IJETIR

ISSN (Online) 2583-0554

International Journal of Emerging Technologies and Innovative Research (IJETIR)

Volume 4, Issue 6, June 2024

Impact Factor: 5.731

Diabetes Prediction using Machine Learning Algorithms

J. D. Jeevaraja[1], P. Kavitha[2], S. Kamalakkannan[3]

PG Student, Department of Computer Applications[1]

Assistant Professor, Department of Computer Applications[2]

Associate Professor, Department of Information Technology[3]

Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai, Tamil Nadu, India

jdjeevaraj@gmail.com, pkavikamal@gmail.com, kannan.scs@velsuniv.ac.in

Abstract: *Diabetic retinopathy (DR) is a disease that damages retinal blood vessels and leads to blindness. Usually, colored fundus shots are used to diagnose this irreversible disease. The manual analysis (by clinicians) of the mentioned images is monotonous and error-prone. Hence, various computer vision hands-on engineering techniques are applied to predict the occurrences of the DR and its stages automatically. However, these methods are computationally expensive and lack to extract highly nonlinear features and, hence, fail to classify DR's different stages effectively. This project focuses on classifying the DR's different stages with the lowest possible learnable parameters to speed up the training and model convergence. The VGG-16, spatial pyramid pooling layer (SPP) is stacked to make a highly nonlinear scale-invariant deep model called the VGG-16 model. The proposed VGG-16 model can process a DR image at any scale due to the SPP layer's virtue. Moreover, the stacking adds extra nonlinearity to the model and tends to better classification. The experimental results show that the proposed model performs better in terms of accuracy, computational resource utilization compared to state-of-the-art methods*

Keywords: Diabetic retinopathy

I. INTRODUCTION

Diabetes is one of the fastest-growing diseases in recent times. Recently, about 382 million people worldwide have diabetes mellitus (DM), and the future projected value of diseases is 592 million by 2025. Based on the causes and symptoms produced, there are two types of DM called type-I and type-II. Moreover, both types of DM affect vital body organs in humans, including the eye. A significant eye illness that has been reported due to DM is known as diabetic retinopathy (DR). Symptoms of DR showed that it produces mutilation of blood vessels in the retina. Among those 382 million of the population of the world, 34.6% are reported to be affected by DR. Apart from DR, proliferative diabetic retinopathy (PDR) and diabetic macular edema (DME) are reported in 7.0% and 6.8% population respectively. By inferring the global situation from these figures, it is estimated that the number of DR cases will rise from 126.6 million to 191.1 million by 2030. However, reports have shown that blindness caused by DR in the Khyber Pakhtunkhwa (KP) province of Pakistan is 4% of the total population. Moreover, the common reason for DR in the province reported to be the BD type-I DR: Risk Factors Awareness and Presentation, Pakistan, 2017). In KP, 30% of the population reported having DM, with 1.6% of the total patients having type-II DM.

II. LITERATURE REVIEW

The analysis of related work gives results on various healthcare datasets, where analysis and predictions were carried out using various methods and techniques. Various prediction models have been developed and implemented by various researchers using variants of data mining techniques, machine learning algorithms or also combination of these techniques. Dr Saravana Kumar N M, Eswari, Sampath P and Lavanya S (2015) implemented a system using Hadoop and Map Reduce technique for analysis of Diabetic data. This system predicts type of diabetes and also risks

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 4, Issue 6, June 2024
Impact Factor: 5.731

associated with it. The system is Hadoop based and is economical for any healthcare organization.[4] Aiswarya Iyer (2015) used classification technique to study hidden patterns in diabetes dataset. Naïve Bayes and Decision Trees were used in this model. Comparison was made for performance of both algorithms and effectiveness of both algorithms was shown as a result. The task of choosing a machine learning algorithm includes feature matching of the data to be learned based on existing approaches. Taxonomy of machine learning algorithms is discussed below-
Machine learning has numerous algorithms which are classified into three categories: Supervised learning, Unsupervised learning, Semi-supervised learning.

