loss function is an important tool for optimizing multi-class classification issues, and the Kaiming and He initialization procedure is used for each network. The efficiency of the classifier is checked to see if adding co-occurrence data from both eyes improves its performance. To minimize logarithmic likelihood, the rules of Maximum Likelihood Estimation (MLE) are used, and the Kappa index (κ) is

used to evaluate the degree of agreement between raters on a population's diagnosis. Optimizing the κ index directly serves as both a loss function and an assessment metric during the model training process. To optimize the loss function associated with output variables, gradient descent techniques can be used for improvement. The standard approach for multi-class classification problems is the optimization of logarithmic loss, which has numerical stability, well-defined derivatives, and empirical efficacy.

II. METHODOLOGY

A variety of approaches may be used to address a single issue. This evaluates the measures of performance for every method. As a result, it evaluates the factors associated with effectiveness.

Optimal implementation of characteristics is commonly emphasised in problem-solving strategies. As a cornerstone of Problem-Solving Methods, this topic demands candid discussion.

Methods for fixing problems are able to be thus efficient because they presuppose certain aspects of the task and the resources at their disposal, such as domain knowledge. Because they provide light on the thinking underlying why Problem-Solving Methods work, clarifying these assumptions is crucial.

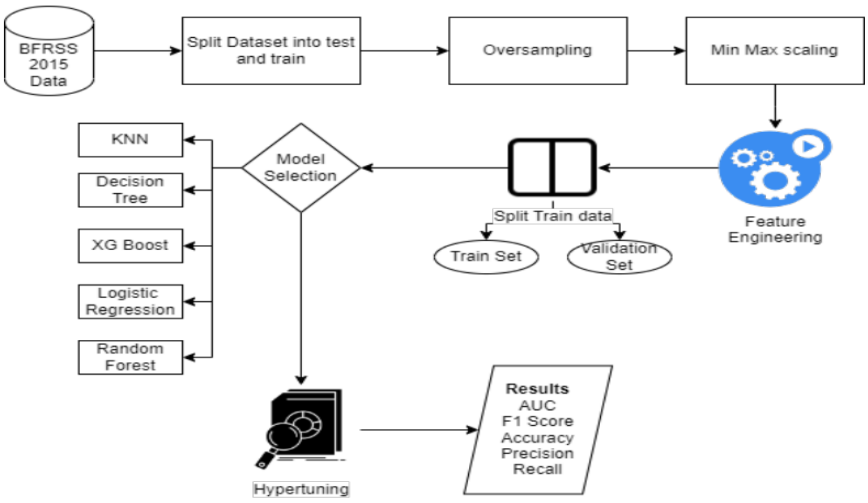
Algorithm-

• Select the optimal hyperplane for class differentiation. Identifying the optimal hyperplane necessitates determining the Margin, defined as the distance between the planes and the data points.

The probability of misclassification escalates as the distance between the classes diminishes and diminishes as the distance grows. We must guarantee that we Select the class with a substantial margin.

The margin is calculated by summing the distances to the positive and negative positions.

The comparison of results allows us to identify the most effective algorithm for diabetes prediction in our specific context. The diagram below shows the flow of the methodology



The first stage of deep learning involves creating separate training and testing sets from the entire dataset. The stratify option is used to maintain the original dataset's class distribution in both sets. The SMOTE methodology is used to achieve class parity by oversampling the training dataset, increasing the representation of under-represented groups. Min-max scaling is used to standardize attributes, ensuring equal influence on the model. The training data is split into a training and a validation set. Deep learning algorithms are selected based on their distinctive features and effectiveness in modeling diabetes risk factors. Metrics like F1-score, accuracy, precision, recall, and the area under the receiver operating characteristic (ROC) curve are used to assess the efficacy of each approach. A thorough analysis of the results helps choose the best diabetes prediction technique for the situation.

