Advanced Predictive Modeling for Diabetes Detection: Enhancing Accuracy and Efficiency

Bavneet kaur, Research Scholar, Swami Vivekanand Institute of Engineering & Technology, Banur, Punjab, India, bona7667@gmail.com

Jyoti Kalia, Assistant Professor, Swami Vivekanand Institute of Engineering & Technology, Banur, Punjab, India, jyotikaliasharma121@gmail.com

Abstract Diabetes Mellitus (DM) is a condition caused by Hyperglycemia, wherein the ability of the body to produce or respond to insulin is restricted. This paper presents an effective and accurate diabetes prediction model that is based on Ensemble learning techniques. The main objective of the proposed approach is to increase the accuracy of diabetes prediction while reducing its complexity and dimensionality issues. To begin with, all the relevant and necessary information is taken from PIMA dataset, upon which pre-processing technique is implemented for making the data more informative and effective. After this, principal Component Analysis (PCA) technique is applied that selects only important and crucial features from the given dataset in order to reduce its dimensionality and complexity. Finally, a Light Gradient Boosting Machine (Light GBM) classifier is used for classifying patients as diabetic or non-diabetic. The performance of the proposed ensemble-bagging approach is analyzed and validated by contrasting it with traditional models in MATLAB software. The simulating outcomes were determined in terms of accuracy, precision, recall, F1-score and Area under Curve (AUC). The results of the simulation showed that the suggested model outperformed all existing models and could obtain an accuracy of0.9, which is far higher than the accuracy of conventional diabetes prediction models

Keywords —biomedical applications, healthcare systems, artificial intelligence, machine learning, diabetes detection, etc.

I. INTRODUCTION

Over the years, chronic and hereditary diseases that harm public health have been significantly on the rise. Among all disease, Diabetes Mellitus is one of the fast-growing fatal disease that damages major organs of body if not treated at right time. Diabetes is a condition that is developed due to the presence of too much sugar in body. Glucose is one of the key forms of energy used by our body to develop muscles and other bodily structures. Diabetes is brought on by an excessive amount of glucose in the bloodstream. In some cases, the pancreas seems to be unable to turn meals into insulin; as a result, sugar is not metabolized, leading to diabetes [1]. Diabetes Mellitus is usually caused by hyperglycemia in which the tendency of a body to produce or respond to insulin is affected [2-3]. The International Diabetes Federation estimates that 463 million people worldwide have diabetes as of 2019 and by the end of 2045, this number is expected to climb by 51% more. Furthermore, it is anticipated that there would be one undiagnosed person for every person who has been given a diabetes diagnosis [4]. Diabetes mellitus can be brought on by obesity, old age, sedentary lifestyle, genetic diabetes, high blood pressure, poor nutrition, and other factors. Diabetes raises the chance of developing conditions including heart disease, stroke, renal failure, nerve damage, vision problems, etc. over time. In addition to this, potential loss of vision due to retinopathy, renal failure due to nephropathy, foot ulcers, amputations, and Charcot joints due to peripheral neuropathy, and gastrointestinal, genitourinary, cardiovascular, and sexual dysfunction due to autonomic neuropathy are some of the long-term complications that can be caused by diabetes [5-6].

Usually, Diabetes Mellitus can be categorized into three types of, Type 1, Type 2 and Gestational diabetes. Type 1 diabetes mellitus, also known as insulin dependent diabetes, is a condition for which a patient requires insulin injections since the body is unable to manufacture enough insulin on its own. The major clinical symptoms are high blood sugar levels, increased thirst, and frequent urination. Type 2 diabetes is usually found in persons with age above 40 years due to the unhealthy and inactive lifestyle. The first symptom of type 2 diabetes is insulin resistance, a condition in which cells do not respond to insulin as they should. On the other hand, Gestational diabetes is usually found in pregnant women with no prior history of diabetes [7].

