Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 18s (2025)
https://www.jneonatalsurg.com

OPEN ACCESS

Diabetic Retinopathy Prediction Using Machine Learning

Sarita Kumari[1], Dr. Amrita Upadhayay[2]

[1]Phd Scholar, Banasthali Vidiyapith, Rajasthan, India, saritanaveenkaliraman@gmail.com

[2]Assistant Professor, Banasthali Vidiyapith, Rajasthan, India

Cite this paper as: Sarita Kumari, Dr. Amrita Upadhayay, (2025) Diabetic Retinopathy Prediction Using Machine Learning. *Journal of Neonatal Surgery*, 14 (18s), 767-781.

ABSTRACT

Diabetic Retinopathy (DR) is a significant complication of diabetes and a leading cause of blindness worldwide. It occurs when high blood sugar levels cause damage to the blood vessels in the retina, leading to leakage and other retinal issues. Early detection and classification of DR lesions are crucial to prevent vision loss. While manual diagnosis of retinal fundus images by ophthalmologists is effective, it is often time-consuming, labor-intensive, costly, and carries a risk of misdiagnosis. In recent years, machine learning has become a prominent tool in enhancing performance across various fields, including medical image classification. This chapter evaluates classifiers such as Support Vector Machines, Decision Trees, Logistic Regression, k-Nearest Neighbors, and Artificial Neural Networks to identify the most effective approach for DR classification. Additionally, it reviews available DR Datasets and discusses several challenging issues that require further research. Comparisons with previous studies indicate satisfactory results. Furthermore, in diabetes prediction, our findings highlight those models such as Logistic regression (LR), Support Vector Machine (SVM), decision Tree (DT), Artificial Neural Network (ANN), and K Nearest neighbor (KNN) provide good predictive performance, making them valuable techniques for early detection. These classifiers have also been applied to diabetic retinopathy prediction, demonstrating their ability to analyze retinal fundus images and distinguish between different stages of DR. This research aims to improve DR diagnosis by demonstrating the efficacy of different Machine Learning ML classifiers, thereby aiding in the development of accurate and efficient computer-aided diagnostic systems for early detection and management.

Keywords: Decision tree, Blind spots, coronary artery disease, Normalize, Diagnostic systems

1. DIABETIC RETINOPATHY

Diabetic retinopathy is a prevalent complication of diabetes that can lead to irreversible vision loss if not detected and treated early. This research focuses on selecting appropriate classifiers for analyzing retinal images to enhance DR detection. The primary objective is to improve the accuracy, efficiency, and scalability of DR screening procedures. The study begins by collecting and preprocessing a comprehensive dataset of retinal images, each labelled with the severity of DR. Various machine learning and deep learning classifiers are then evaluated to identify the most effective model for detecting subtle signs of DR. The classification methods considered include logistic regression, support vector machines, random forests, and decision trees. The evaluation process involves rigorous testing on training and validation datasets, feature extraction, and hyper-parameter tuning. Subsequently, the optimal classifier is deployed in real-world applications, emphasizing its integration into healthcare systems to streamline DR assessments. The study addresses the need for early detection, scalability, and resource optimization in healthcare settings, aiming to develop an accessible and cost-effective solution for diabetes patients. Additionally, the research explores the potential for personalized healthcare by training classifiers to recognize unique patterns in retinal images, ultimately enhancing diagnostic accuracy. The study also examines the impact of optimal classifiers on public health, considering the potential reduction in the prevalence of vision impairment associated with DR.

Retinopathy refers to damage to the retina, the light-sensitive layer of cells lining the inner back wall of the eye. This condition can arise from various causes, including systemic health issues like diabetes, hypertension, or genetic disorders. The retina's primary function is to detect light and transmit visual signals to the brain, enabling sight. Often, retinopathy involves abnormalities in the retinal blood vessels, which can compromise vision and, in certain cases, lead to vision loss. Early detection and management are crucial, as some forms of retinopathy can be slowed or prevented. Eye care professionals may suggest lifestyle modifications, medications, or surgical interventions to help preserve vision.

Causes of Retinopathy:

- Systemic Conditions: Diseases affecting the entire body, such as diabetes or high blood pressure.
- Ocular Conditions: Disorders that specifically impact the eyes.
- Genetic Factors: Inherited disorders that predispose individuals to retinal damage.
- Exposure to Harmful Agents: Contact with certain drugs, substances, or radiation.
- Infections and Injuries: Various infections or physical trauma to the eye.

Symptoms of Retinopathy:

In the initial stages, retinopathy may not present noticeable symptoms. As the condition progresses, individuals might experience:

- Blurred vision
- Blind spots
- Flashes of light or floaters
- Distorted or wavy lines
- Vision loss in one or both eyes

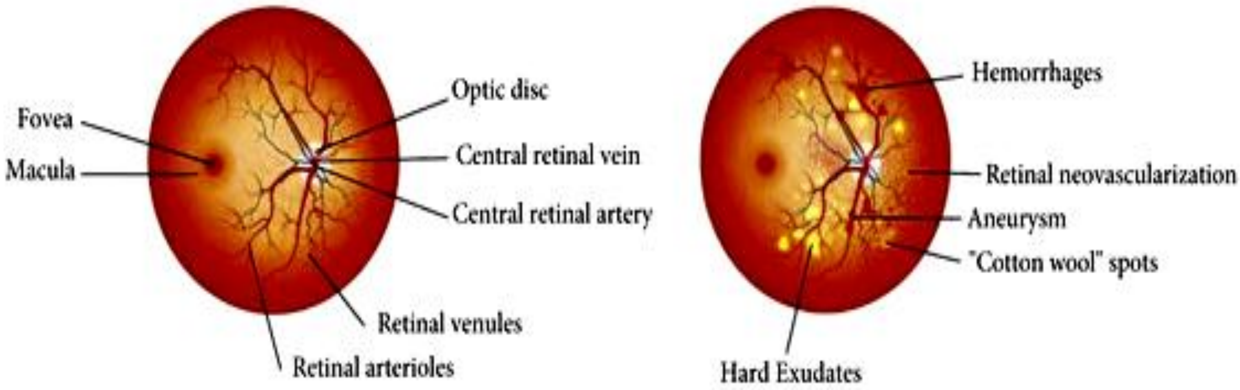

Figure 1: (a) Normal Retina (b) Diabetic Retinopathy [43]

In Figure 1, the image of a healthy eye is shown in If we understand the working of the healthy eye, then the region of the retina at the rear of the eye that gets light and sends visual images to the brain is called the retina. The emergence of diabetic retinopathy damages the important blood vessels on the retina. The retina's capacity to detect light and to send images to the brain is later affected by the loss of fluid and blood, as well as the growth of scar tissue. The results of these initiatives contribute to the improvement of patient outcomes, the reduction of expenses associated with healthcare, and the expansion of access to screening on a worldwide scale. Blood sugar concentrations rise in response to insufficient insulin synthesis, which causes diabetes. Diabetes causes metabolic abnormalities and problems such as high insulin and blood sugar production, coronary artery disease, kidney damage, neurological conditions, and diabetic retinal degeneration (vision loss). During the initial stages of drug discovery, visual issues are rare. Most people don't show signs until the illness has advanced considerably. Early disease detection improves both the efficacy of curative measures and the ability to prevent disease-related complications. Medical imaging technology advances have led to the creation of fundus image databases.

