Grenze International Journal of Engineering and Technology, June Issue

Forecasting Diabetes: A Machine Learning Approach for Predictive Analysis

Nikita Manohar Sable[1], Vijaya Kamble[2], Dr. Poonam Chaudhari[3] and Purva Gogte[4]
[1]Yeshwantrao Chavan college of Engineering, Department of Computer Science & Engineering, Nagpur, India,
Email: nikitaprasadgiradkar@gmail.com
[2]S.B.Jain Institute of Technology, Management & Research, Department of Computer Science & Engineering, Nagpur,India
Email: researchvijaya7@gmail.com
[3]GES's R. H. Sapat College of Engineering , Mangement Studies and Research, Department of MCA Engineering , Nashik, India
Email: poonam.chaudhari@ges-coengg.org
[4]Shri Ramdeobaba College of Engineering and Management, Nagpur, India, Department of Computer Science and Engineering, Nagpur, India
Email: purva.gogte1@gmail.com

***Abstract*— Diabetes is a serious condition that is primarily brought on by aging, high blood pressure, pregnancy, insulin, body mass index, and the diabetes gene. Remember that if diabetes is not treated, it can cause damage to other organs and lead to major health issues like high blood pressure, kidney, eye, and heart damage. For the best possible care, diabetes must be detected early. Our research uses a range of machine learning techniques to improve the accuracy of early diabetes identification in patients or persons. To anticipate when diabetes may manifest, we would employ machine learning, ensemble classification algorithms, and a dataset. Naïve Bayesian network, logistic regression (LR), K-Nearest Neighbor, random forests (RF), and support vector machines (SVM). The research results in a model that clearly demonstrates its ability to predict diabetes. Our results show that the Random Forest model performs more accurately than other machine learning techniques.**

***Index Terms*— Diabetes, Machine Learning Algorithms, Prediction, Dataset, Ensemble, Predictive analysis.**

I. INTRODUCTION

Diabetes stands as a detrimental global affliction, often triggered by factors such as pregnancies, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, age and elevated blood glucose levels [1]. It disrupts insulin, a pivotal hormone, leading to irregular carbohydrate metabolism and escalated blood sugar readings. The onset of diabetes emerges when the body's insulin production falls short. According to the World Health Organization (WHO), approximately 422 million individuals grapple with diabetes, predominantly prevalent in low-income or sedentary nations. This figure is projected to escalate to 490 million by 2030. However, the prevalence of diabetes extends across diverse regions, including Canada, China, and India. With India's population surpassing 1 billion, the diabetic populace in India alone has exceeded 40 million [2][3].
Diabetes ranks as a prominent contributor to global mortality. Proactively forecasting diseases like diabetes holds the potential to manage and preserve human lives. This undertaking thus delves into the prognosis of diabetes via

Grenze ID: 01.GIJET.10.2.104

an array of attributes associated with the condition. To this end, the Pima Indian Diabetes Dataset serves as the foundation, leveraging an assortment of machine learning classifiers and techniques for diabetes prediction. Machine learning serves as a strategy for training computers or machines to perform specific tasks. Employing diverse machine learning techniques yields noteworthy outcomes, fostering knowledge acquisition by erecting a variety of classification and ensemble models based on amassed datasets. This amassed data holds the promise of facilitating diabetes prognosis. While numerous machine learning approaches may offer predictive capabilities, selecting the most optimal technique remains a challenge [4]. Consequently, this endeavor employs widely accepted classification and ensemble methodologies on the dataset to ascertain the most effective predictive approach.

II. Literature Review

K. VijiyaKumar and colleagues [2] introduced a Random Forest algorithm aimed at forecasting diabetes, with the objective of constructing a system that can accurately anticipate diabetes onset for patients through the utilization of the Random Forest algorithm within the framework of machine learning. The model they proposed yielded superior outcomes in predicting diabetes, demonstrating its efficacy in accurately and swiftly identifying the disease. Similarly, Nonso Nnamoko and collaborators put forth an ensemble-based supervised learning strategy for predicting the onset of diabetes. They harnessed five extensively employed classifiers for these ensembles and employed a meta-classifier to amalgamate their outputs. The study's findings were showcased and juxtaposed against analogous research employing the same dataset. The results illuminated that the proposed method significantly enhances the accuracy of predicting diabetes onset.