Diabetes is becoming more prevalent in people's daily

 DOI : 10.35291/2454-9150.2024.0052 © 2024, IJREAM All Rights Reserved.

lives as living standards rise. Therefore, it is worthwhile to research how to rapidly and effectively identify and assess diabetes. According to fasting blood glucose, glucose tolerance, and random blood glucose levels, diabetes is diagnosed medically [8]. Various risk categories have been developed for the first diagnosis of diabetes. The Finnish Diabetes Risk Score was deemed to be the best helpful tool for the first diagnosis of diabetes by the investigators. However, this system may be subject to human error because it calls for human intervention in the selection of the criteria and score. Due to the fact that Diabetes Mellitus is influenced by several additional elements and has profound socioeconomic consequences, a significant amount of data is eventually produced. As a result, machine learning (ML) and data mining Techniques (DMT) in DM are crucial, particularly when it comes to clinical administration concerns including diagnosis, management, and other related issues. Lately, a number of researchers are working on developing an effective and accurate diabetes prediction model for saving human lives.

The upcoming sections of this paper are organized as: Section 2 reviews and analyzes various ML based Diabetes prediction models followed up by gaps in them. Section 3 discusses proposed work and its working. The results obtained for the propose model are discussed in section 5 and finally a conclusion is written in section 6.

II. Related Work

During the tenure of last five to ten years, a significant number of ML and DL based Diabetes prediction models were proposed by various scholars. Some of the recently published works are discussed briefly here; G. Iswaria et al. [9], utilized ML techniques on the PIMA database for identifying and diagnosing Gestational Diabetes in patients. The efficacy of the suggested technique was elaborated in terms of accuracy, ROC, AUC as well as confusion matrix. U M Butt et al. [10], proposed an effective diabetes prediction model that worked in three phases of classification, detection and Prediction. The authors used ML classifiers which included LR, RF and MLP for categorization purpose, while as LSTM, MA and LR were used for predicting disease. The performance of their model was analyzed on PIMA dataset that revealed MLP achieves an accuracy of 86% for classification and LSTM achieved an accuracy of 87% for prediction. Z Quan, et al. [11], utilized and analyzed the performance of three ML algorithms i.e. DT, RF and NN on a real-world diabetes dataset for predicting diabetes in patients. Moreover, they used PCA along with minimum redundancy maximum relevance (mRMR) for reducing the dimensionality of database. Through extensive experiments, it was observed that RF outperforms other classifiers with an accuracy of 80%. K Saloni, et al. [12], proposed a diabetes prediction model wherein they used RF, LR and NB classifiers. The efficacy of the system was validated by comparing the performance of proposed approach with traditional models under various dependency factors. The results revealed that proposed model yielded highest accuracy of 79%. S Shahriare et al. [13], proposed an effective hybrid ML based model that was specifically designed for detecting Type 2 diabetes. The authors analyzed the performance of AdaBoost, NB, MLP, LDA, KNN, J48 and RF classifiers on the PIMA dataset. Results revealed that RF yielded highest accuracy of 99% and outperformed all other classifiers. I Md Merajul, et al. [14], analyzed the efficacy of six ML techniques like SVM, RF, LDA, LR, KNN and Bagged CART for identifying and detecting diabetes in patients. The results obtained revealed that bagged CART model outperforms all other existing approaches with an accuracy of 94%. Devi, R et al. [15], proposed an effective diabetes prediction model in which they used Farthest First clustering technique along with SMO algorithm. The effectiveness of the suggested scheme was analyzed on PIMA dataset with 768 samples and achieved an accuracy of 99%. A Karamath, et al. [16], proposed an improved diabetes prediction model in which they used PSO along with updated MLP and back propagation network for identifying and categorizing diabetes in patients. R Rohollah et al. [17], presented Logistic Adaptive Network-based Fuzzy Inference System (LANFIS) diabetes prediction model that was based on LR and ANFIS for detecting and categorizing patients as diabetic or non-diabetic. The suggested system yielded an accuracy of 88% which was better than other similar models. P. Roy et al. [18], proposed a model for detecting diabetes in patients in which they integrated CNN along with the dictionary-based techniques that include pathology specific images patterns. Simulating outcomes revealed that proposed model yielded quadratic kappa score (κ 2 = 0.86) which was far better than traditional models.

From the above literatures, it is observed that Diabetes is a chronic disease that needs timely and accurate detection in order to prevent any fatality. In the last few years, researchers have been working continuously for proposing an effective Diabetes prediction model. However, after analyzing few existing models we observed that they undergo through some limitations that degrade their overall accuracy. Majority of these models doesn't use any pre-processing technique for balancing and normalizing data, which leads to skewness and hence less accurate results. Furthermore, we also observed that most of the researchers were extracting less informative features which increased the processing and computations time of the model. In addition to this, it was also analyzed that current ML based Diabetes Prediction models are yielding low accuracy rate of 70 to 80%, which can be increased further. Furthermore, we observed that existing models were unable to handle large datasets and didn't contain any real world data. This

 DOI : 10.35291/2454-9150.2024.0052 © 2024, IJREAM All Rights Reserved.

leads to overfitting issues that further decreased model's efficacy. Keeping these findings in mind, a new and effective Diabetes prediction model is proposed in this paper for overcoming these issues.