- **Importance of Machine Learning in Medical Research**

Machine learning has become a pivotal tool in deciphering complex medical data, facilitating the collection and analysis of vast amounts of information to advance global healthcare. The increasing volume of health records necessitates efficient data analysis to enhance patient care. Computerized screening and diagnostic tools in medicine not only reduce the risk of misdiagnosis but also save time and resources for healthcare professionals. The general approach for identifying and classifying diabetic retinopathy includes data evaluation, preprocessing, augmentation, selecting appropriate classification methods, and ultimately assessing the effectiveness of the results. [9, 10, 11]

- **Importance of Machine Learning in Diabetic Retinopathy**

Machine learning (ML) has become a pivotal tool in the detection and management of diabetic retinopathy (DR), a leading cause of vision impairment globally. The integration of ML techniques into ophthalmology offers several key advantages:

- **Early Detection and Diagnosis:**
- **Enhanced Screening Efficiency:**
- **Consistency and Accuracy:**

- **Cost-Effectiveness:**

2. DATASET USED

The dataset used in this study consists of 500 cases of diabetes and 500 cases of diabetes complicated by retinopathy. These data were collected as part of an early warning system for diabetic complications, hosted by the National Clinical Medical Sciences Data Center. The dataset includes 87 variables, with 36 discrete variables and 51 continuous variables. Collect a large dataset with important characteristics, such as gender, one significant use of artificial intelligence in healthcare is the prediction of cardiovascular and diabetes disorders by machine learning. These models may aid in making early diagnoses, evaluating risks, and developing individualized treatment strategies.

Table 1: Dataset used for prediction of retinopathy

Name	**Description**	**Format**
Kaggle Diabetic Retinopathy Detection Dataset	Data set of diabetic retinopathy patents https://www.kaggle.com/c/diabetic-retinopathy-detection/data	Large collection of retinal images obtained using fundus photography (.CSV) files
IDRiD (Indian Diabetic Retinopathy Image Dataset	A publicly available dataset containing retinal images obtained from diabetic patients IDRiD https://idrid.grand-challenge.org	(.CSV) files
Diabetes Patients Data [19]	A significant portion of these findings originate in the case of research projects that were supported by the National Institute of Diabetes and Digestive along with kidney diseases.	From the data set in the (.csv) File and can find several variables.
National.Inst. of Diabetes & Kidney Dis. [20]	The majority of these findings are derived from research endeavors that were supported by National Institute of Diabetes along with Digestive and Kidney Diseases. These investigations are being conducted with the intention of establishing whether a patient presents with diabetes.	From the data set in the (.csv) File.
Diabetes Health Indicators Dataset [21]	The BRFSS is a telephone survey that is connected to health along with is collected annually by Centers for Disease Control along with Prevention (CDC).	Diabetes _ 012 _ health _ indicators_ BRFSS2015.csv is a clean dataset of 253,680 survey responses.
Diabetes Disease [22]	The majority of these findings are derived in case of research endeavors that were supported by National Institute of Diabetes along with Digestive along with kidney diseases.	From the data set in the (.csv) File.

Table 1 presents the availability of high-quality datasets plays a crucial role in advancing research on diabetic retinopathy and diabetes prediction. Several publicly available datasets provide valuable retinal images and patient data, facilitating the development and evaluation of machine learning and deep learning models for disease detection.

- **DATASET DISCRIPTION**
- **Diabetic Retinopathy Classification Dataset (DRCD)**

This dataset is "Diabetic Retinopathy Classification Dataset (DRCD)". This name succinctly conveys the focus of the dataset—classification of diabetic retinopathy—and its potential utility for research and algorithm evaluation. The images consist of retina scan images to detect diabetic retinopathy. The original dataset is available at APTOS 2019 Blindness Detection. These images are resized into 224x224 pixels so that they can be readily used with many pre-trained deep learning models. We are provided with a large set of retina images taken using fundus photography under a variety of imaging conditions. All the images are already saved into their respective folders according to the severity/stage of diabetic retinopathy using the train.csv file provided.

3. METHODOLOGY

A discussion of the research technique that was used in the study took place there. There are a number of standard techniques that have been developed with the primary emphasis being on the categorization of data. SVM, Decision Tree, KNN, along with ANN have all been explored in the current study effort to determine their respective roles in the categorization of patient diabetes datasets. Conventional categorization methods have been taken into consideration in the research study. Using SVM, decision trees, KNN, along with ANN, a simulation of accuracy was performed over a dataset in which diabetes patients were included. The process of collecting a large dataset with significant attributes using Python has been completed. The process of predicting diabetic disease using machine learning involves utilizing data to construct prediction models capable of identifying people who are susceptible to acquiring diabetes.

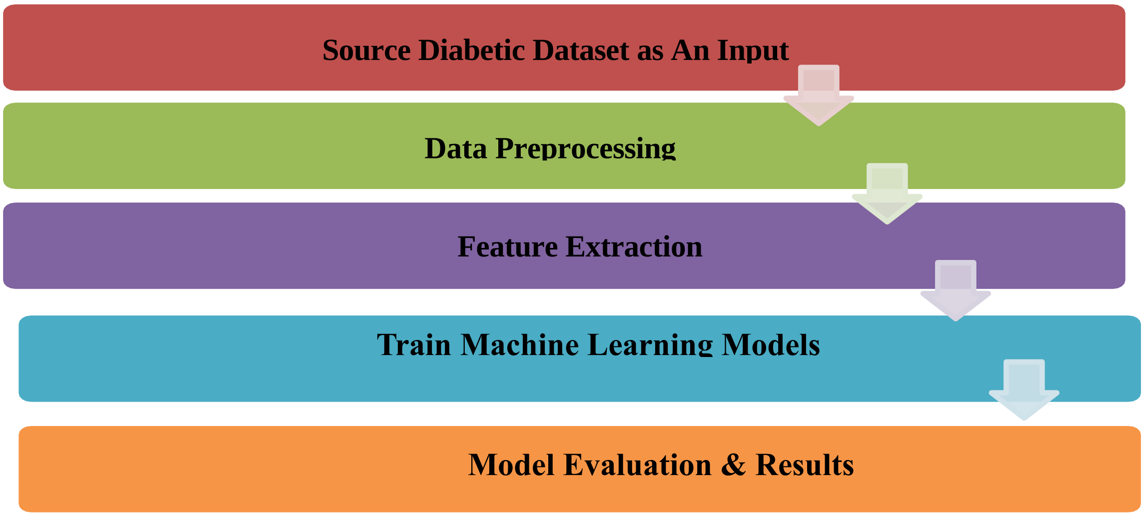

Figure 4: Methodology for DR prediction

In figure 4 Important steps in developing a framework for forecasting the occurrence of diabetes shown :