Tejas N. Joshi et al [4]. presented a study titled 'Diabetes Prediction Using Machine Learning Techniques,' with the objective of forecasting diabetes utilizing three distinct supervised machine learning approaches: Support Vector Machine (SVM), Logistic Regression, and Artificial Neural Networks (ANN). This project puts forth an efficient method for the early detection of diabetes ailments. Dheeraj Shetty and colleagues proposed the utilization of data mining for predicting diabetes diseases, culminating in an Intelligent Diabetes Disease Prediction System. This system leverages a diabetes patient database to conduct an analysis of diabetes conditions. The authors advocate the incorporation of algorithms such as Bayesian and K-Nearest Neighbor (KNN) for application on diabetes patient databases, scrutinizing diverse diabetes attributes to predict the disease. Muhammad Azeem Sarwar and fellow researchers embarked on a study involving the prediction of diabetes through machine learning algorithms in the realm of healthcare. They employed six distinct machine learning algorithms, delving into the performance and accuracy of these applied techniques, subsequently drawing comparisons among them. By evaluating various machine learning methods, the study pinpointed the most suitable algorithm for diabetes prediction.

The anticipation of diabetes has garnered substantial attention from researchers, focusing on training programs to distinguish whether patients are diabetic through the implementation of appropriate classifiers on datasets. Building upon prior research, it becomes evident that there is room for enhancement within the classification process. Consequently, there arises a need for a system in the domain of computers to address the challenges identified in previous research, recognizing Diabetes Prediction as a pivotal sphere warranting attention.

In 2019, Mujumdar and Vaidehi[19] presented a study titled "Utilization of Machine Learning Algorithms for Diabetes Prediction" in the Procedia Computer Science journal (Volume 165, Pages 292-299). In this study, the authors introduced an advanced diabetes prediction model that incorporates supplementary criteria alongside the conventional ones, such as glucose levels, body mass index (BMI), age, and insulin levels. The utilization of these additional criteria resulted in an enhanced ability to categorize diabetes accurately when compared to previous approaches using a different dataset. Furthermore, they introduced a pipeline model dedicated to diabetes prediction, contributing further to the improvement of classification accuracy.

S M Hasan Mahmud et al.[5] A framework that is capable of efficiently tracking and monitoring people's diabetes and health conditions within an application perspective is required. We presented a real-time diabetes prediction, monitoring, and application framework in this study (DPMA). Our goal is to create an effective and efficient machine learning (ML) application that can accurately identify and forecast the state of diabetes.

Veena Vijayan.V et[6] al. AdaBoost technique for classification using Decision Stump as the basic classifier. In addition, the AdaBoost method uses Decision Trees, Naive Bayes, and Support Vector Machines as foundation classifiers to verify accuracy. When using Decision Tree as the basic classifier, the AdaBoost method yielded an accuracy of 80.72%, which is higher than Support Vector Machine, Naive Bayes, and Decision Tree combined.

III. PROPOSED METHODOLOGY

This text outlines the process of constructing a function model based on MLA (Machine Learning Algorithms) and the subsequent creation of performance evaluation datasets. The workflow employed in this research is illustrated in Figure 1. The dataset utilized for data analysis is derived from clinical diabetic data, detailing the sequential steps necessary for building a practical framework through ensemble learning-driven machine learning techniques.

Phase 1: Data preparation was executed to ensure data balance, encompassing both testing and training aspects.

Phase 2: Subsequently, model evaluation took place, with a specific focus on employing the 10-fold cross-validation And percentage distribution methodologies to assess the models' efficacy.

Phase 3: Patient symptom data, included within the dataset, was subjected to prediction algorithms, including RF, MLP, SVC, NB, and LR algorithms. Subsequently, a system was developed for end-users, utilizing the optimal algorithm and data repository.

Phase 4: Following the selection of the most precise classification models, the implementation of ensemble learning methods was executed.

Phase 5: This algorithm serves as a tool for users to predict risks based on symptom data. It comprehensively covers the process of implementing MLA for disease diagnosis, encompassing everything from data collection to achieving specific objectives.