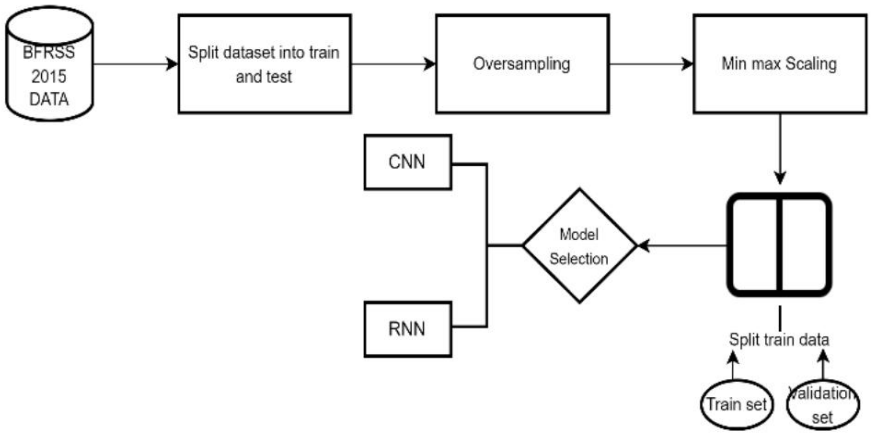


Figure 2. Flow Chart of Deep Learning Approach

The research utilized Python libraries and functions for data analysis, machine learning, and deep learning on the dataset. Pandas was used for data manipulation and analysis, while NumPy provided a centralized namespace for mathematical functions and support for arrays, matrices, and queues. Matplotlib was used for data visualization, offering static and animated representations of the loaded dataset. Data preprocessing and feature selection were performed using MinMaxScaler, SelectKBest, and chi2 from the sklearn.feature_selection library. The dataset was analyzed from the CDC's 2015 Behavioural Risk Factor Surveillance System (BRFSS) survey to identify Americans' health-related habits and potential dangers. The dataset included around 253,680 rows and 21 attributes, significantly enhancing the machine learning model's capacity to identify diabetes. Feature engineering was essential for enhancing the machine learning model's performance, with the Select Best approach using the chi-square statistical test. The research aimed to extract key information needed for accurate diabetes predictions, with 17 attributes identified that could improve the model. The chi-square test was used to evaluate the select best method, preserving crucial features for model prediction. Training and validation sets were created using a synthetic dataset relevant to the problem statement. After training on 90% of the dataset, the machine learning models successfully identified correlations between attributes and the target variable, diabetes status. To test the model's effectiveness, a 10% subset of the training dataset was used to generate the validation set. The validation dataset hyper-tuned parameters based on their performance on the validation dataset, optimizing all models' behavior to achieve the best possible results on new, unseen instances. A validation set was created using 10% of the overall dataset tuples to enhance the overall accuracy of the model. In the context of healthcare, Data Mining plays a significant role in anticipating infections, particularly diabetes, which is the leading global health concern.

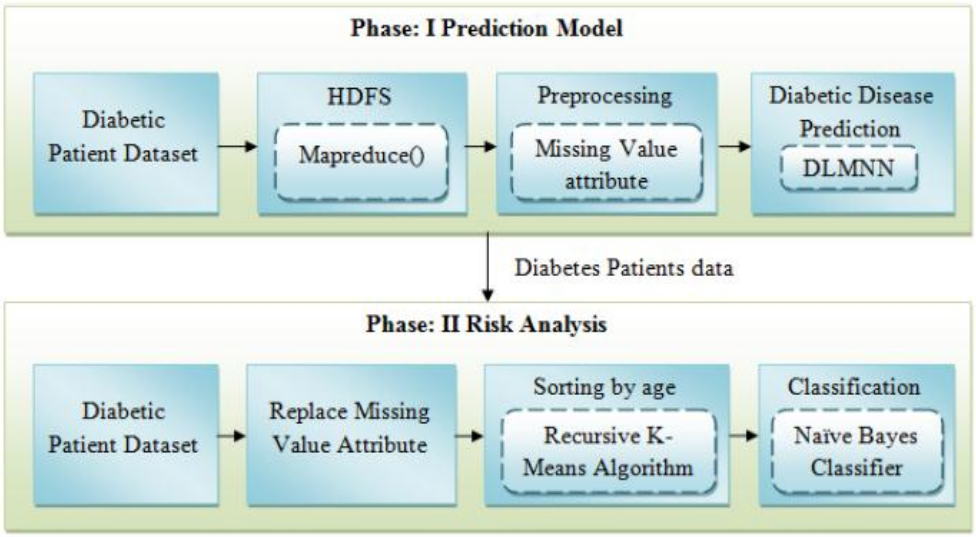


Figure 3 System Architecture

III. DIABETIC PATIENT DATASET

The Diabetes patient dataset, sourced from the PIDD repository, is used to analyze the likelihood of developing diabetes in individuals based on specific diagnostic criteria. The dataset includes 768 patient records with 8 diabetes-related characteristics. To eliminate redundant data, HDFS is used for deduplication, reducing storage requirements and optimizing network traffic. Entity resolution in massive datasets is achieved using MapReduce, which divides input data into blocks of similar records. Dedoop is used to eliminate duplicate models and efficiently allocate map-and-reduce tasks to the appropriate nodes within the cluster. MapReduce tasks are used to handle large volumes of data and provide analytical capabilities for analyzing data. The MapReduce framework simplifies data processing by dividing input data into subtasks and implementing sort or merge based on distributed computing. This approach allows for easy scalability and efficient data processing across multiple computing nodes.

