iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** XII **Month of publication:** December 2023

DOI: https://doi.org/10.22214/ijraset.2023.57364

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue XII Dec 2023- Available at www.ijraset.com

Binary Classification of Diabetes in Pima Indian Dataset: A Deep Learning Perspective

Asst. Prof. Moumita Dey[1], Akhand Pratap Singh[2], Shiwanshi[3], Waquif Akhtar[4], Shiva Rao[5]
Durgapur Institute of Advanced Technology & Management, India

Abstract: ***In this study, we develop into the application of deep learning methodologies for diabetes prediction utilizing the Pima Indian dataset. Employing Keras with Theano as the backend, we establish a binary classification model to effectively forecast the presence or absence of diabetes in individuals. Our research aims to enhance the precision and reliability of diabetes diagnosis, ultimately contributing to improved healthcare decision-making. Our investigation leverages Keras, a high-level neural networks API, in conjunction with Theano, to conduct binary classification on the Pima Indian diabetes dataset. Our study provides valuable insights into the field of medical data analysis, showcasing the effectiveness of deep learning techniques in advancing diagnostic tools for proactive healthcare management. Diabetes mellitus, a prevalent chronic disease globally, necessitates the development of a system for early type 2 diabetes mellitus (T2DM) diagnosis. Multiple machine learning and data mining techniques, including ANN, SVM, KNN, decision trees, and Extreme Learning Machines, have emerged and been employed as aids in diabetes detection. Consequently, we introduce Deep Learning, a subfield of machine learning, which can effectively handle smaller datasets through efficient data processing techniques. This paper presents an in-depth review of Diabetic Retinopathy, covering its features, causes, various ML models, DL models, challenges, comparisons, and future directions for early DR detection. Diabetes mellitus is a global health concern with a rapidly increasing prevalence. In this context, machine learning technologies prove invaluable for early disease identification and diagnosis. The focus of this study is to identify the most effective ML algorithm for diabetes prediction.***

Keywords: Diabetes Prediction, Binary Classification, Deep Learning, Python(Keras Theano), Data Mining algorithms, Neural Networks, Pima Indian Dataset, Data Analysis.

I. INTRODUCTION

Diabetes should be also called as silent killer. Now a day's diabetes spreading in all over the world and the effect of diabetes is showing a major condition in human beings directly or indirectly. Due to diabetes various severely affected and ill functioned which may cause heart stroke, blindness, brain dead, kidney failure e.t.c. More than 422 million people are suffering from diabetes as per world health organization index (WHO) [1]. Deep learning now days plays a crucial rolefor detection and prediction of medical diseases at an early stage of safe human life.Type 1 (T1DM), type 2 (T2DM) and gestational diabetes (GDM) are the three types of diabetes. In this, type 1 is a condition where the pancreas stoppedsecreting the insulin so that external injection of it is necessary to maintain the insulin level in the body [2]. In type 2 diabetes,insulin is not utilized properly by the body which needs proper diet, exercises and in some cases they take tablets orally to compensate the insulin level. Type 3 diabetes comprises of high level ofblood glucose during pregnancy and disappear after it [2]. From all these types, T2DM is more prominent among the people. Earlier prediction of this type of diabetes will help the people to get rid of it when proper diet and healthy life style were followed [3].This research work represents comprehensive studies done on the PIMA datasets using data mining algorithms like DT, NB, ANN, and DL [4].The comparison of algorithms is represented in a logical and well-organized manner from which DL provides more effective and prominent results. DL is a technology that self-learns from data and is used effectively for predicting diabetes nowadays. A DL network is a technique that uses ANN properties in which neuronsare interconnected to each other with lot of representation layers. These machinelearning methods tend to improve the accuracy of the available methods. But DLand ANN provide the best results as they are more reliable, robust and accurate in terms of prediction of the disease [5].

II. PROJECT OBJECTIVES

These project objectives are as follows:

1) *Development of a Deep Learning Model:* The primary objective of this project is to design and implement a robust deep learning model specifically tailored for the binary classification of diabetes using the Pima IndianDataset. The model will leverage advanced neural network architectures, with a focus on achieving high accuracy and robust performance in the challenging task of diabetes diagnosis.