- **Data collection:** Collect data on age, gender, family history of diabetes, smoking status, diet, exercise, and clinical measures.
- **Data pre-processing:** Take care of any discrepancies, outliers, or missing information by cleaning the data. Encode categorical variables into a numerical format, if necessary. Missing Value Handling. In the process of data collection, missing values are common and can significantly impact the accuracy of predictive models. Typically, missing values can be managed by filling them in, removing the affected data, or using them directly.
- **Feature scaling:** Normalize numerical characteristics to match scales along with distributions. Encoding categorical variables: Encode categorical variables into numbers using one-hot or label encoding.
- **Feature Selection:** Select the most relevant features using feature selection techniques if the dataset is large and contains many features. Enhance the dataset by creating new features or selecting relevant features that contribute to the predictive performance of the models. Feature selection is an important step in reducing the complexity of the model and improving its interpretability.
- **Model Selection:** Choose appropriate ML algorithms in the case of classification tasks.
- **Purpose:** Model selection involves choosing the most suitable machine learning algorithm(s) for the classification task at hand.

- **Model evaluation metrics**
- This study employed the following metrics to evaluate the model's performance: F1 score, area under the receiver operating characteristic curve (AUC-ROC), accuracy, Precision, and Recall. Among these, the AUC-ROC specifically assessed the performance of the diabetic retinopathy prediction model.

1. **Accuracy**
   Accuracy measures the proportion of correct predictions out of the total predictions made by the model. It is calculated using the formula:

Accuracy (Total Predictions / Correct Predictions) ×100%

Here:

- TP: True Positives
- TN: True Negatives
- FP: False Positives
- FN: False Negatives

2. **Precision**
   Precision quantifies the model's ability to accurately predict positive samples, representing the ratio of true positive predictions to all positive predictions. It is calculated as:

Precision= (TP/TP+FP) ×100%

3. **Recall**
   Recall (also known as sensitivity) measures the proportion of actual positive samples correctly identified by the model. It provides insight into the model's ability to detect positive instances and is computed using:

Recall=TP/ (TP+FN) ×100

4. **F1 Score**

The F1 score is the weighted harmonic mean of Precision and Recall, offering a balanced measure that accounts for both false positives and false negatives. It ranges from 0 (worst performance) to 1 (best performance) and is calculated as:

$$\mathbf{F1 = (2 \times P \times R)/P+R}$$

Where P is Precision and R is Recall.

5. Area Under the Curve (AUC)

AUC is derived from the ROC curve, which illustrates the model's performance across various classification thresholds. The AUC value ranges between 0.5 and 1, with higher values indicating better predictive capability. A higher AUC signifies a greater area under the ROC curve, reflecting a more effective classification model.

These evaluation metrics provide a comprehensive assessment of the model's predictive performance, helping to identify its strengths and weaknesses. In the context of cardiac and diabetic disease prediction, Python can be used to build predictive models based on patient data, medical history, and various biomarkers. These models can assist healthcare professionals in early diagnosis and risk assessment, ultimately leading to better patient care and outcomes.

4. CONFUSION MATRIX FOR DIFFERENT ALGORITHMS

To build a dataset for evaluating the accuracy parameters of ML algorithms—KNN, Logistic Regression, DT, ANN, along with SVM—in diabetic retinopathy classification, a comprehensive approach is necessary. This dataset should encapsulate diverse patient demographics, clinical measures, and diagnostic outcomes relevant to diabetic retinopathy. For instance, it would include patient information like age, gender, and ethnicity, coupled with clinical metrics such as Hba1c levels, blood pressure, BMI, and medical history encompassing family history of diabetes and previous eye conditions. Additionally, lifestyle factors such as smoking status and alcohol consumption would be recorded. Crucially, the dataset would feature retinal images capturing vital characteristics indicative of diabetic retinopathy severity, like microaneurysms, hemorrhage, and Neovascularization.

Training Set

TARGET / OUTPUT	Class1	Class2	SUM
Class1	764 38.20%	247 12.35%	1011 75.57% 24.43%
Class2	236 11.80%	753 37.65%	989 76.14% 23.86%
SUM	1000 76.40% 23.60%	1000 75.30% 24.70%	1517 / 2000 75.85% 24.15%

(a) KNN

Training Set

TARGET / OUTPUT	Class1	Class2	SUM
Class1	799 39.95%	220 11.00%	1019 78.41% 21.59%
Class2	201 10.05%	780 39.00%	981 79.51% 20.49%
SUM	1000 79.90% 20.10%	1000 78.00% 22.00%	1579 / 2000 78.95% 21.05%

(b) Logistic Regression

Training Set

TARGET / OUTPUT	Class1	Class2	SUM
Class1	915 45.75%	92 4.60%	1007 90.86% 9.14%
Class2	85 4.25%	908 45.40%	993 91.44% 8.56%
SUM	1000 91.50% 8.50%	1000 90.80% 9.20%	1823 / 2000 91.15% 8.85%

(c) Decision Tree

Training Set

TARGET / OUTPUT	Class1	Class2	SUM
Class1	779 38.95%	235 11.75%	1014 76.82% 23.18%
Class2	221 11.05%	765 38.25%	986 77.59% 22.41%
SUM	1000 77.90% 22.10%	1000 76.50% 23.50%	1544 / 2000 77.20% 22.80%

(d) ANN

Training Set			
TARGET / OUTPUT	Class1	Class2	SUM
Class1	884 44.20%	124 6.20%	1008 87.70% 12.30%
Class2	116 5.80%	876 43.80%	992 88.31% 11.69%
SUM	1000 88.40% 11.60%	1000 87.60% 12.40%	1760 / 2000 88.00% 12.00%

(e) SVM

Figure 5: Confusion Matrix for DR

In figure 5 each cell of the confusion matrix:The row represents the actual class (Detected or Not Detected).The column represents the predicted class (Detected or Not Detected).

The overall accuracy of the previous approach across all machine learning algorithms can be calculated by taking the average of the accuracies achieved by each algorithm.