III. PRESENT WORK

In order to overcome the limitations of conventional diabetes Prediction models, a new and highly accurate Diabetes prediction model is proposed in this work that is based on ensemble learning techniques. The main goal of the proposed diabetes prediction method is to increase the accuracy rate of diabetes diagnosis and reducing the complexity of model. Just like other automated diagnosis systems, the proposed diabetes detection model also undergoes through phases like Data collection, pre-processing for normalizing data, Feature selection for reducing dataset dimensionality and finally classification. The proposed model basically tries to improve the detection accuracy rate by modifying two phases of Feature selection and Classification. At the very beginning, data is taken from PIMA Indian Diabetes Dataset, which is one of the most widely used datasets used for detecting diabetes. Since, this data contains a lot of unnecessary and redundant and noisy data that can degrade the efficacy of our approach, therefore, it becomes important to pre-process the raw data. During the pre-processing phase, all the null values are removed, repeated values are deleted and string values are converted into numeric for better data representation. This step is followed up by the Feature selection phase wherein, Principal Component Analysis (PCA) is applied to processed data for selecting only critical and important features, so that the computational time and dimensionality of dataset is reduced. One of the important reason for using PCA in our is that it extracts patterns of only those features that are highly important and informative. The PCA extracts patterns of only those features that are highly important and informative. Additionally, by substituting the ensemble learning techniques for the particular classical ML classifiers, the results of proposed diabetes prediction model is improved. The bagging ensemble learning technique is used in the proposed study to help increase the model's overall recognition accuracy rate. The proposed model employs the Light Gradient Boosting Machine as the classifier (Light GBM). By employing the aforementioned strategies, this classifier evaluates the provided data successfully and predicts whether a patient has diabetes or not. The next portion of this paper discusses the suggested ensemble learning-based diabetes prediction approach's step-by-step operation.

A. Methodology

In this section, the detailed and sequential working of proposed Ensemble-bagging Diabetes prediction model is given. Each phase of diabetes prediction model is explained briefly below.

Step 1: The relevant information is primarily extracted from the PIMA dataset, which is essentially an Indian Diabetes Database of women who are 21 years of age or older. There are 768 entries in all, 268 of which are positive for diabetes, while the other 500 are negative. The dataset contains important features like total number of pregnancies, Glucose level, BP, Skin thickness, insulin, BMI, Diabetes pedigree function, age and Outcome that are shown in graphical form in figure 1.

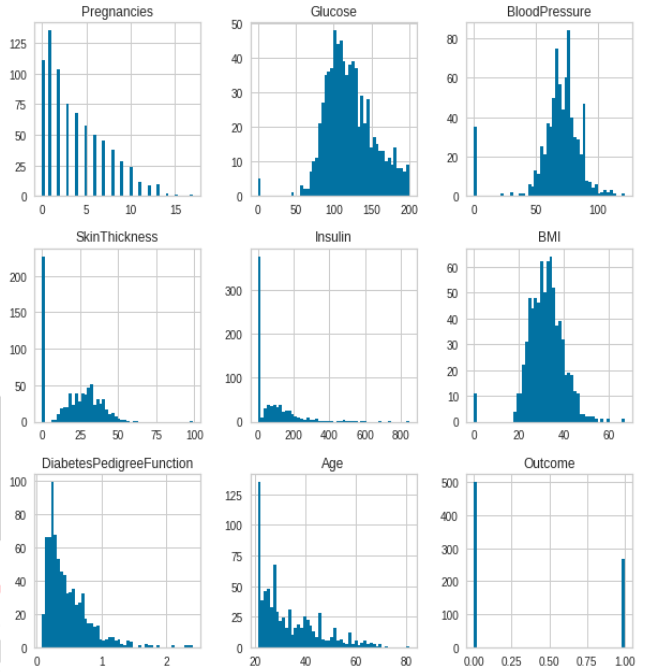

Figure 1. Attributes of PIMA dataset

Step 2: As mentioned earlier that original dataset contains lot of unnecessary data that can degrade the model's performance. Therefore, we implemented data pre-processing technique on raw data so that all the null, empty and redundant data is eliminated. Moreover, the string data present in dataset is converted into numeric values so that patterns can eb generated effectively. The aim of pre-processing is to make data more informative and effective.