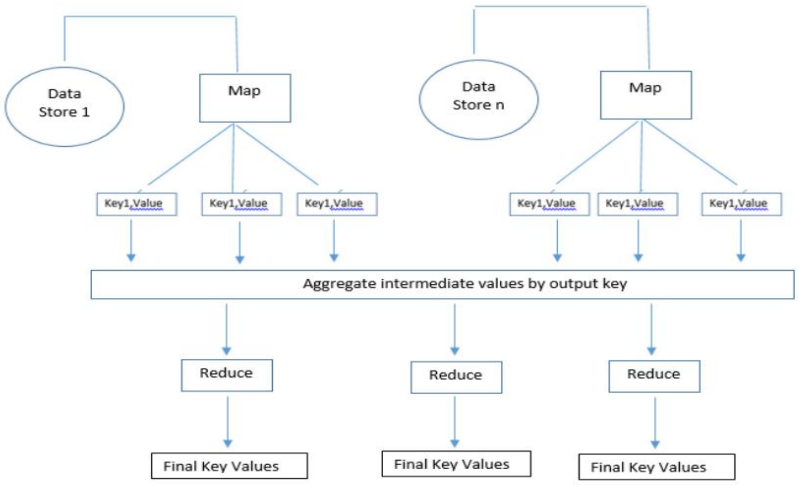


Figure 4 Map Reduce Architecture

Optimization Algorithm for Cuckoo Search

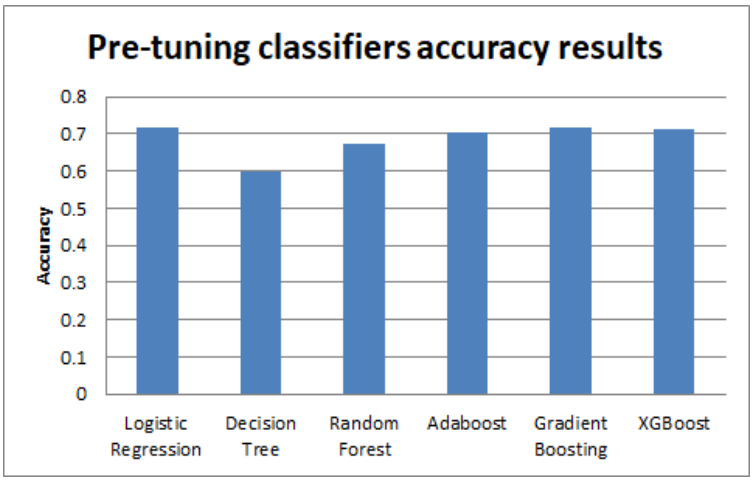
The article presents a global optimization method for cuckoo species, CSA, to improve dataset outcomes. The algorithm involves randomly selecting a nest and depositing eggs within it, ensuring superior quality. The model was tested on a different dataset, with the optimal configuration determined through hyperparameter adjustments and a 10-fold cross-validation approach. The study also explores the cost matrix for error classification, establishing a 3:1 ratio between false negatives and false positives. The model's performance was assessed through 10-fold cross-validation and the misclassification rate. The study also discusses logistic regression, a classification technique used in machine learning. Decision Trees are rule-based data classification systems that handle numerical and categorical data, making them easy to understand and visualize. They can handle both numerical and categorical data, but may become unstable due to their complexity. Random Forests use meta-estimators to create multiple decision trees from various datasets, increasing prediction accuracy and decreasing overfitting. AdaBoost is a revolutionary approach to boosting algorithms, used for binary classification problems and marketing. It combines shallow decision trees and AdaBoost with shallow decision trees to evaluate the importance of the next tree. Gradient Boosting is a method for training models that combines sequential and additive techniques, overcoming limitations of weak learners like decision trees. The optimized Gradient Boosting classification model includes DIQ010_2.0, a dataset for determining if a patient has diabetes.