©IJRASET: All Rights are Reserved

2) *Feature Analysis:* Conduct a comprehensive analysis of health-related features within the Pima Indian dataset, exploring their individual and collective impact on diabetes prediction.
3) *Data Preprocessing:* Implement robust data preprocessing techniques to handle missing values, normalize features, and ensure the dataset's suitabilityfor training a binary classification model.
4) *Model Robustness:* Develop a robust neural network model using Keras and Theano, considering the unique characteristics of the Pima Indian dataset to enhance the accuracy and reliability of diabetes predictions.
5) *Optimization:* Optimize the model parameters, including activation func- tions, layer architecture, and hyper parameters, to achieve an optimal balance between model complexity and generalization.
6) *Performance Metrics:* Evaluate the model's performance using a rangeof metrics, including accuracy, precision, recall, and F1 score, to provide a comprehensive assessment of its predictive capabilities.
7) *Theano Integration:* Explore and leverage the capabilities of Theano as the computational backend for Keras, ensuring efficient execution of neural network computations [6].
8) *Interpretability:* Strive for an interpretable model that allows for a deeper understanding of the relationships between input features and the likelihood of diabetes, facilitating insights for healthcare practitioners.
9) *Generalization:* Assess the model's ability to generalize to new, unseen data, ensuring its practical utility beyond the training dataset and increasingits applicability in real-world scenarios.
10) *Documentation:* Provide clear and comprehensive documentation of the entire process, including data preprocessing steps, model architecture, train- ing procedures, and evaluation metrics, to facilitate understanding and replication by other researchers or practitioners.
11) *Contribution to Healthcare:* Contribute valuable insights to the health- care domain by demonstrating the effectiveness of machine learning in predicting diabetes within a specific population, potentially informing pre- ventive measures and personalized healthcare strategies [7].

III. LITERATURE REVIEW

The list is not comprehensive but represents a selection of key sources that informedour understanding of the topic.

1) Swapna G [5] and others made a study that Machine learning practice has proven useful and efficient to construct a prediction model for diabetes using HRV signals in the DL approach. The author was motivated through the deaths caused by diabetes every year in the world which necessitated avoiding the complication ofthe disease. The author developed a new predictive model using a convolutional neural network (CNN), long short-term memory (LSTM) and an ensemble model for detecting compound chronological characteristics of the input HRV data. Then SVM has been applied to those detected characteristics for classifying the data. The proposed system can be useful for healthcare officials and clinicians to analyze diabetes using ECG signals.
2) Nesreen Samer El Jerjawi and Samy S. Abu Naser [8] proposed a prediction model for diabetes using ANN (Artificial Neural Network) that can be very useful for healthcare official and practitioners. The author was motivated by the highly dangerous complication of the disease. He developed an ANN model for minimizing the error function in the training. So the average error function calculated was 0.01% and accuracy attained through ANN was 87.3%.
3) Dey, Samrat Kumar, Ashraf Hossain, and Md Mahbubur Rahman [9]The authors created an architecture that can predict if a patient is diabetic or not. They build a web application based on the higher precision of a strong learning algorithm. They utilized the Pima Indian benchmark dataset, which can predict diabetes on the basis of diagnoses. A precision rate of 82.35% they achieve by using Artificial Neural Network.
4) Faruque, Md Faisal, and Iqbal H. Sarker [10] the authors apply four well-known machine learning algorithms to adult population data to help predict the presence of diabetes. These algorithms are Naive Bayes, C4.5 Decision Tree, K NearestNeighbor, and SVM. They found that their experimental results demonstrated that the C4.5 algorithm made 72% more accurate predictions than other machine learning algorithms. As a result, higher rates were obtained than those acquired by others, splitting the dataset into training and testing sets. Then using ML algorithms (LR, NB, and KNN), the system shows that (LR) produced the best results, the measures: recall, precision, and F1-measure are applied to reach thiscomparison.
5) In the year 2017 Carrera et al [11] suggested a computer assisted methodology for the detection of diabetic retinopathy, based on the digital signals processing of retinal images. The major aspiration of this proposed approach is the categorization of the position of non-proliferative diabetic retinopathy at any of the retinal image. The main advantage of

this approach is that it is robust in nature but precision and accuracy are needed to improve for the documented application matter. Diabetes retinopathy is chronic and has become the leading lifestyle ailment. A long run of this disease can cause heart failure, kidney failure; improper functioning of stomach, prolonged elevated blood sugar levels and many more.