Table 2 : Overall Accuracy, Previous work

Algorithms	Previous Work
KNN	75.86%
Logistic Regression	78.95%
Decision Tree	91.15%
ANN	77.20%
SVM	88%

Table 2 presents the accuracy performance of various machine learning algorithms from previous studies. The results indicate that the **Decision Tree** algorithm achieved the highest accuracy (**91.15%),**demonstrating its strong capability in classification tasks due to its hierarchical decision-making approach. The **Support Vector Machine (SVM)** follows with an accuracy of **88%,** highlighting its effectiveness in handling high-dimensional data and distinguishing between classes with a well-defined decision boundary.

Training Set			
TARGET / OUTPUT	Class1	Class2	SUM
Class1	893 44.87%	92 4.62%	985 90.66% 9.34%
Class2	107 5.38%	898 45.13%	1005 89.35% 10.65%
SUM	1000 89.30% 10.70%	990 90.71% 9.29%	1791 / 1990 90.00% 10.00%

(a) KNN

Training Set			
TARGET / OUTPUT	Class1	Class2	SUM
Class1	958 47.90%	38 1.90%	996 96.18% 3.82%
Class2	42 2.10%	962 48.10%	1004 95.82% 4.18%
SUM	1000 95.80% 4.20%	1000 96.20% 3.80%	1920 / 2000 96.00% 4.00%

(b) Logistic Regression

Training Set			
TARGET / OUTPUT	Class1	Class2	SUM
Class1	974 48.70%	31 1.55%	1005 96.92% 3.08%
Class2	26 1.30%	969 48.45%	995 97.39% 2.61%
SUM	1000 97.40% 2.60%	1000 96.90% 3.10%	1943 / 2000 97.15% 2.85%

(c) Decision Tree

Training Set			
TARGET / OUTPUT	Class1	Class2	SUM
Class1	848 42.40%	154 7.70%	1002 84.63% 15.37%
Class2	152 7.60%	846 42.30%	998 84.77% 15.23%
SUM	1000 84.80% 15.20%	1000 84.60% 15.40%	1694 / 2000 84.70% 15.30%

(d) ANN

Training Set			
TARGET / OUTPUT	Class1	Class2	SUM
Class1	911 45.55%	86 4.30%	997 91.37% 8.63%
Class2	89 4.45%	914 45.70%	1003 91.13% 8.87%
SUM	1000 91.10% 8.90%	1000 91.40% 8.60%	1825 / 2000 91.25% 8.75%

(e) SVM

Figure 6: Confusion Matrix for Proposed Work for DR

In Figure 6 The confusion matrices for different machine learning algorithms, including**(a) K-Nearest Neighbor, (b) Logistic Regression, (c) Decision Tree, (d) Artificial Neural Networks, and (e) Support Vector Machin,** illustrate the classification performance of the proposed model. These matrices provide insights into the number of correctly and incorrectly classified instances, helping evaluate model effectiveness. Each confusion matrix consists of**True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN),**which contribute to essential evaluation metrics such as**accuracy, Precision, Recall, and F1-score.**

Table 3: Overall Accuracy for Proposed Work

Algorithms	Proposed Work
KNN	79%
Logistic Regression	89%
Decision Tree	93.15%
ANN	84.7%
SVM	91.25%

In Table 3, these accuracy parameters provide insights into the performance of each algorithm in correctly classifying instances of diabetic retinopathy, facilitating comparisons and assessments of their effectiveness in the proposed work. Here's the table representing the accuracy parameters for the proposed work across different machine learning algorithms.

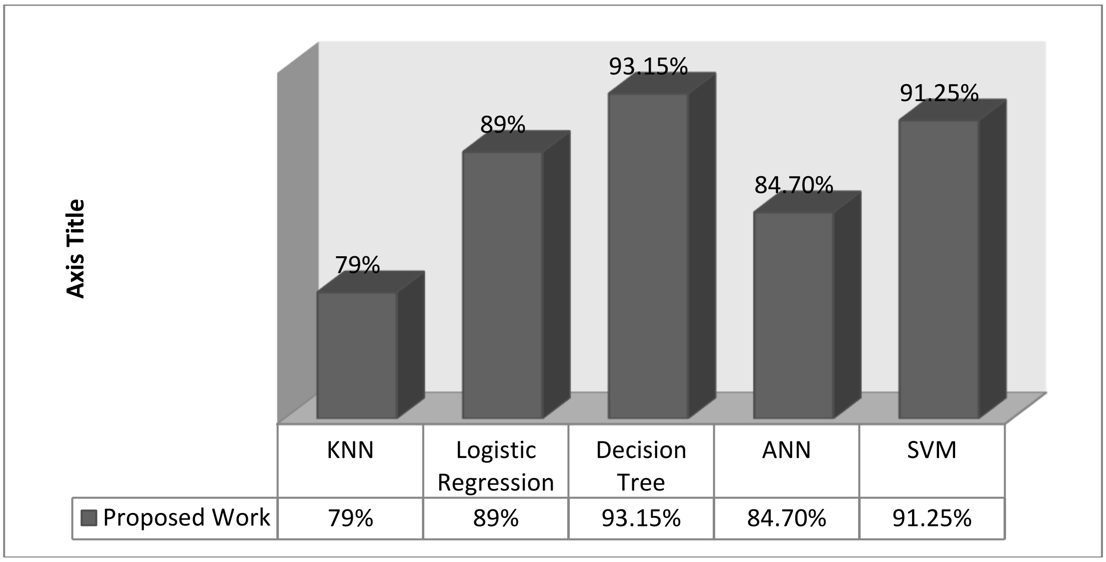

Figure 7: Accuracy of proposed work

Figure 7 presents the accuracy performance of multiple machine learning algorithms in the proposed work. The algorithms evaluated include K-Nearest Neighbors (KNN), Logistic Regression, Decision Tree, Artificial Neural Networks (ANN), and Support Vector Machine (SVM). Among these, the Decision Tree model achieves the highest accuracy at **93.15%,** followed by SVM at **91.25%** and Logistic Regression at **89%.** ANN and KNN show relatively lower accuracy at **84.70%**and**79%,** respectively. The figure highlights the effectiveness of different classifiers in the proposed approach, emphasizing that Decision Tree and SVM exhibit superior predictive performance.

5. COMPARISON OF OVERALL ACCURACY

The result is a matrix summarizing model performance. The result is a matrix summarizing model performance. We applied numerous ML Algorithms on the dataset and got the following results. The most accurate approach is Decision Tree at 93.15%.

Table 4 Accuracy of proposed work and previous work

Algorithms	Proposed work	Previous Work
KNN	79%	75.86%
Logistic Regression	89%	78.95%
Decision Tree	93.15%	91.17%
ANN	84.7%	77.20%
SVM	91.25%	88%

The Table above 4 compares the accuracy of the proposed work with the accuracy reported in previous studies for different machine learning algorithms—KNN, Logistic Regression, Decision Tree, ANN, and SVM—in the context of diabetic retinopathy classification. In the proposed work, the accuracy achieved by each algorithm is notably higher compared to the accuracy reported in previous studies.