Step 3: In the next step of proposed approach, PCA technique is used for selecting only crucial and important features among the nine features given in original dataset. The PCA reduces the dimensionality of dataset as well as makes the process of training classifier more effective. A heat map is created, as shown in figure 2 for nine features to examine which characteristics are more effective at predicting diabetes in a patient.

Step 4: The final featured dataset thus obtained is categorized into training and testing data. the training data is utilized for training the classifier whereas, its performance is tested by passing the testing data to it.

Step 5: After this, the process of analyzing and classifying data initiates in which bagging technique is implemented. A light Gradient Boosting Machine (Light GBM) classifiers is used in the proposed model for analyzing the testing data and predicting it as diabetic

 DOI : 10.35291/2454-9150.2024.0052 © 2024, IJREAM All Rights Reserved.

positive or negative.

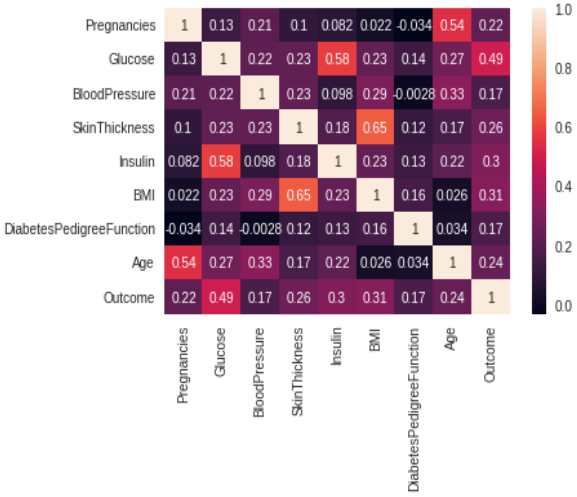

Figure 2. Heat Map for given features

Step 6: In the last step, the efficacy of the proposed approach on given dataset is analyzed and compared with few standard models in terms of various performance dependency factors that are explained in next section of this paper.

IV. Results and Discussions

The usefulness and effectiveness of the suggested ensemble-bagging diabetes prediction model is analyzed and validated by putting it in comparison with traditional models in MATLAB software. The simulating outcomes were obtained in terms of accuracy, precision, recall, Fscore and AUC respectively. The detailed description of these results are explained in followed up section.

A. Performance Evaluation

The suggested Ensemble-bagging diabetes prediction model's efficacy is initially evaluated in terms of its AUC curve. Figure 3 depicts the resulting curve, where the x-axis and y-axis are used to calibrate the False Positive rates and True Positive rates, respectively. The AUC curve is used to assess the efficacy of a proposed technique in distinguishing between diabetic and non-diabetic individuals. Following analysis of the provided graph, it is discovered that the suggested approach's AUC score for the PIMA dataset is 0.8965%.