6) Zou et al [12] had implemented three classifiers namely neural network, random forest, and decision tree. Authors had carried out comparison of classifiers on Luzhou and PIMA dataset and analyzed it.The study highlights that random forest method was superior to decision tree and neural network methods. The dimensionality of dataset has been reduced with the use of principal componentanalysis (PCA) and minimum redundancy maximum relevance (MRMR). Based onthe experimental analysis, authors had concluded that the accuracy of classification by utilizing PCA was inferior to the accuracy obtained using all the features.

7) Alalwan [13] have used various data mining algorithms like SVM, multilayer perceptron (MLP), na¨ıve bayes, random forest, logistic regression and J.48 and self-organizing maps (SOM) to develop predictive model for diabetes on the PIMA dataset. Authors had suggested self-organizing maps to improve upon the accuracy of the prediction.

8) Kopitar et al [14] compared various machine learning based predictive models, such as, random forest, regularized generalized linear model, extreme gradient boosting, light gradient boosting and commonly employed regression approaches to predict T2DM.

IV. METHODOLOGY

The project consists of seven chapters, and the organization of the project is as follows: Our study methodology comprises several stages as presents in Fig. 1, first, we collect the Pima Indian Diabetes dataset (PIDD). Second, we prepro- cess the (PIDD) dataset in order to construct the prediction model. Third, we employ a variety of Deep learning algorithms to the training (PIDD) dataset. Finally, a test dataset is used to evaluate the approaches' performance in order to choose the best classifier for diabetes prediction. We will discuss these stages below.

1) *Data Collection:* The primary dataset for this project is the Pima Indian Dataset, a well-established repository of health-related information for Pima Indian individuals. The dataset includes fea- tures such as glucose levels, BMI, age, and pregnancy history, with binary labels indicating the presence or absence of diabetes.The PIMA Indian Dataset, established as part of the long-term co- hort study initiated by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) since 1965, is particu- larly significant due to the heightened risk of diabetes within this population.

2) *Data pre-processing:* Prior to model training, a thorough preprocessing phase will be conducted. This includes handling missing values, normalizing numerical features, and encoding categorical variables [5]. Given the imbalanced nature of the dataset, techniques such as over sampling or under sampling will be explored to address class imbalances and ensure the model's robustness.

3) *Deep Learning Model Architecture:* The chosen deep learning architecture for this project is a Convolutional Neural Network (CNN). CNNs have demonstrated effectiveness in capturing spatial dependencies within data, making them well-suited for tasks such as image classification. The architecture will comprise multiple convolutional and pooling layers followed by fully connected layers[15].

4) *Naive Bayes (NB):* Naive Bayes classifier is constructed from a family of machine learning classifier based on Bayes theorem. NB classifier develops a probabilistic model which assumes inter dependency of features to each other for prediction of outcome and also used for PIDD [4]

5) *Decision Tree Classifier:* Decision tree classifier is one of the best classifier to classify binary data. Decision tree classifier follows divide and conquer approach in which the dataset is represented in a tree structure .The structure is formed using the concept of information gain and depending on the importance of information the classifier performs the classification which is adopted by many researchers for PIDD [4]

6) *Hyperparameter Tuning:* To optimize model performance, hy- perparameter tuning will be conducted. This includes fine-tuning parameters such as learning rates, dropout rates, and the number of hidden units in the layers. A systematic approach, possibly using grid search or random search, will be employed to explore the hyperparameter space efficiently [16].