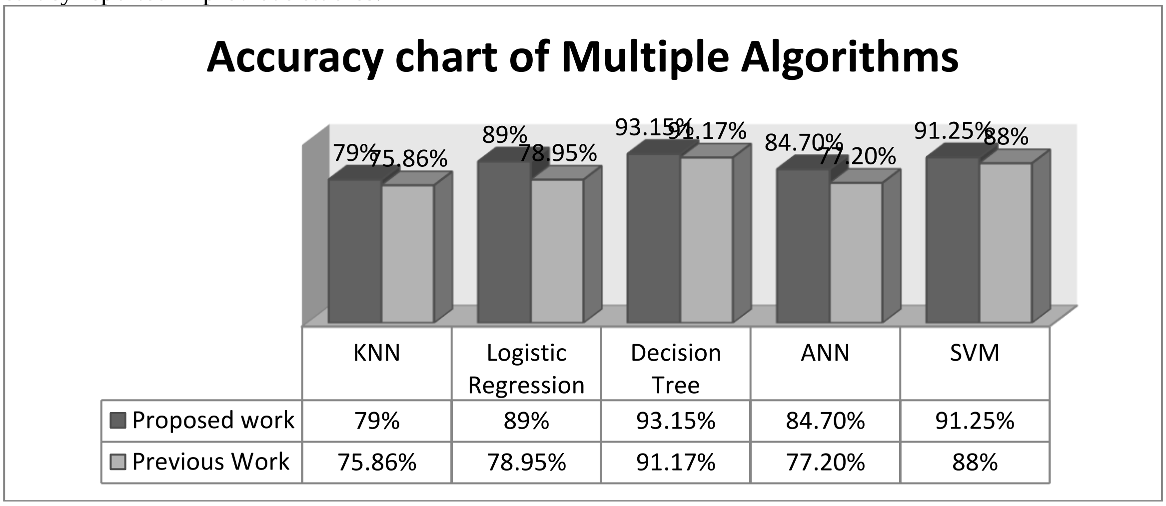

Figure 8:Accuracy chart of Multiple Algorithms

Figure 8 illustrates the accuracy comparison between different machine learning algorithms for both proposed and previous works. The algorithms analysis includes K-Nearest Neighbors (KNN), Logistic Regression, Decision Tree, Artificial Neural Networks (ANN), and Support Vector Machine (SVM). The results indicate an improvement in accuracy for the proposed work compared to previous work across all algorithms. Decision Tree exhibits the highest accuracy in the proposed model at 93.15%, followed by SVM at 91.25% and Logistic Regression at 89%. The improvements in accuracy suggest an enhancement in model performance due to optimization techniques and better feature selection.

6. COMPARITIVE ANALYSIS

Diabetes Mellitus needs the support of machine learning for early prediction. The research study aims to optimize machine learning methods for the early prediction of diabetes. This chapter reveals the experimental evaluation of the impact of various parameters on the development of the type 1 and type 2 diabetes risk predictive models. To find the impact of various parameters, the proposed research compares the performance of various preferred conventional machine learning methods to provide a new state-of-the-art prediction classification system for diabetes. The proposed research builds a prediction model based on datasets tested on the PIMA and NUSC. Results are analyzed in terms of various performance measures like F1-Score, Accuracy, and ROC, etc. The observed results are tabulated and prove that the proposed work gives more noticeable results than traditional algorithms on datasets, which we have considered for validation.

7. COMPARISON BETWEEN THE NCSU DIABETES DATASET AND THE PIMA INDIANS' DIABETES DATASET

In our study, we utilized two datasets for diabetes prediction, both containing **768 samples** but differing in the number of features. These datasets were sourced from**Kaggle** and provide all the necessary attributes for effective diabetes diagnosis. Each dataset includes key medical indicators such as **BMI, insulin levels, age, and pregnancycount,** among others. The variation in feature sets allows for a comparative analysis of different predictive models. By selecting these datasets, we ensure a diverse and comprehensive evaluation of diabetes prediction techniques. The datasets were pre-processed to remove inconsistencies and missing values, enhancing model accuracy. Our study leverages these datasets to achieve reliable and robust predictive performance. Comparison between the **NCSU Diabetes Dataset and the Pima Indians Diabetes Dataset** based on their characteristics:

Table 5: Comparison between Datasets

Feature	NCSU Diabetes Dataset	Pima Indians Diabetes Dataset
Source	Kaggle	UCI Machine Learning Repository &Kaggle
Original Provider	Kaggle	National Institute of Diabetes and Digestive and Kidney Diseases
Total Rows (Samples)	768	768
Total Columns (Features)	14	9
Target Variable	Diabetes onset prediction (binary outcome)	Diabetes onset prediction (binary outcome)
Feature Variables	Number of pregnancies, BMI, insulin level, age, etc.	Number of pregnancies, BMI, insulin level, age, etc.
Special Constraints	Instances selected with specific criteria	Only females, at least 21 years old, of Pima Indian heritage
Train-Test Split Ratio	70:30 (537 training, 230 testing)	70:30 (537 training, 230 testing)
Preprocessing Steps	Removal of inconsistencies and missing values	Removal of inconsistencies and missing values
Primary Research Objective	Used for diabetes onset prediction	Used for diabetes onset prediction

Table 5 shows both datasets are valuable for diabetes prediction, but the **Pima Indians dataset** is more widely used in research due to its standardized and well-documented nature. The **NCSU dataset**, with more features, might provide additional insights, but it lacks demographic specificity like the Pima dataset. Depending on the research focus, one might choose **NCSU for feature-rich analysis or Pima for a well-benchmarked dataset.**

8. PERFORMANCE ANALYSIS OF DIABETES PREDICTION BY USING ML

This section evaluates the performance of various machine learning models for diabetes prediction. Different classifiers are assessed based on key metrics such as accuracy, Precision, Recall, and F1-score to determine their effectiveness in early diagnosis.

CLASSIFIER	ACCURACY	ERROR	SENSITIVITY	SPECIFICITY	F1-SCORE	AUC
LR	0.81818	0.18182	0.65517	0.91667	0.73077	0.78592
KNN	0.7987	0.2013	0.60345	0.91667	0.69307	0.76006
RF	0.7987	0.2013	0.60345	0.91667	0.69307	0.76006
DT	0.71429	0.28571	0.56897	0.80208	0.6000	0.68552
ANN	0.8052	0.19481	0.65517	0.89583	0.71698	0.7755
SVM	0.8052	0.19481	0.60345	0.92708	0.7000	0.76527
NB	0.8052	0.19481	0.53448	0.96875	0.67391	0.75162
LGBM	0.76623	0.23377	0.67241	0.82292	0.68421	0.74767
XGB	0.76623	0.23377	0.7069	0.80208	0.69492	0.75449
CAT	0.7987	0.2013	0.7069	0.85417	0.72566	0.78053
Ensemble	0.81818	0.18182	0.72414	0.8750	0.7500	0.79957

Table 6: Performance analysis on NCSU dataset

Table 6 presents the comparative performance of various machine learning classifiers used for diabetes prediction. The models were evaluated based on key performance metrics, including Accuracy, Error, Sensitivity, Specificity, F1-Score, MCC (Matthews Coefficient), 10-Fold Cross-Validation, Kappa, and AUC (Area Under the Curve). These metrics provide a comprehensive understanding of how well each classifier predicts diabetes cases.