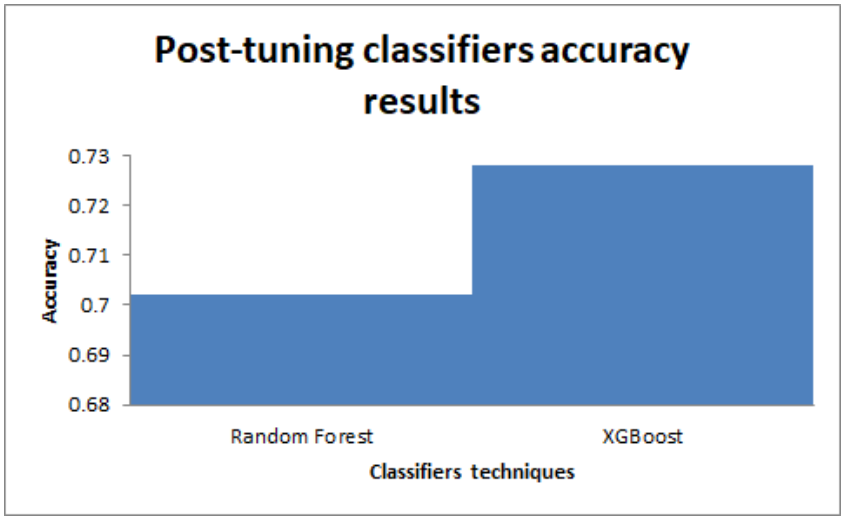
XGBoost

XGBoost is an advanced ensemble technique in machine learning that uses decision trees as part of the gradient boosting framework. It is particularly notable for its DIQ010_2.0 feature, which indicates the presence of diabetes in the patient's blood. The enhanced XGBoost model achieved high performance in non-diabetic patients, with a recall of 0.83, accuracy of 0.78, and F1-score of 0.81. The pre-diabetes metric set had an F1-score of 0.60, accuracy of 0.55, and recall of 0.57. The proposed solution involves data collection, data cleaning, training a machine learning model, evaluating its performance, and providing predictive assistance for food quality classification specifically tailored for diabetes management. The National Health and Nutrition Examination Survey (NHANES) dataset was used to analyze the data, and a rule-based system was developed to categorize individuals with diabetes into three distinct classifications: No Diabetes, Pre-Diabetes, and Diabetes. The accuracy formula is calculated by dividing the total number of potential outcomes by the total number of actual results. The recall metric quantifies the number of favorable results relative to the total outcomes observed. A classifier that

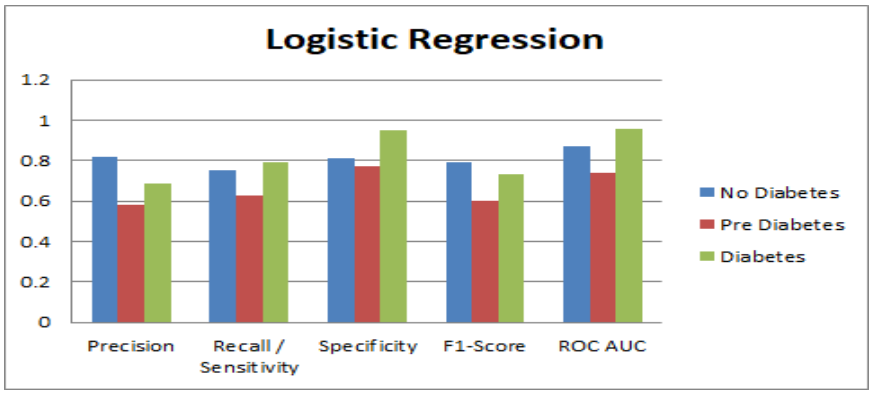
consistently attains a recall and accuracy of 1 indicates that it generates no false positives or negatives. The F1-score is a robust metric that effectively balances precision and recall, making it a valuable statistic in performance evaluation. Enhancing accuracy and memory efficiency results in a superior F1 score, representing the harmonic mean of recall and precision, offering a more nuanced metric.

You may divide the experimental results of classification methods into two groups: pre-tuning and post-tuning. In the image, we can see the results of comparing the accuracy of the classifier both before and after the adjustment.