The Supervised Learning/Predictive Models
Supervised learning algorithms are used to construct predictive models. A predictive model predicts missing value using other values present in the dataset. Supervised learning algorithm has a set of input data and also a set of output, and builds a model to make realistic predictions for the response to new dataset. Supervised learning includes Decision Tree, Bayesian Method, Artificial Neural Network, Instance based learning, Ensemble Method. These are booming techniques in Machine learning.[3]

Unsupervised Learning / Descriptive Models
Descriptive models are developed using unsupervised learning method. In this model we have known set of inputs but output is unknown. Unsupervised learning is mostly used on transactional data. This method includes clustering algorithms like k-Means clustering and k-Medians clustering. [3]

Semi-supervised Learning
Semi Supervised learning method uses both labeled and unlabeled data on training dataset. Classification, regression techniques come under Semi Supervised Learning. Logistic Regression, Linear Regression are examples of regression techniques.

III. METHROLOGY SECTION

Data Collection
Data collection is the process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes. The main purpose of data collection is to gather information in a measured and systematic manner to ensure accuracy and facilitate data analysis. Since the data collected is meant to provide content for data analysis, the information gathered must be of the highest quality for it to be of value. Here for this project we are using CSV file for the process. The dataset is collected from the kaggle web site.

Data Visualization
EDA is applied to investigate the data and summarize the key insights. It will give you the basic understanding of your data, it's distribution, null values and much more. You can either explore data using graphs or through some python functions. There will be two type of analysis. In the graphical approach, you will be using plots such as scatter, box, bar, density and correlation plots.

Data Pre-Processing
EDA is applied to investigate the data and summarize the key insights. It will give you the basic understanding of your data, it's distribution, null values and much more. You can either explore data using graphs or through some python functions. There will be two type of analysis. In the graphical approach, you will be using plots such as scatter, box, bar, density and correlation plots.

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 4, Issue 6, June 2024
Impact Factor: 5.731

Prediction

By using this machine learning algorithms like SVM, naïve Bayes, random forest, logistic regression k-neighbors classifier and decision tree. By using all these algorithm we are predicting the values like accuracy, classification report and confusion matrix.

We are finding the best model by using the accuracy values

SYSTEM ARCHITECTURE

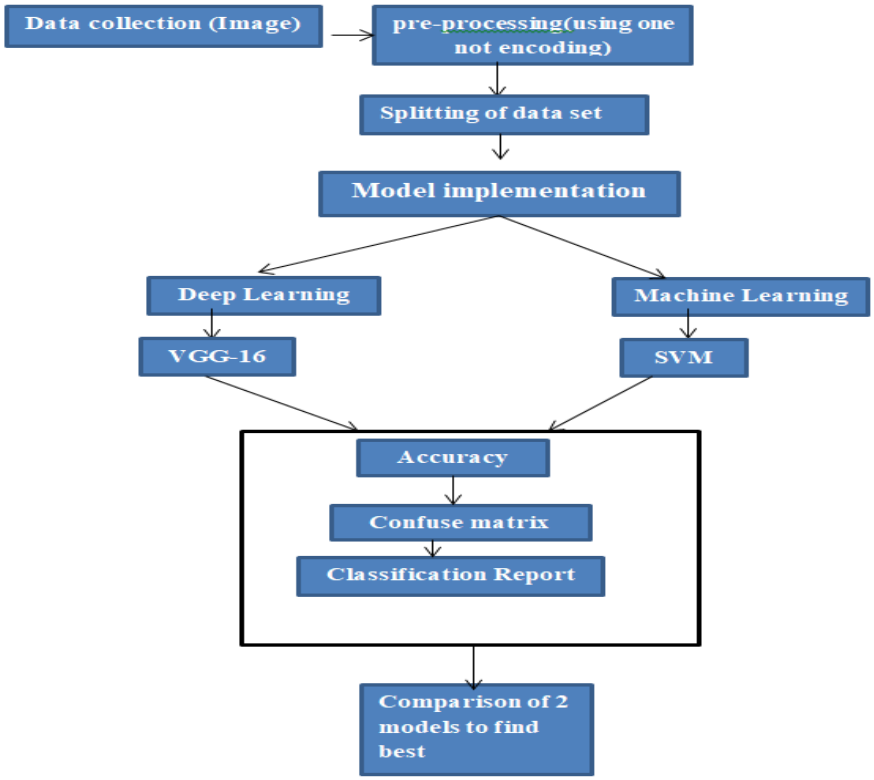

Fig1. System Architecture

IV. MODEL IMPLEMENTATION

In this Paper, we implemented some of the machine learning models. We compared the accuracy metrics with all results of the algorithms. The algorithms like random forest, SVM, naïve bayes, decision tree, logistic regression and k neighbors classifier.