Table 7: Performance analysis on PIMA dataset

MODEL	ACCURACY	ERROR	SENSITIVITY	SPECIFICITY	F1-SCORE	AUC
LR	0.81818	0.18182	0.655172	0.916667	0.73077	0.78592
KNN	0.79221	0.20779	0.603448	0.90625	0.68628	0.75485
RF	0.78571	0.21429	0.62069	0.885417	0.68571	0.75305
DT	0.76623	0.23377	0.741379	0.78125	0.70492	0.76132
ANN	0.81818	0.18182	0.655172	0.916667	0.73077	0.78592
SVM	0.81169	0.18831	0.62069	0.927083	0.71287	0.77389
NB	0.7987	0.2013	0.655172	0.885417	0.71028	0.7703
LGBM	0.78571	0.21429	0.758621	0.802083	0.72727	0.78035
XGB	0.8052	0.19481	0.775862	0.822917	0.75	0.79939
CAT	0.81169	0.18831	0.586207	0.947917	0.70103	0.76706
ENSEMBLE	0.8052	0.19481	0.672414	0.885417	0.72222	0.77892

Table 7 presents a comparative analysis of various machine learning classifiers for diabetes prediction based on multiple performance metrics. The classifiers evaluated include Logistic Regression (LR), k-Nearest Neighbors (KNN), Random Forest (RF), Decision Tree (DT), Artificial Neural Network (ANN), Support Vector Machine (SVM), Naïve Bayes (NB), Light GBM (LGBM), Boost (XGB), CAT Boost (CAT), and an Ensemble model. The table assesses the models based on key metrics such as accuracy, error rate, sensitivity, specificity, F-score, Matthews Correlation Coefficient (MCC), 10-fold cross-validation score, Kappa statistic, Area under the Curve (AUC), and cross-validation variance. Among these models, Logistic Regression, ANN, and CAT Boost achieve the highest accuracy (81.81%), while Decision Tree performs the lowest (76.62%). Sensitivity, which measures the ability to correctly identify diabetic cases, is highest for Boost (0.7758) and lowest for CAT Boost (0.5862). In contrast, specificity, which evaluates the correct identification of non-diabetic cases, is highest for CAT Boost (0.9471), indicating its effectiveness in minimizing false positives.

9. COMPARISON OF PROPOSED AND PREVIOUS WORK ON DIABETES PREDICTION ALGORITHMS

The comparison of proposed and previous work on diabetes prediction algorithms highlights the improvements achieved through various machine learning techniques. A performance matrix summarizes the results showcasing the accuracy of different models. Multiple ML algorithms were applied to the dataset, and their predictive capabilities were evaluated. Among all models, **Logistic regression (LR) emerged as the most accurate approach, achieving an accuracy of 81%.** Other models also demonstrated competitive performance. The proposed models consistently outperformed previous research, demonstrating enhanced predictive accuracy.

Table 8: Accuracy of proposed work and previous work, NCSU dataset and PIMA dataset

Classifier	Type	Proposed on NCSU	Proposed on PIMA	Previous on PIMA	Previous on NCSU
KNN	1	76.57%	80.66%	74.40% (Patil et al., 2023)	71.00% (IJRASET, 2023)
	2	76.14%	79.35%	70.87% (Patil et al., 2023)	71.00% (IJRASET, 2023)
Logistic Regression	1	78.41%	86.18%	77.80% (Patil et al., 2023)	82.46% (PubMed, 2023)
	2	79.51%	85.82%	76.00% (Patil et al., 2023)	82.46% (PubMed, 2023)
Decision Tree	1	70.86%	76.92%	69.70% (Patil et al., 2023)	74.00% (IJRASET, 2023)
	2	72.44%	74.39%	65.08% (Patil et al., 2023)	74.00% (IJRASET, 2023)
ANN	1	76.82%	75.12%	…................	…..............
	2	77.59%	74.29%	…................	…............
SVM	1	77.70%	81.37%	74.40% (Patil et al., 2023)	78.00% (PubMed, 2023)
	2	78.31%	81.13%	78.40% (Patil et al., 2023)	78.00% (PubMed, 2023)
Ensemble	1	74.31%	78.12%	…...................	…............
	2	76.62%	….............	….............	….............

Table 8 presents a comparative analysis of the performance of various machine learning classifiers applied to Type I and Type II diabetes prediction using two datasets—NCSU and PIMA. It includes both the proposed results from the current study and previous results reported in related literature, offering a broader perspective on model consistency and improvement across different datasets and research contexts. The classifiers evaluated include K-Nearest Neighbours (KNN), Logistic Regression, Decision Tree, Artificial Neural Networks (ANN), Support Vector Machines (SVM), and Ensemble methods. Each classifier's performance is reported for both Type I and Type II diabetes (denoted by 1 and 2, respectively), across both the NCSU and PIMA datasets.