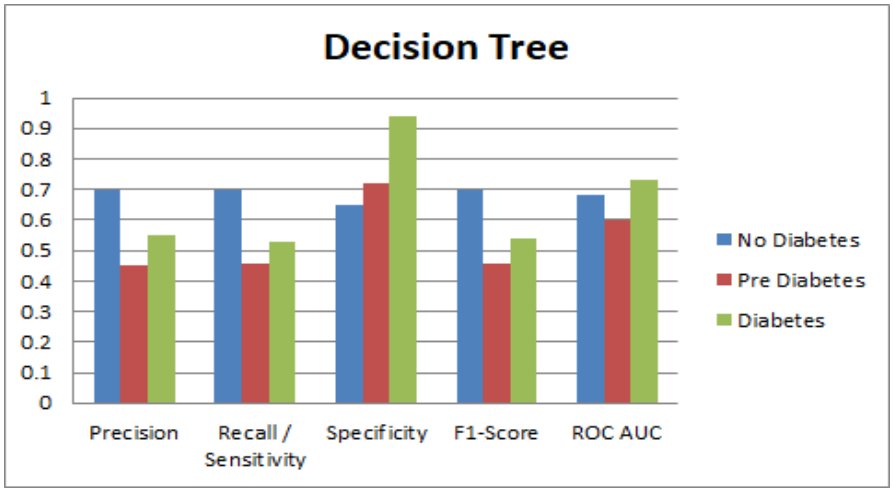




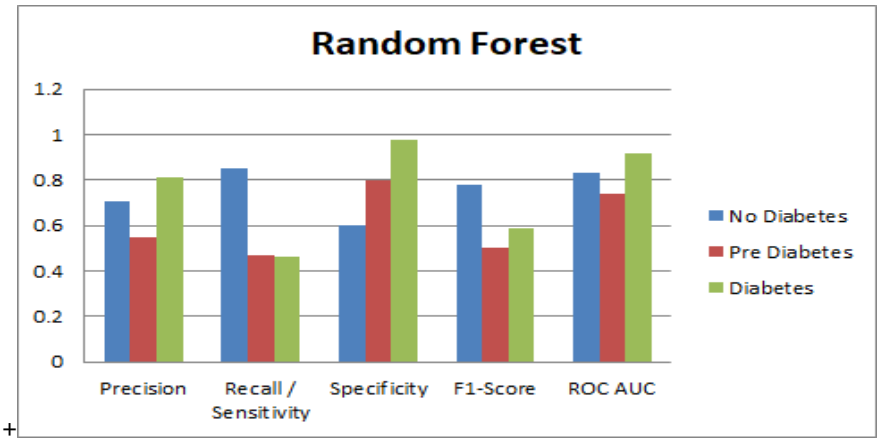
With an F1-score of 0.79 and a recall of 0.75, the logistic regression model achieves an overall accuracy of 0.72. For the group that does not have diabetes, the accuracy score is 0.82. The accuracy, recall, and F1-score of the pre-diabetes scale are 0.58, 0.63, and 0.61, respectively. Diabetes has an F1-score of 0.73, recall of 0.79, and accuracy of 0.69. When compared to prediabetes, this model's ability to forecast diabetes is much higher. Compared to other methods, prediabetes prediction does not significantly outperform them.



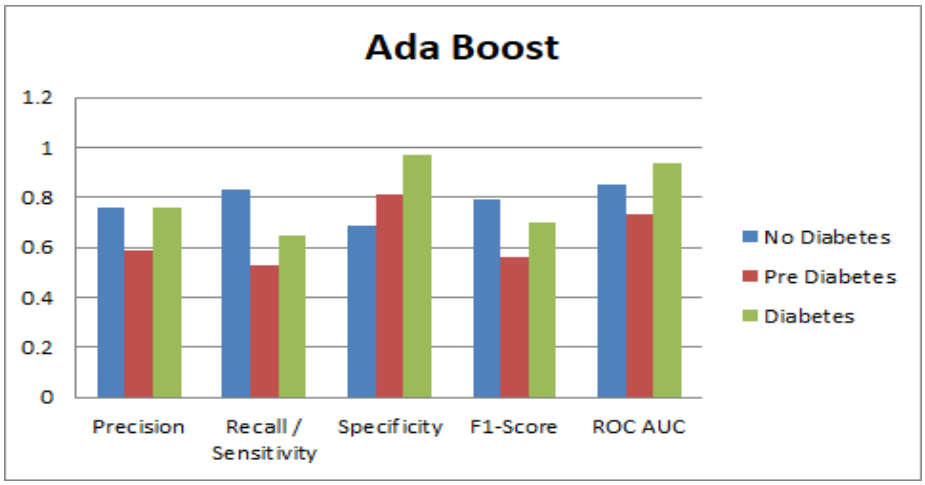
The decision tree model confirms diabetes diagnosis with a recall rate of 0.52% and accuracy rate of 0.54%, ensuring accuracy in healthcare. However, the One Vs Rest Classifier may make errors due to NaN parameters. Manual one-versus-rest processing is advantageous.