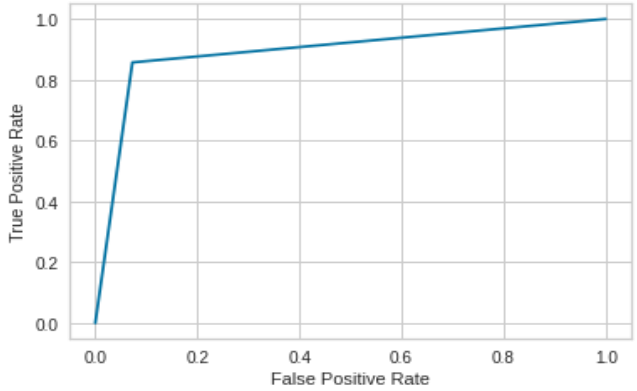

Figure 3 AUC in proposed Ensemble-bagging model

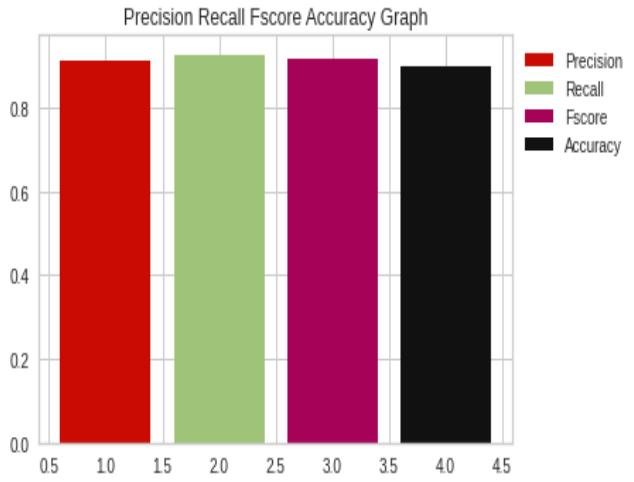

Figure 4. Performance matrices in proposed model

Additionally, parameters like accuracy, precision, recall, and Fscore are used to investigate and evaluate how well the suggested ensemble-bagging diabetes prediction model performs. The graph obtained for these parameters is shown in figure 4. After closely examining the provided graphs, it is evident that the precision, recall, and Fscore values all above 0.9, whereas the accuracy of proposed model came in at 0.9. These figures demonstrate the effectiveness and efficiency of the suggested strategy.

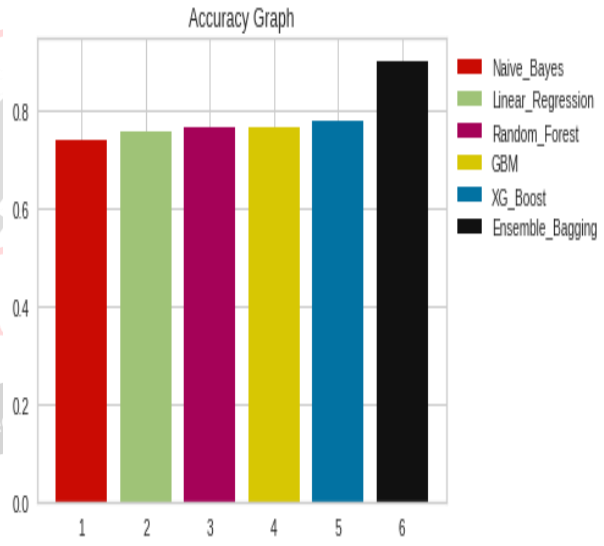

Figure 5. Accuracy comparison graph for proposed and traditional models

Furthermore, in order to prove the superiority of our ensemble-bagging diabetes prediction model, we analyzed and compared its performance with conventional models in terms of their accuracy value. The comparison graph obtained for the same is illustrated in figure 5. From the graph, it is demonstrated that the accuracy of standard Naive Bayes, Linear Regression, Random Forest, GBM, and XG-Boost models is only 0.73, 0.75, 0.76, 0.76, and 0.77, correspondingly. On the other hand, when the value of accuracy was determined for proposed ensemble-bagging diabetes prediction approach, it came out to be 0.9 that is far better than other existing models. The specific accuracy value are recoded in tabular form and is give in

 DOI : 10.35291/2454-9150.2024.0052 © 2024, IJREAM All Rights Reserved.

table 1.

Table 1: Comparison for accuracy values

Sr. No	Algorithms	Values
1	Naïve Bayes	0.73
2	Linear Regression	0.75
3	Random Forest	0.76
4	GBM	0.76
5	XG-Boost	0.77
6	Proposed Ensemble-Bagging	0.9

From the above given tables and graphs, it can be concluded that proposed Ensemble-bagging Diabetes Prediction model is yielding more productive and accurate results and hence can be used in real world application for determining whether a patient is diabetic or not.

V. CONCLUSION

This paper presents an effective and highly accuracy diabetes prediction model in which PCA and Light-GBM techniques have been used. The efficacy and supremacy of the suggested approach is examined and assessed in the MATLAB software by comparing it with few traditional models in terms of accuracy. The results simulated that proposed ensemble-bagging diabetes prediction approach is able to detect diabetes in patients with an accuracy of 0.9% whereas, it was only 0.73 in Naïve Bayes, 0.75 in Linear Regression, 0.76 in Random Forest and 0.76 and 0.77 in GBM and XGBoost models. Moreover, the performance of proposed Ensemble-Bagging diabetes prediction approach was also analyzed in terms of AUC, whose value came out to be 0.896. In addition to this, the proposed approach yields better results in terms of other parameters like precision, recall and Fscore with an average value of 0.99%. These stats confirm the supremacy and dominance of our ensemble-bagging diabetes prediction approach and hence can be used in real world for practical applications.