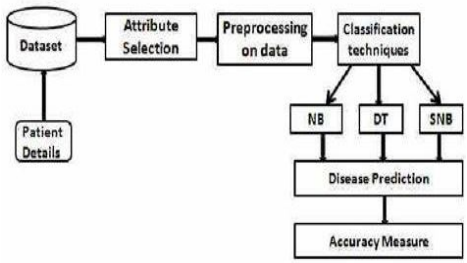

Figure 1. Building framework through ensemble learning-driven machine learning technique

About Dataset: The information is sourced from a UCI repository, specifically the Pima Indian Diabetes Dataset, featuring a multitude of attributes pertaining to 768 patients. To safeguard individuals' privacy, the dataset intentionally excludes personal details such as names or identification numbers. It's important to note that the dataset exhibits an unbalanced nature. The input features are used to predict the diabetes outcome. Once the model is trained, it can be used to make predictions on new, unseen data.

TABLE I: DIABETES DETECTION PATIENT DATASET

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
0	6	148	72	35	0	33.6	0.627	50
1	1	85	66	29	0	26.6	0.351	31
2	8	183	64	0	0	23.3	0.672	32
3	1	89	66	23	94	28.1	0.167	21
4	0	137	40	35	168	43.1	2.288	33

TABLE II: DIABETES PATIENT FEATURES DETAIL

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedig
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	

```
Out[45]: Pregnancies                 0
         Glucose                     0
         BloodPressure               0
         SkinThickness               0
         Insulin                     0
         BMI                         0
         DiabetesPedigreeFunction    0
         Age                         0
         Outcome                     0
         dtype: int64
```

Few data are visible for the columns Glucose, Insulin, skin thickness, BMI and Blood Pressure which have value as 0. That not possible, one cannot have 0 values for these as you can quickly search. Let's take care of that. We can either exclude such data or just substitute it with the appropriate mean values. Now substituting the column mean for zero.

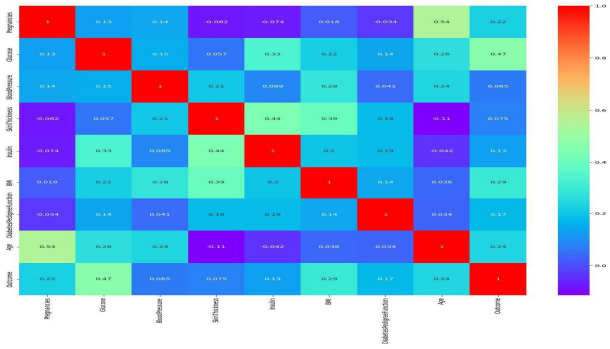

Figure 2. Correlation Plot

*Data cleaning***:** Data cleaning involves extracting raw information and employing various techniques to refine the data, including removing duplicates and irrelevant entries.
Data balancing: Data balancing is crucial since imbalanced classifications can pose challenges for predictive modeling.
Employing Machine Learning - Once the data is prepared, the application of machine learning techniques ensues. A diverse array of classification and ensemble methods is employed for diabetes prediction, all of which are applied to the Pima Indian diabetes dataset. The overarching objective revolves around the application of machine learning techniques to assess the usefulness of these methods and ascertain their correctness.
Support Vector Machine (SVM): The Support Vector Machine (SVM) involves the creation of a hyperplane that spans across different classes or entities. The hyperplane's dimensions are determined by computing the extent of the problem space. Additionally, SVM enables data dimension reduction to achieve equilibrium. By leveraging support vectors and class boundary points, the marginal distance between classes is computed from the hyperplane's center. SVM employs vital variables including kernels, coefficients, and intercepts. A pivotal facet of Support Vector Classification (SVC) resides in the kernel.These kernels are adaptable to the nature of the data they handle, with adjustments made accordingly [3]. The usage of linear and Gaussian kernels in this study is substantiated by the data conforming to linearity and radial basis function (RBF).