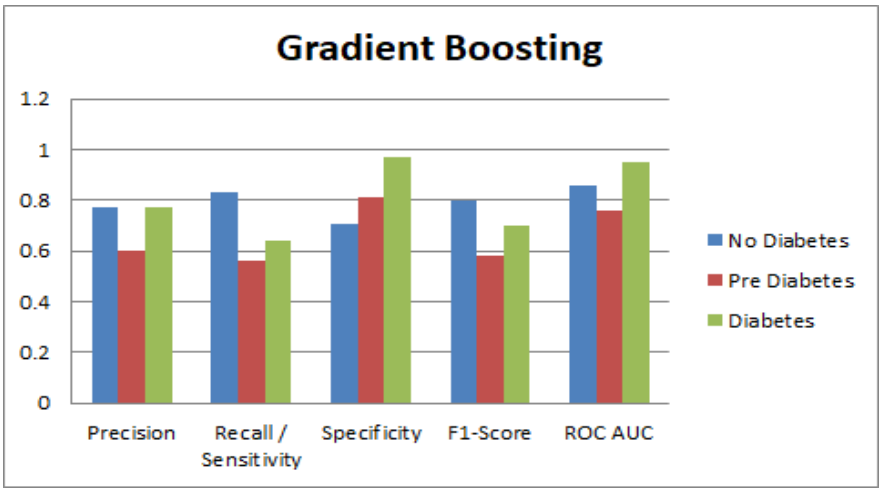
The Random Forest technique achieves a 0.67 accuracy rate in diabetes classification. The no-diabetes group has an F1-score of 0.78, while pre-diabetes has an F1-score of 0.50, recall of 0.46, and accuracy of 0.55. Diabetes has an F1-score of 0.59 and recall of 0.47, with a specificity of 0.98 for diabetes and 0.81 for prediabetes. Random Forest outperforms Decision Tree in performance.



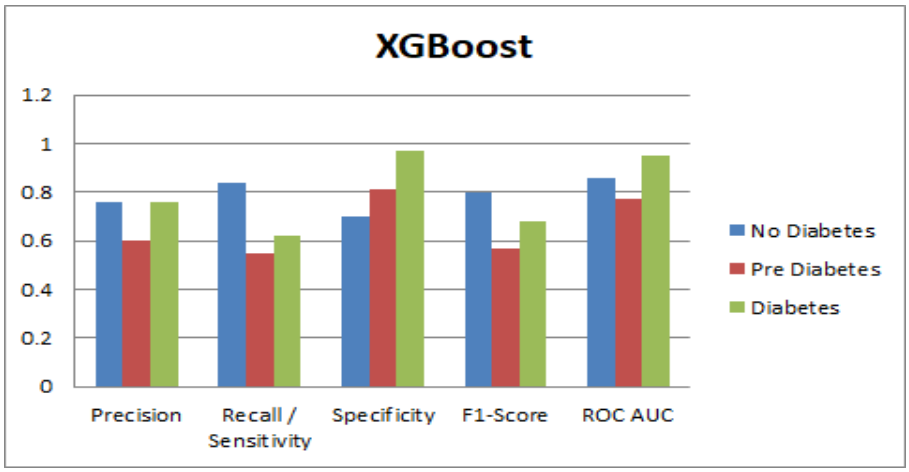
As a whole, the accuracy when using Ad boost for classification is 0.70, with non-diabetic scenarios yielding an F1-score of 0.79, recall of 0.83, and precision of 0.76. Accuracy of 0.59, recall of 0.53, and F1-score of 0.56 are the metrics used to identify pre-diabetes. Recall is 0.65, F1-score is 0.70, and accuracy is 0.76 for the diabetes model. For the absence of diabetes, the specificity value is 0.69; for prediabetes, it is 0.81; and for diabetes, it is 0.96. Compared to a Random Forest, Ad boost performs better when the latter is not optimized.



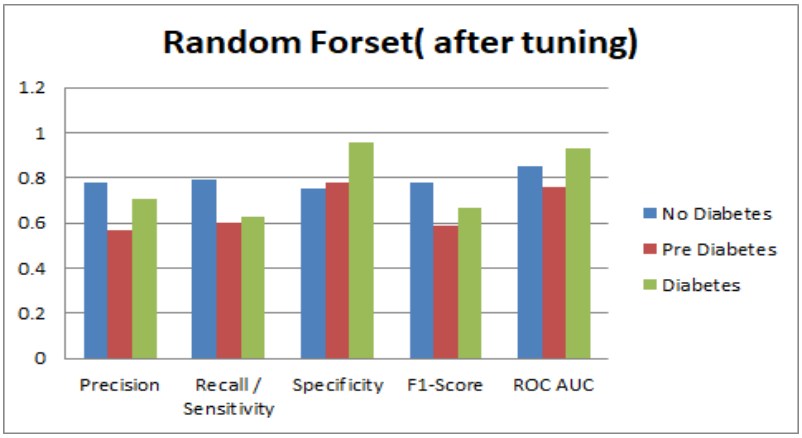
With respect to the no-diabetes category, the gradient boosting classification model achieves recall of 0.83, F1-score of 0.80, accuracy of 0.71, and precision of 0.77. Recall for pre-diabetes is 0.56, F1-score is 0.58, and accuracy is 0.60. The diabetes model achieves an overall accuracy of 0.77 with a recall of 0.64 and an F1-score of 0.70. The absence of diabetes is indicated by a specificity of 0.71, prediabetes by a specificity of 0.81, and the occurrence of diabetes is confirmed by a specificity of 0.97. When it comes to performance, Gradient Boosting is very much like the optimized Random Forest model.