REFERENCES

[1] Roshi Saxena, Sanjay Kumar Sharma, Manali Gupta, G. C. Sampada, "A Comprehensive Review of Various Diabetic Prediction Models: A Literature Survey", Journal of Healthcare Engineering, vol. 2022, Article ID 8100697, 15 pages, 2022.

[2] R. Williams, S. Karuranga, B. Malanda et al., "Global and regional estimates and projections of diabetes-related health expenditure: results from the international diabetes federation diabetes atlas," Diabetes Research and Clinical Practice, vol. 162, Article ID 108072, 2020.

[3] American Diabetes Association, "Diagnosis and classification of diabetes mellitus," Diabetes Care, vol. 37, no. Supplement 1, pp. S81-S90, 2014.

[4] International Diabetes Federation. Diabetes. Brussels: International Diabetes Federation; 2019.

[5] Ormazabal, Valeska, et al. "Association between insulin resistance and the development of cardiovascular disease." Cardiovascular diabetology 17.1 (2018): 1-14.

[6] Jaspinder Kaur, "A Comprehensive Review on Metabolic Syndrome", Cardiology Research and Practice, vol. 2014, Article ID 943162, 21 pages, 2014.

[7] Diabetes, the Lancet. "COVID-19 and diabetes: a co-conspiracy?." The Lancet. Diabetes & Endocrinology 8.10 (2020): 801.

[8] Katsarou, A., Gudbjörnsdottir, S., Rawshani, A., Dabelea, D., Bonifacio, E., Anderson, B. J., ... & Lernmark, Å. (2017). Type 1 diabetes mellitus. Nature reviews Disease primers, 3(1), 1-17.

[9] Gnanadass, Iswaria. "Prediction of gestational diabetes by machine learning algorithms." IEEE Potentials 39.6 (2020): 32-37.

[10] Umair Muneer Butt, Sukumar Letchmunan, Mubashir Ali, Fadratul Hafinaz Hassan, Anees Baqir, Hafiz Husnain Raza Sherazi, "Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications", Journal of Healthcare Engineering, vol. 2021, Article ID 9930985, 17 pages, 2021.

[11] Zou, Quan, et al. "Predicting diabetes mellitus with machine learning techniques." Frontiers in genetics (2018): 515.

[12] Kumari, Saloni, Deepika Kumar, and Mamta Mittal. "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier." International Journal of Cognitive Computing in Engineering 2 (2021): 40-46.

[13] Satu, Shahriare, Syeda Tanjila Atik, and Mohammad Ali Moni. "A novel hybrid machine learning model to predict diabetes mellitus." Proceedings of International Joint Conference on Computational Intelligence. Springer, Singapore, 2020.

[14] Islam, Md Merajul, et al. "Automated detection and classification of diabetes disease based on Bangladesh demographic and health survey data, 2011 using machine learning approach." Diabetes & Metabolic Syndrome: Clinical Research & Reviews 14.3 (2020): 217-219.

[15] Devi, R. Delshi Howsalya, Anita Bai, and N. Nagarajan. "A novel hybrid approach for diagnosing diabetes mellitus using farthest first and support vector machine algorithms." Obesity Medicine 17 (2020): 100152.

[16] Ateeq, Karamath, and Gopinath Ganapathy. "The novel hybrid Modified Particle Swarm Optimization-Neural Network (MPSO-NN) Algorithm for classifying the Diabetes." International Journal of Computational Intelligence Research 13.4 (2017): 595-614.

[17] Ramezani, Rohollah, Mansoureh Maadi, and Seyedeh Malihe Khatami. "A novel hybrid intelligent system with missing value imputation for diabetes diagnosis." Alexandria engineering journal 57.3 (2018): 1883-1891.

[18] P. Roy et al., "A novel hybrid approach for severity assessment of Diabetic Retinopathy in colour fundus images," 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), 2017, pp. 1078-1082, doi: 10.1109/ISBI.2017.7950703.

 DOI : 10.35291/2454-9150.2024.0052 © 2024, IJREAM All Rights Reserved.