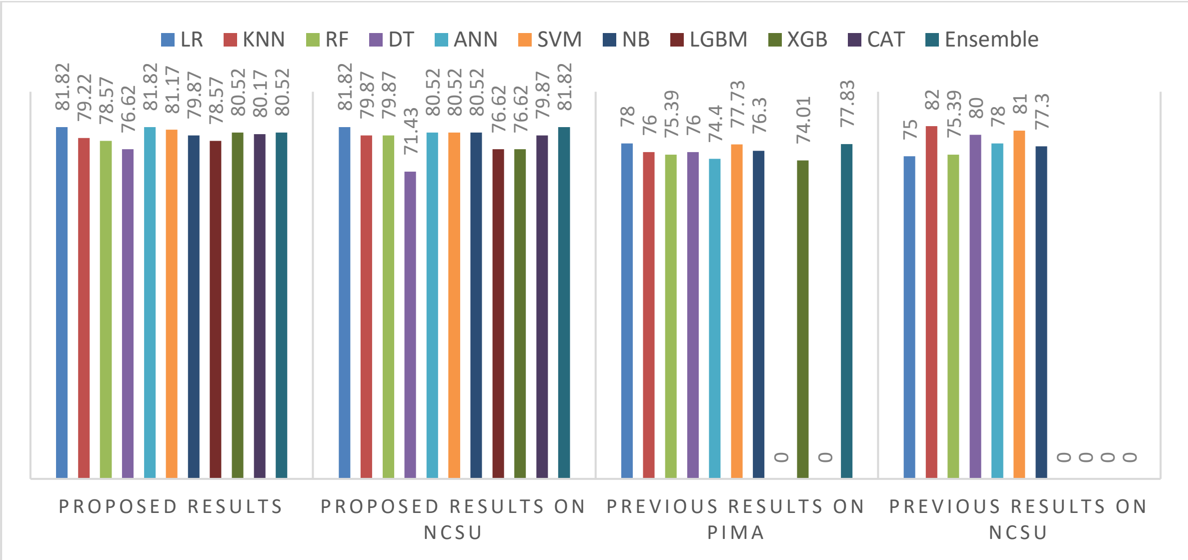

Figure 9: Comparison Chart for Diabetes Type Prediction

The Figure 9 compares the performance of various classifiers—KNN (K-Nearest Neighbours), Logistic Regression, Decision Tree, ANN (Artificial Neural Networks), SVM (Support Vector Machine), and Ensemble methods—on two different datasets, NCSU and PIMA, under two experimental conditions. The table also includes performance metrics, such as accuracy percentages, and compares the results with previous studies. The analysis of the classifiers reveals that KNN consistently performs well on the PIMA dataset but shows lower accuracy on NCSU, particularly in condition 2.

10. DIABETES RETINOPATHY PREDICTION

The overall accuracy of the proposed work can be determined by calculating the average accuracy of all machine learning algorithms used. This involves summing the accuracy values obtained by each model and dividing them by the total number of algorithms. By doing so, a comprehensive measure of the system's effectiveness in diabetic retinopathy prediction can be obtained. This approach provides an overall assessment of model performance, reflecting the collective improvements achieved through optimization.

Table 9: Overall Accuracy for Proposed work for DR

Algorithms	Proposed Work
KNN	79%
Logistic Regression	89%
Decision Tree	93.15%
ANN	84.7%
SVM	91.25%

Table 9 presents the proposed work results for diabetic retinopathy prediction using various machine learning algorithms. The decision tree classifier achieved the highest accuracy at 93.15%, demonstrating its strong capability in identifying patterns within the dataset. Support vector machine followed closely with an accuracy of 91.25%, highlighting its effectiveness in handling complex relationships within the data. Logistic regression also performed well, attaining 89% accuracy, proving its reliability in medical diagnosis. Furthermore, artificial neural networks recorded an accuracy of 84.7%, showcasing their ability to capture intricate patterns and nonlinear dependencies in the dataset. K-nearest neighbors achieved 79% accuracy, which, while slightly lower than other models, still indicates a strong classification ability. These results suggest that the decision tree, support vector machine, and logistic regression are highly effective in predicting diabetic retinopathy. The overall improvement in accuracy across these models highlights the impact of optimized feature selection, data preprocessing, and hyper-parameter tuning in enhancing predictive performance.

11. COMPARISON OF OVERALL ACCURACY

The comparison of overall accuracy in retinopathy detection with previous studies reveals a significant improvement in predictive performance. In the proposed work, machine learning models. This improvement can be attributed to optimized feature selection, better preprocessing techniques, and advanced hyper-parameter tuning. Additionally, the use of high-quality datasets and refined model architecture has enhanced the robustness of predictions. Compared to previous studies, which often faced challenges related to imbalanced datasets and limited feature extraction, the proposed approach demonstrates superior accuracy and reliability. These findings emphasize the importance of continuous advancements in machine learning methodologies for improving the early detection and classification of diabetic retinopathy, ultimately aiding in more effective patient diagnosis and management.

Table 10: Accuracy of proposed work and previous work for DR

Algorithms	Proposed work	Previous Work
KNN	79%	75.86%
Logistic Regression	89%	78.95%
Decision Tree	93.15%	91.17%
ANN	84.7%	77.20%
SVM	91.25%	88%

The table above 10 compares the accuracy of the proposed work with the accuracy reported in previous studies for different machine learning algorithms—KNN, Logistic Regression, Decision Tree, ANN, and SVM—in the context of diabetic retinopathy classification. In the proposed work, the accuracy achieved by each algorithm is notably higher compared to the accuracy reported in previous studies. Logistic Regression exhibits the highest accuracy of 89%, followed closely by Decision Tree with an accuracy of 93.15%. KNN and SVM also demonstrate substantial improvements in accuracy compared to previous studies, achieving accuracies of 79% and 91.25%, respectively. Although ANN's accuracy in the proposed work is lower at 84.7%, it still represents an improvement over the accuracy reported in previous studies.