Logistic Regression:

Logistic regression is basically a supervised classification algorithm. In a classification problem, the target variable(or output), y, can take only discrete values for a given set of features(or inputs), X.Contrary to popular belief, logistic regression is a regression model. The model builds a regression model to predict the probability that a given data entry belongs to the category numbered as "1". Just like Linear regression assumes that the data follows a linear function, Logistic regression models the data using the sigmoid function.

Decision tree:

Decision Tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 4, Issue 6, June 2024
Impact Factor: 5.731

Random Forest:

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. It performs better results for classification problems.

K-Nearest Neighbor

K-Nearest Neighbors is one of the most basic yet essential classification algorithms in Machine Learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining and intrusion detection. It is widely disposable in real-life scenarios since it is non-parametric, meaning, it does not make any underlying assumptions about the distribution of data.

SVM:

Support vector machines are a set of supervised learning methods used for classification, regression, and outliers detection. All of these are common tasks in machine learning.You can use them to detect cancerous cells based on millions of images or you can use them to predict future driving routes with a well-fitted regression model.

V. IMPLEMENTATION RESULTS

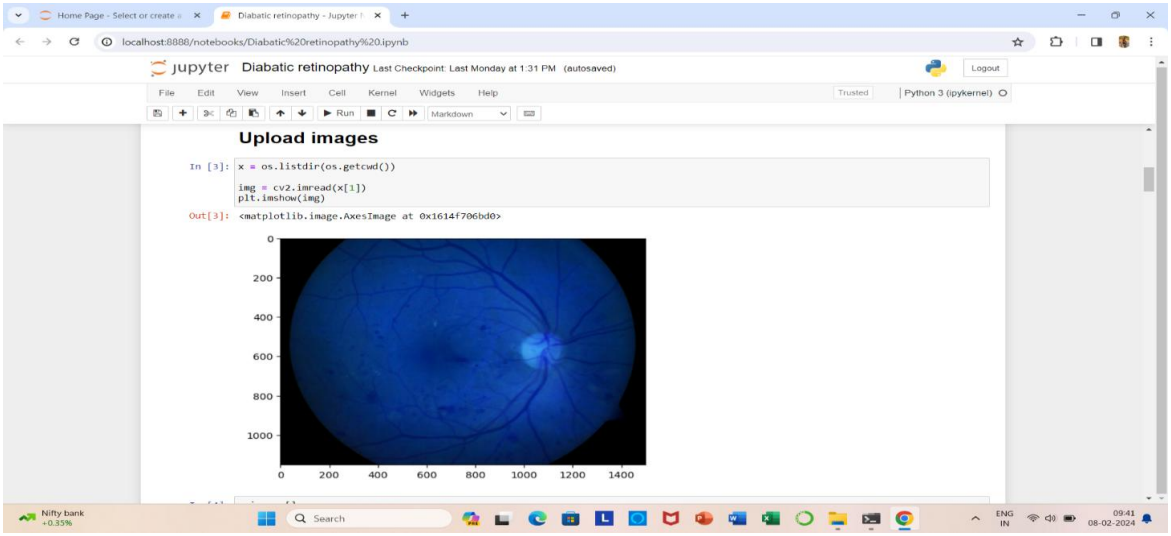

Figure.2 Dataset load

IJETIR

ISSN (Online) 2583-0554

International Journal of Emerging Technologies and Innovative Research (IJETIR)