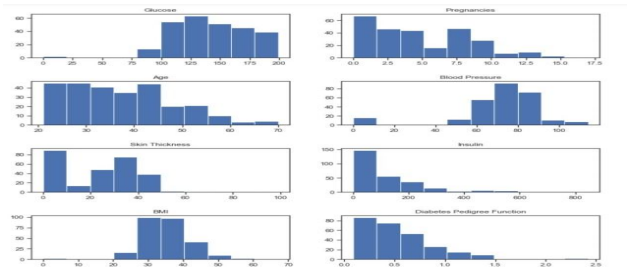

Figure 3. Histogram of each attribute

Algorithm:

Opt. for the hyperplane that achieves improved class division. To ascertain the optimal hyperplane, calculate the margin, which represents the distance between the hyperplanes and the data. When the inter-class distance is small, the likelihood of misclassification increases, and conversely. Thus, prioritize the class with a greater margin.

Margin = Distance to positive point + Distance to negative point.

Random Forest (RF): - In the realm of RF ensemble learning methodology, applicable to classification and various other tasks, decision trees are built through training alongside optimal tree approximations [2]. During the classification phase, the majority voting approach is employed to yield significant outcomes in the identification of diabetic conditions. Outcomes for a non-afflicted patient can be categorized as indicative of good health or a medical ailment.

Algorithm: Initial stage involves choosing "R" features from the overall feature set "m," where R is considerably smaller than M. From the selected "R" features, the node employs the most optimal splitting point. The node is divided into sub-nodes using this optimal split. Iterate through steps a to c until the desired count of nodes "l" is achieved. Construct a forest by replicating steps a to d "a" times, resulting in the creation of "n" trees.
Logistic Regression: Logistic regression, similarly to supervised classification techniques, is employed to assess the likelihood of a binary outcome, relying on single or multiple predictor variables, which can take on continuous or discrete values. This approach finds utility when the objective is to categorize data points into distinct groups. The binary classification nature involves assigning either a 0 or 1, often applied in scenarios like determining diabetes positivity in patients. Within the Logistic Regression framework, the sigmoid function assumes a pivotal role by assigning each data point a place an S-shaped curve. The sigmoid function's representation is outlined in the following equation:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad \text{..... (1)}$$

The main objective of logistic regression is to top match what is accountable on for describing the connection among the target and the predictor variable.
K-Nearest Neighbor: KNN also falls under the umbrella of supervised machine learning methodologies. It is adept at addressing both classification and regression tasks. Functioning as a lazy prediction technique, KNN operates under the assumption that entities sharing similarities tend to be situated in close proximity. This proximity translates to data points with resemblances often being spatially adjacent. KNN facilitates the categorization of novel instances based on measures of likeness. The KNN algorithm retains a comprehensive dataset of records, using them to classify based on degrees of similarity. A tree lecture approach aids in calculating distances between points. When predicting outcomes for a new data point, the algorithm recognizes the nearest data points in the training dataset, referred to as its optimal neighbors. In this context, "K" represents the amount of nearest neighbors, and a positive integer that guides the selection. The neighbor with the most votes from a set of classes is chosen. The proximity calculation primarily employs the Euclidean distance metric. For two points P (p1, p2, … pn) and Q (q1, q2, … qn), the Euclidean distance is defined by the subsequent equation:

Algorithm:

1. Begin by selecting a representative dataset consisting of columns and rows, denoted as the Pima Indian Diabetes dataset.
2. Acquire a testing dataset containing attributes and corresponding rows.
3. Utilize a formula to compute the Euclidean distance.
4. Proceed to select a random value for K, representing the number of nearest neighbors to consider.
5. Utilize the minimum distance in conjunction with the Euclidean distance to determine the values for the nth column of each entity.
6. Ultimately, deduce the corresponding output values in line with the above steps.

Naïve Bayes (NB): This segment presents the experimental outcomes achieved through the training of multinomial and Gaussian naive Bayes classifiers using a dataset of individuals with diabetes.
Confusion Matrix: In the realm of machine learning, the Confusion Matrix plays a pivotal role in assessing the effectiveness of a classification algorithm. This matrix assumes a tabular form where the rows correspond to the Actual class labels and the columns correspond to the Predicted class labels.