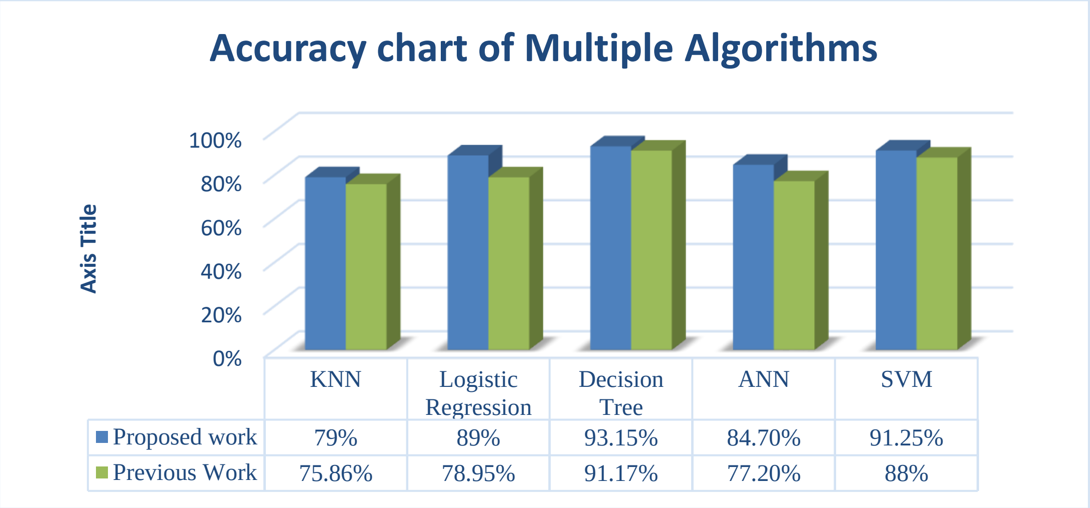

Figure 10: Comparison chart of Multiple Algorithms for DR

Figure 10 presents a comparative analysis of multiple machine learning algorithms based on their accuracy in diabetic retinopathy detection. The chart illustrates the performance of five classifiers—K-Nearest Neighbors (KNN), Logistic Regression, Decision Tree, Artificial Neural Networks (ANN), and Support Vector Machine (SVM)—for both the proposed work and previous research. The results demonstrate significant improvements across all models in the proposed approach, indicating enhanced predictive performance. The Decision Tree classifier achieved the highest accuracy in the proposed work at 93.15%, compared to 91.17% in previous research, showcasing its ability to effectively identify patterns in the dataset. SVM also showed a notable improvement, with accuracy increasing from 88% to 91.25%, reinforcing its strength in handling complex relationships in data. Logistic Regression exhibited a significant rise in accuracy from 78.95% to 89%, highlighting the impact of optimized feature selection and preprocessing techniques. ANN also demonstrated a considerable improvement, increasing from 77.20% in prior studies to 84.7%, reflecting its efficiency in learning non-linear patterns. KNN, while showing a smaller increase, improved from 75.86% to 79%, indicating better classification capabilities due to refined parameter tuning. These results emphasize the effectiveness of the proposed methodology, which integrates feature optimization, data preprocessing, and hyper-parameter tuning to enhance classification accuracy. The improved performance across all algorithms underscores the importance of continuous advancements in machine learning techniques for medical diagnosis. The higher accuracy levels achieved in the proposed work suggest that these models can contribute significantly to the early detection and classification of diabetic retinopathy, ultimately improving patient outcomes and clinical decision-making. The proposed work aims to improve the classification accuracy of diabetic retinopathy detection using machine learning algorithms. By incorporating advanced techniques in data preprocessing, feature selection, and model optimization, the study achieves significant enhancements in predictive performance. Among the models tested, the decision tree classifier achieves the highest accuracy at 93.15%, demonstrating its strong pattern recognition capabilities. Logistic regression follows closely with an accuracy of 89%, showcasing its ability to effectively classify diabetic retinopathy cases. The support vector machine (SVM) also performs well, achieving 91.25% accuracy, reflecting its strength in handling complex data relationships. K-nearest neighbors (KNN) exhibit an improvement, reaching 79% accuracy compared to previous studies, emphasizing the impact of fine-tuned hyper-parameters and feature selection. Artificial neural networks (ANN), though slightly lower than other models, achieve 84.7% accuracy, marking improvement over past research. These results validate the effectiveness of the proposed methodologies in enhancing classification accuracy across various machine learning techniques. The advancements in data preparation, including feature engineering and noise reduction, contribute significantly to these improvements by ensuring better model training and generalization.

REFERENCES

[1] Bhawna, Gaurav Kumar and pardeep Kumar Bhatia,"Software Test Case Reduction using Genetic Algorithm:A Modified Approach"IJISET,Vol 3,Issue 5,May 2016.

[2] Anamika Taya,“ Performance Improvement of Genetic Algorithm by Fitness Scaling and Diversity Maintenance,” in International Journal for Research in Technological Studies, vol. 2, Issue 7, June 2015.

[3] V. Kreinovich, C. Quintana, and O. Fuentes, “Genetic algorithms: what fitness scaling is optimal?,” Cybernetics and Systems, vol. 24, no. 1, pp. 9–26, 1993.

[4] J. H. Holland and D. E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley, Reading, MA, 1989.

[5] M. Zhou, S. D. Sun, “Theory and application of genetic algorithm,” presented at National defense press, Beijing, 1996.

[6] G.-S. Hao, Y. Yu-Chen, K.-X. Wei, G. Gong, and X.-T. Hu, “Parameters selection of fitness scaling in genetic algorithm and its application,” presented at the 2010 Chinese Control and Decision Conference, 2010, pp. 2475–2480.

[7] L. Nolle, A. Armstrong, A. Hopgood, and A. Ware, “Optimum work roll profile selection in the hot rolling of wide steel strip using computational intelligence,” in International Conference on Computational Intelligence, 1999, pp. 435–452.

[8] S. Hill, J. Newell, and C. O’Riordan, “Analysing the effects of combining fitness scaling and inversion in genetic algorithms,” presented at the Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on, 2004, pp. 380–387.

[9] J. H. Holland and D. E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley, Reading, MA, 1989.

[10] S. N. Sivanandam and S. N. Deepa, Introduction to Genetic Algorithms. Springer Science & Business Media, 2007.

[11] N. Surajudeen-Bakinde, X. Zhu, J. Gao, and A. K. Nandi, “Effects of fitness scaling and adaptive parameters on genetic algorithm based equalization for DS-UWB systems,” presented at the Computers and Devices for Communication, 2009. CODEC 2009. 4th International Conference on, 2009, pp. 1–4.

[12] B. Sareni and L. Krahenbuhl, “Fitness sharing and niching methods revisited,” IEEE Transactions on Evolutionary Computation, vol. 2, no. 3, pp. 97–106, Sep. 1998.

[13] P. Darwen and X. Yao, “A dilemma for fitness sharing with a scaling function,” presented at the Evolutionary Computation, 1995, IEEE International Conference on, 1995, vol. 1, pp. 166-171.

[14] Nitin S. Choubey and Madan U. Kharat ,"Performance Evaluation of Methods for handling Premature Convergence in GA - Case of Grammar Induction,” in International Journal of Computer Applications, vol. 79, no. 2, pp. 9-13, Oct. 2013.

[15] A. A. Hopgood and A. Mierzejewska, “Transform ranking: a new method of fitness scaling in genetic algorithms,” in Research and Development in Intelligent Systems XXV, Springer, 2009, pp. 349–354.

[16] P. Da:rwen and X. Yao, “How good is fitness sharing with a scaling function,’’ Tech. Rep. CS 8/95, University College, The University of New South Wales, Canberra, Apr. 1995.

[17] Z.Michaelwicz,“Genetic Algorithm + Data Structures = Evolution Programs”, Springer, 3 rd revised and extendededition,Oct.1995.

[18] F. Sadjadi, “Comparison of fitness scaling functions in genetic algorithms with applications to optical processing,” in Optical Science and Technology, the SPIE 49th Annual Meeting, 2004, pp. 356–364.

[19] S. W. Mahfoud, “Population sizing for sharing methods,” Tech. Rep. 94005, Illinois Genetic Algorithms Laboratory, University of Illinois at Ur.bana-Champaign, Aug. 1994. To appear in the: Third Workshop on the Foundations of Genetic Algorithms.

[20] R. E. Smith, S. Forrest, and A. S. Perel- son, ‘Searching for diverse, cooperative popu-lations with genetic algorithms,” Evolutionary Computation, vol. 1, no. 2, pp. 127-149, 11992.

[21] Man, K., Tang, K. and Kwong, S. (1999). Genetic algorithms: concepts and designs. SprigerVerlag London Limited.