Volume 4, Issue 6, June 2024

Impact Factor: 5.731

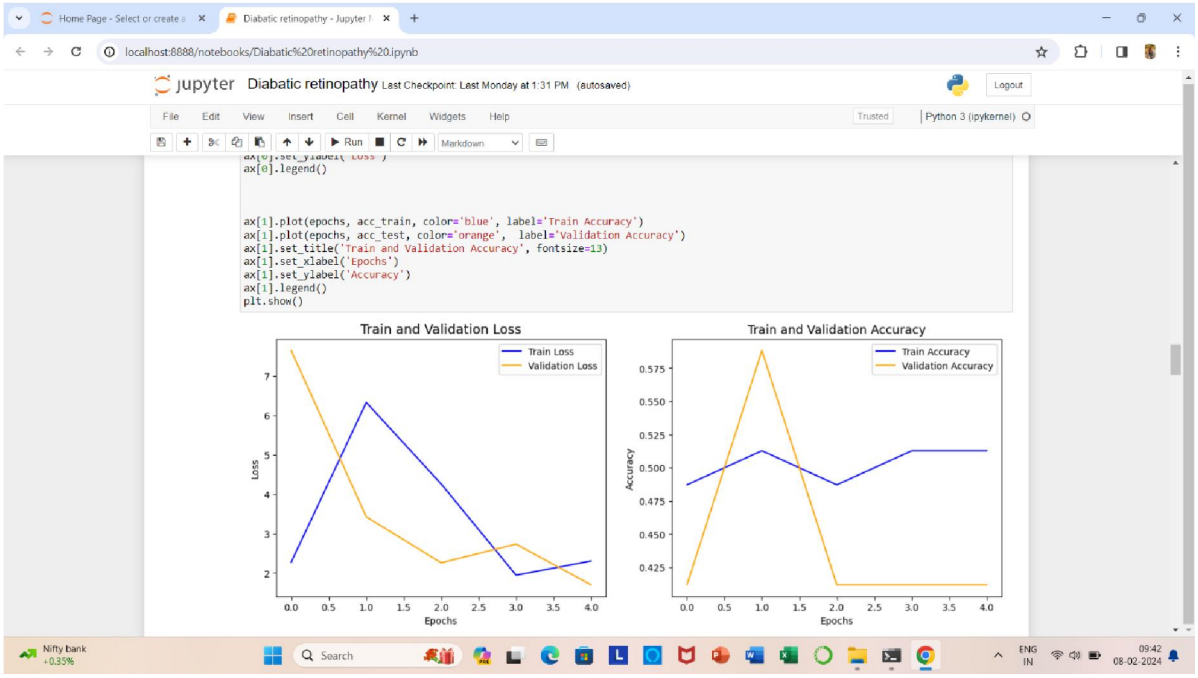

Figure. 3 Train and loss validation output

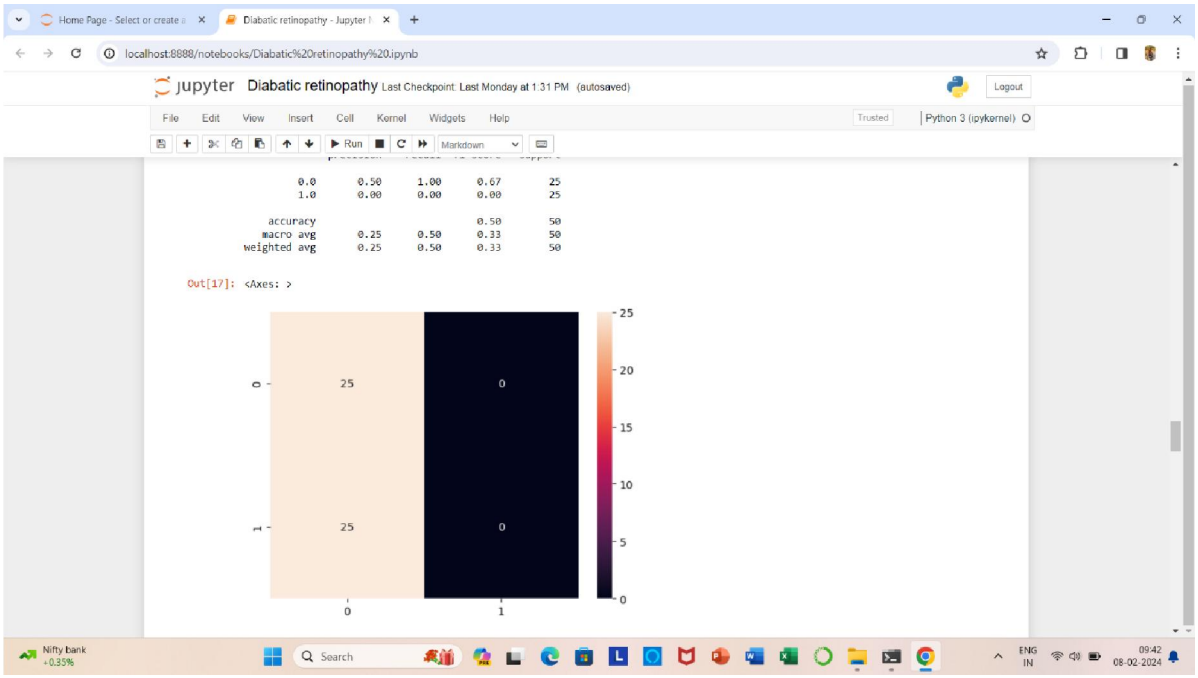

Figure.4 Heatmap

IJETIR
ISSN (Online) 2583-0554
International Journal of Emerging Technologies and Innovative Research (IJETIR)
Volume 4, Issue 6, June 2024
Impact Factor: 5.731