CONFUSION MATRIX STRUCTURE			
Total No. of Instances		**Predicted Class**	
		No (0)	**Yes (1)**
Actual Class	**No (0)**	True Negative	False Positive
	Yes (1)	False Negative	True Positive

The following definitions pertain to specific terminology within the comprehensive Confusion Matrix Structure. These terms will subsequently serve as the foundation for assessing the efficacy of each classifier.
Actual Class: The class label that signifies the original class before constructing the classifier.
Predicted Class: The class label that signifies the anticipated class after constructing the classifier.
True Positive: The count of instances accurately predicted as positive.

False Positive: The count of instances incorrectly predicted as positive when they are actually negative.
True Negative: The count of instances correctly predicted as negative.
False Negative: The count of instances incorrectly predicted as negative when they are actually positive

IV. Experimental Result

Several stages were undertaken within this study. The recommended methodology employs diverse classification and ensemble techniques, all implemented using the Python programming language. These approaches represent established machine learning methods utilized to attain optimal data accuracy. The findings of this study indicate that the random forest classifier outperforms other methodologies, producing superior results. Collectively, the study harnessed top-tier machine learning techniques to anticipate outcomes and achieve exceptional accuracy. The illustration depicts the outcomes derived from these machine learning methodologies.
This characteristic assumed a significant part in the estimate process associated with the random forest algorithm. The cumulative significance of each prominent attribute contributing to diabetes prediction was graphed, with the importance of individual attributes depicted along the X-axis and the corresponding name along the Y-axis.

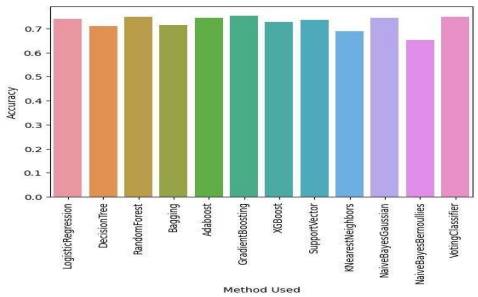

Figure 4. Accuracy Result of Machine learning methods

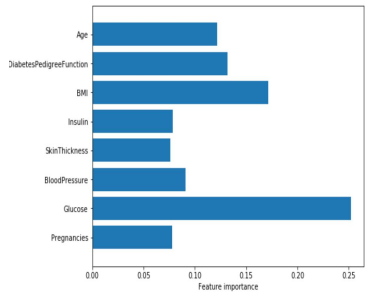

Figure 5. Feature Importance plot for Random Forest

IV. Conclusion

The primary objective of this endeavor centered around formulating and executing a diabetes prediction system through the utilization of machine learning techniques, with an emphasis on evaluating their effectiveness. This objective was successfully met. The devised approach encompasses diverse classification and ensemble learning methods, encompassing SVM, KNN, Random Forest, Naïve Bayes, and Logistic Regression classifiers. Ultimately, a classification accuracy rate of 85% was attained. The outcomes of the experimentation potentially pave the way for early healthcare prognostics and informed decisions for diabetes treatment, contributing to the preservation of human lives.

References

[1] Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analyzing Feature Importance's for Diabetes Prediction using Machine Learning". IEEE, pp 942-928, 2018.
[2] K.VijiyaKumar, B.Lavanya, I.Nirmala, S.Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes ".Proceeding of International Conference on Systems Computation Automation and Networking, 2019.
[3] Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus". International Conference on Electrical, Computer, and Communication Engineering (ECCE), 7-9 February 2019.
[4] Tejas N. Joshi, Prof. Pramila M. Chawan, "Diabetes Prediction Using Machine Learning Techniques".Int. Journal of Engineering Research and Application, Vol. 8, Issue 1, (Part -II) January 2018, pp.-09-1
[5] S M Hasan Mahmud "Machine Learning Based Unified Framework for Diabetes Prediction" BDET 2018: 2018 International Conference on Big Data Engineering and Technology Chengdu China August 25 - 27, 2018.
[6] V. Veena Vijayan " Prediction and Diagnosis of diabetes mellitus – A Machine learning approach" 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS) 10-12 Dec 2015.
[7] Nonso Nnamoko, Abir Hussain, David England, "Predicting Diabetes Onset: an Ensemble Supervised Learning Approach ". IEEE Congress on Evolutionary Computation (CEC), 2018.
[8] Dheeraj Shetty, Kishor Rit, Sohail Shaikh, Nikita Patil, "Diabetes Disease Prediction Using Data Mining "International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2017.
[9] Nahla B., Andrew et al,"Intelligible support vector machines for diagnosis of diabetes mellitus. Information Technology in Biomedicine", IEEE Transactions. 14, (July. 2010), 1114-20.