The XGBoost model achieves an F1-score, recall, and accuracy of 0.76 for the non-diabetic sample. The pre-diabetes characteristics yielded an F1-score of 0.57, a recall of 0.55, and an accuracy of 0.60. The F1-score for diabetes, calculated using a recall of 0.62 and an accuracy of 0.76, results in a value of 0.68. With a specificity of 0.97, diabetes can be accurately identified; a specificity of 0.81 indicates the potential presence of prediabetes; and a specificity of 0.70 confirms that diabetes is absent.

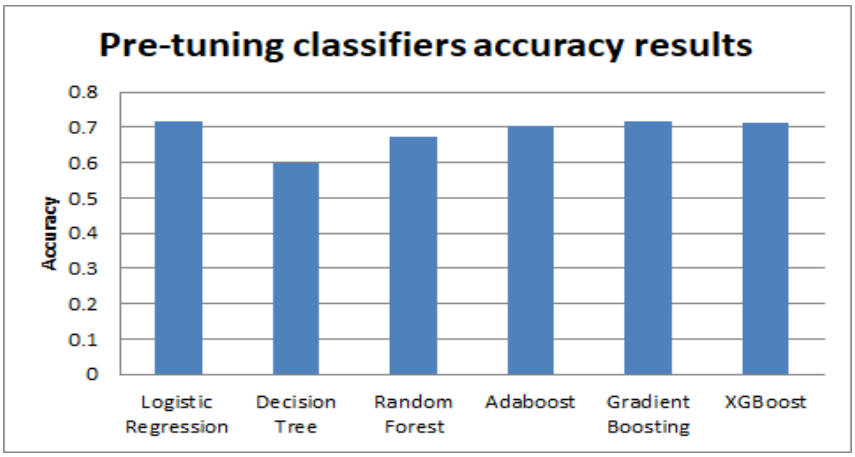


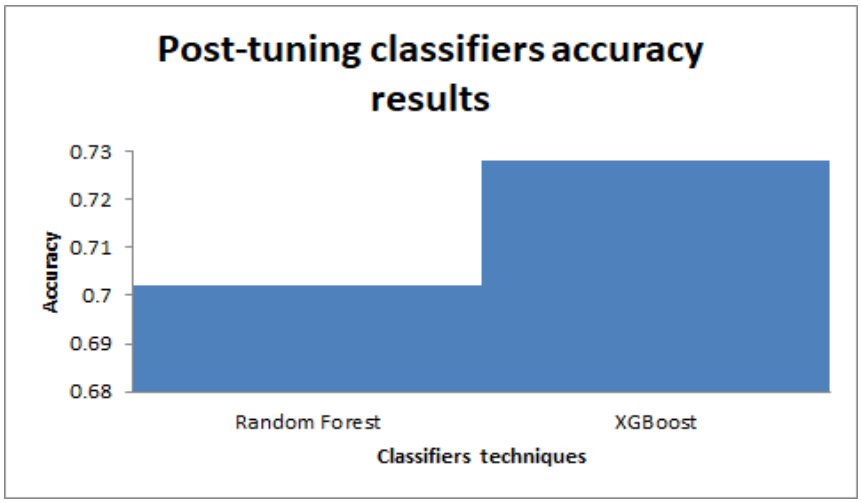
The optimized random forest classifier performed well in the diabetes-free patient group, with F1-scores of 0.78, recall of 0.79, precision of 0.78, and accuracy of 0.70. The diabetes model had F1-scores of 0.67, overall accuracy of 0.71, and recall rate of 0.63. The Random Forest approach surpassed the Decision Tree method, but not yet sufficient for accurately predicting prediabetes.