Table 1. Comparison between accuracies of PIMA Diabetes Dataset and Diabetes Dataset

Algorithms	Accuracy with PIMA Dataset	Accuracy with Diabetes Dataset used in this paper
Logistic Regression	76%	96%
Gradient Boost Classifier	77%	93%
LDA	77%	94%
AdaBoost Classifier	77%	93%
Extra Trees Classifier	76%	91%
Gaussian NB	67%	93%
Bagging	75%	90%
Random Forest	72%	91%
Decision Tree	74%	86%
Perceptron	67%	76%
SVC	68%	60%
KNN	72%	90%

Table 2. Pipelining Results

Algorithms	Accuracy
AdaBoost Classifier	98.8%
Gradient Boost Classifier	98.1%
Random Forest Classifier	98.1%
Logistic Regression	97.5%
Extra Trees Classifier	96.3%

VI. CONCLUSION

This project is an extension of our work in which we proposed the deep learning-based ensemble approach for diabetic retinopathy detection. The major drawback of the ensemble model is the number of learnable parameters. In this paper, we brought architectural changes in existing CNN to enhance the efficiency and accuracy of classification of the DR's stages in colour fundus images and reduce the number of learnable parameters. We used imbalanced versions of the Kaggle dataset to validate the performance measures of the proposed model. The results depict that the proposed model is low in computation and better than other state-ofthe-art ensemble and non-ensemble methods. In the future, we plan to bring some other productive changes in the existing model's architecture and some pre-processing techniques and discuss how these changes affect the working of a model on the classification of DR's stages, especially the early ones.

REFERENCES

[1] S. Jan, I. Ahmad, S. Karim, Z. Hussain, M. Rehman, and A. A. Shah, ''Status of diabetic retinopathy and its presentation patterns in diabetics at ophthalomogy clinics,'' J. Postgraduate Med. Inst. (Peshawar-Pakistan), vol. 32, no. 1, 2018.

[2] L. Math and R. Fatima, ''Adaptive machine learning classification for diabetic retinopathy,'' Multimedia Tools Appl., vol. 80, pp. 5173–5186, Oct. 2020.

[3] A. He, T. Li, N. Li, K. Wang, and H. Fu, ‘‘CABNet: Category attention block for imbalanced diabetic retinopathy grading,’’ IEEE Trans. Med. Imag., vol. 40, no. 1, pp. 143–153, Jan. 2021.
[4] L. Andersen and P. Andersson, ‘‘Deep learning approach for diabetic retinopathy grading with transfer learning,’’ Tech. Rep., 2020.
[5] N. Congdon, Y. Zheng, and M. He, ‘‘The worldwide epidemic of diabetic retinopathy,’’ Indian J. Ophthalmol., vol. 60, no. 5, p. 428, 2012.
[6] W. R. Memon, B. Lal, and A. A. Sahto, ‘‘Diabetic retinopathy,’’ Prof. Med. J., vol. 24, no. 2, pp. 234–238, 2017.
[7] R. Sarki, K. Ahmed, H. Wang, and Y. Zhang, ‘‘Automatic detection of diabetic eye disease through deep learning using fundus images: A survey,’’ IEEE Access, vol. 8, pp. 151133–151149, 2020.
[8] R. E. Putra, H. Tjandrasa, and N. Suciati, ‘‘Severity classification of non-proliferative diabetic retinopathy using convolutional support vector machine,’’ Int. J. Intell. Eng. Syst., vol. 13, no. 4, pp. 156–170, Aug. 2020.
[9] S. Qummar, F. G. Khan, S. Shah, A. Khan, S. Shamshirband, Z. U. Rehman, I. Ahmed Khan, and W. Jadoon, ‘‘A deep learning ensemble approach for diabetic retinopathy detection,’’ IEEE Access, vol. 7, pp. 150530–150539, 2019.
[10] R. Ghosh, K. Ghosh, and S. Maitra, ‘‘Automatic detection and classification of diabetic retinopathy stages using CNN,’’ in Proc. 4th Int. Conf. Signal Process. Integr. Netw. (SPIN), Feb. 2017, pp. 550–554.