[10] A.K., Dewangan, and P., Agrawal, “Classification of Diabetes Mellitus Using Machine Learning Techniques,” International Journal of Engineering and Applied Sciences, vol. 2, 2015.
[11] S. Kumari, D. Kumar, and M. Mittal, “An ensemble approach for classification and prediction of diabetes mellitus using a soft voting classifier,” International Journal of Cognitive Computing in Engineering, vol. 2, pp. 40–46, 2021.
[12] E. O. Olaniyi and K. Adnan, “Onset diabetes diagnosis using artificial neural network,” International Journal of Scientific
[13] S. Gupta, H. K. Verma, and D. Bhardwaj, “Classification of diabetes using na¨ıve Bayes and support vector machine as a technique,” Lecture Notes on Multidisciplinary Industrial Engineering, Springer, Singapore, pp. 365–376, 2021.
[14] J. P. Kandhasamy and S. Balamurali, “Performance analysis of classifier models to predict diabetes mellitus,” Procedia Computer Science, vol. 47, pp. 45–51, 2015.
[15] D. K. Choubey, M. Kumar, V. Shukla, S. Tripathi, and V. K. Dhandhania, “Comparative analysis of classification methods with PCA and LDA for diabetes,” Current Diabetes Reviews, vol. 16, no. 8, pp. 833–850, 2020.
[16] S. Perveen, M. Shahbaz, A. Guergachi, and K. Keshavjee, “Performance analysis of data mining classification techniques to predict diabetes,” Procedia Computer Science, vol. 82, pp. 115–121, 2016.
[17] S. Gujral, “Early diabetes detection using machine learning: a review,” International Journal for Innovative Research in Science & Technology, vol. 3, no. 10, 2017.
[18] Ayush Anand and Divya Shakti,” Prediction of Diabetes Based on Personal Lifestyle Indicators”, 1st International Conference on Next Generation Computing Technologies, 978-1-4673-6809-4, September 2015.
[19] Aishwarya Mujumdar, V Vaidehi,” Diabetes Prediction using Machine Learning Algorithms,” INTERNATIONAL CONFERENCE ON RECENT TRENDS IN ADVANCED COMPUTING “ 2019 , ICRTAC 2019 27 February 2020, Version of Record 27 February 2020.
[20] Meshram, A., & Dembla, D. (2023, February). Multistage Classification of Retinal Images for Prediction of Diabetic Retinopathy-Based Deep Learning Model. In International Conference on Emerging Trends in Expert Applications & Security (pp. 213-226). Singapore: Springer Nature Singapore.
[21] MCBM: IMPLEMENTATION OF MULTICLASS AND TRANSFER LEARNING ALGORITHM BASED ON DEEP LEARNING MODEL FOR EARLY DETECTION OF DIABETIC RETINOPATHY Meshram, A., Dembla, D. ASEAN Engineering Journal, 2023, 13(3), pp. 107–116
[22] R. Bhave, P. G. Vyawahare and S. Shamkuwar, "“A approach for Hybrid Security Management in Health Care for Secure Cloud Storage: A Review”," 2022 Second International Conference on Next Generation Intelligent Systems (ICNGIS), Kottayam, India, 2022, pp. 1-3, doi: 10.1109/ICNGIS54955.2022.10079740.
[23] Multistage Classification of Retinal Images for Prediction of Diabetic Retinopathy-Based Deep Learning Model Meshram, A.,Dembla, D.
[24] Dembla, D., Meshram, A., & Anooja, A. (2024). Enhanced Diabetic Retinopathy Detection through Deep Learning Ensemble Models for Early Diagnosis. International Journal of Intelligent Systems and Applications in Engineering, 12(15s), 26-38
[25] Gauri D. Kalyankar, Shivananda R. Poojara and Nagaraj V. Dharwadkar,” Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop”, International Conference On I-SMAC, 978-1-5090-3243-3, 2017.