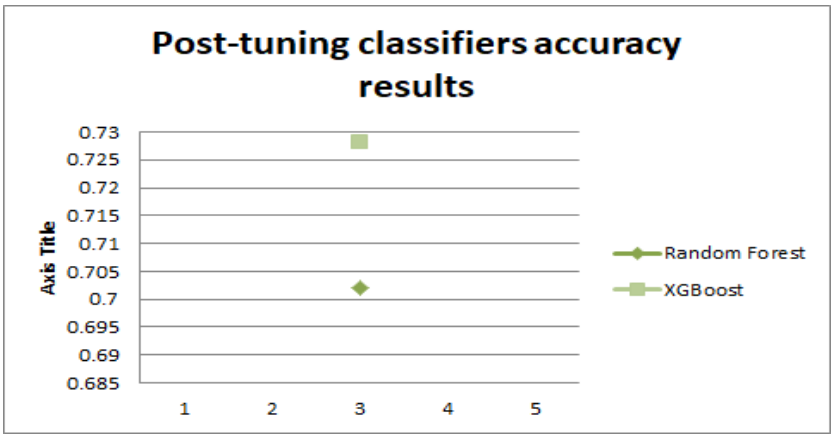


The modified XGBoost classifier achieves a recall of 0.83, an F1-score of 0.81, an accuracy of 0.72, and a precision of 0.78 for non-diabetic patients. An F1-score of 0.60, a recall of 0.59, and an accuracy of 0.61 suggest the presence of pre-diabetes. The recall stands at 0.64, while the F1-score is calculated to be 0.71, resulting in an accuracy value of 0.81 for diabetes detection. The specificity metrics for No Diabetes, Prediabetes, and Diabetes are 0.73, 0.81, and 0.98, respectively.

Experimental results from classification methods can be categorized into two types: pre-tuning and post-tuning. The image illustrates a comparative analysis of classifier accuracy pre- and post-adjustment.







IV. CONCLUSION

The study demonstrates the effectiveness of post-tuning classifiers when combined with Random Forest and XGBoost. Diabetes is a growing concern, affecting individuals of all ages and stages of life. The Internet of Things (IoT) in healthcare is a direct result of constant data collection from state-of-the-art systems. Predictive modelling techniques can help detect diabetic symptoms early and increase public awareness of the condition. Utilizing big data for diabetes prediction enhances patient comprehension compared to traditional methods. The proposed system's diabetes model prediction technique is more thorough, utilizing a variety of feature elements to illuminate patient's historical data and dietary patterns. Spark RDD and advanced machine learning approaches were used in the framework's creation.

REFERENCES

[1]. Peter H Scanlon, Stephen J Aldington and Irene M Stratton, Epidemological Issues in Diabetic retinopathy, Middle East Afr J Opthalmol; 2013; 20 (4), 293-300.

[2]. Rema M, Sujatha P, Pradeepa R., Visual outcomes of panretinal photocoagulation in diabetic retinopathy at one-year follow-up and associated risk factors, Indian J Ophthalmol; 2005; 53: 93-9.

[3]. Anuja Kumari, Kang, P, Ko, T, Cho, S, Rhee, SJ & Yu, KS 2013, 'An efficient and effective ensemble of support vector machines for antidiabetic drug failure prediction', Expert Systems with Applications, vol. 42, no. 9, pp. 4265-4273.

[4]. Begum, SA, Afroz, R, Khanam, Q, Khanom, A & Choudhury, TS 2014, 'Diabetic Mellitus and gestational Diabetic Mellitus', Journal of Paediatric Surgeons of Bangladesh, vol. 5, no. 1, pp. 30-35.

[5]. Chikh, M, Saidi, M & Settouti, N 2012, 'Diagnosis of diabetes diseases using an Artificial Immune Recognition System 2 (AIRS2) with fuzzy K-nearest neighbor', Journal of Medical Systems, vol. 36, no. 5, pp. 2721-2729.

[6]. Dipti N. Punjani & Kishor Atkotiya 2017, 'Data Mining and life science: a survey', International Journal on Recent and Innovation Trends in Computing and Communication, vol. 5, no. 7, pp. 633-636.

[7]. Durairaj, M & Ranjani, V 2013, 'Data Mining applications in healthcare sector: a study', International Journal of Scientific & Technology Research, vol. 2, no. 10, pp. 29-35.

[8]. Karthikeyani, V, Parvin Begum, I, Tajudin, K & Shahina Begam, I 2012, 'Comparative of Data Mining classification algorithm (CDMCA) in diabetes disease prediction', International Journal of Computer Applications, vol. 60, no. 12.

[9]. Pradeep, KR & Naveen, NC 2016, 'Predictive analysis of diabetes using J48 algorithm of classification techniques', 2nd International Conference on Contemporary Computing and Informatics (IC3I), IEEE, pp. 347-352.

[10]. Wernick, M.N., Yang, Y., Brankov, J.G., Yourganov, G. and Strother, S.C., 2010. Machine learning in medical imaging. IEEE signal processing magazine, 27(4), pp.25-38.

[11]. HR Ravikumar & Prasadu Peddi. (2024). An investigation into the use of deep learning and image processing in the domain of diabetic medical care. Internation Journal Of Advance Research And Innovative Ideas In Education, 10(4), 3377-3384.