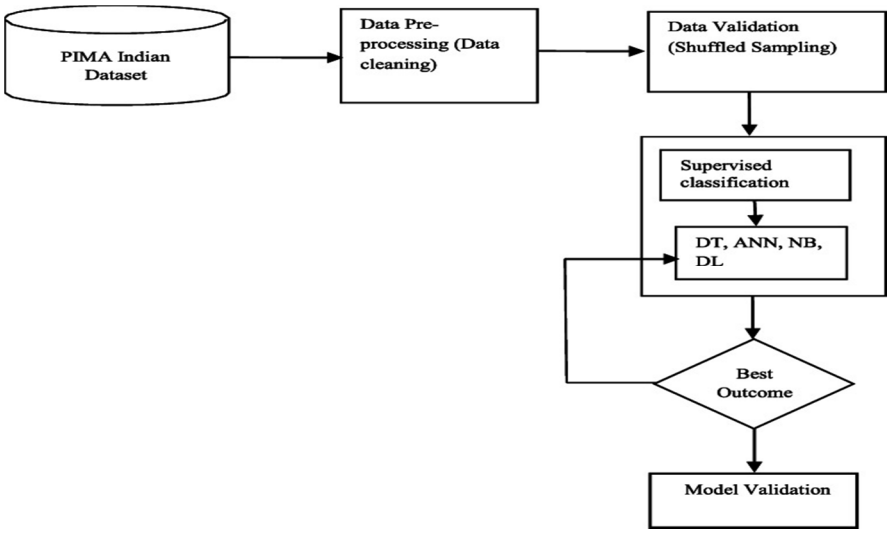


Figure 1: Flow Chart

Sr. no.	Selected Attributes from PIMA Indian dataset	Description of selected attributes	Range
1.	Pregnancy	Number of times a participant is pregnant	0–17
2.	Glucose	Plasma glucose concentration a 2 h in an oral glucose tolerance test	0–199
3.	Diastolic Blood pressure	It consists of Diastolic blood pressure (when blood exerts into arteries between heart) (mm Hg)	0–122
4.	Skin Thickness	Triceps skinfold thickness (mm).It concluded by the collagen content	0–99
5.	Serum Insulin	2-Hour serum insulin (mu U/ml)	0–846
6.	BMI	Body mass index (weight in kg/(height in m)^2)	0–67.1
7.	Diabetes pedigree Function	An appealing attributed used in diabetes prognosis	0.078–2.42
8.	Age	Age of participants	21–81
9.	Outcome	Diabetes class variable, Yes represent the patient is diabetic and no represent patient is not diabetic	Yes/No

Figure 2: Description of Pima India Dataset

V. TRAINING AND VALIDATION

1) *Dataset Splitting:* The dataset will undergo a process of division into three distinct sets: training, validation, and test sets. This standard practiceinvolves allocating 80% of the dataset for training, 10% for validation, and the remaining 10% for testing. Each set serves a specific purpose in the model development and evaluation process.
2) *Model Training:* The Convolutional Neural Network (CNN) model will undergo training on the designated training set employing the backpropagation algorithm and stochastic gradient descent. During model training, the iterative process involves cycling through the dataset multiple times, commonly referred to as epochs. Throughout these epochs, the model's weights are dynamically adjusted to minimize the loss function, enhancing its ability to accurately capture patterns within the data [15].
3) *Model Evaluation:* The assessment of the model's effectiveness will involve the utilization of established classification metrics, ensuring a thorough evaluation. These metrics encompass accuracy, precision, recall, F1 score, and the area under the Receiver Operating Characteristic (ROC) curve [16].
4) *Result and Analysis:* The final step involves interpreting the results, analyzing the model's performance, and drawing conclusions based on theevaluation metrics. The insights gained will be compared with existing literature and may inform recommendations for future research or potential real-world applications.

VI. CONCLUSION

1) This paper aimed to implement a prediction model for the risk measurement of diabetes. As discussed earlier, a large part of the human population is in the hold of diabetes disease. If remains untreated, then it will create a huge risk for the world. Therefore In our proposed research, we have put into practice diverse classifiers on the PIMA dataset and proved that data mining and machine learning algorithm can reduce the risk factors and improve the outcome in terms of efficiency and accuracy.

2) Employing convolutional neural networks (CNNs), the objective was to advance the precision and transparency of diabetes diagnosis. Leveraging a diverse dataset, complemented by strategic sampling methods and robust model evaluation metrics, the study aimed to offer a thorough analysis. DL is considered as the most efficient and promising for analyzing diabetes withan accuracy rate of 98.07%.

VII. FUTURE WORK

Write the future scopes of the project.

In the future, we intend to develop a robust system in the form of an app or a website that can use the proposed DL algorithm to help healthcare specialists in the early detection of diabetes.

1) Tailoring the model to specific populations or demographics, beyond the Pima Indian population, could broaden its applicability. This involves training the model on diverse datasets to ensure its effectiveness across different patient groups.

2) Collaborate with healthcare professionals to integrate the model into clinical decision-making processes, ensuring its practical utility and relevance in real-world healthcare scenarios [17].

3) Develop a user-friendly interface or application that allows healthcare practi- tioners to easily input patient data and obtain predictions, fostering practical implementation in clinical settings.

4) Longitudinal Analysis: Conducting a longitudinal analysis by incorporating time-series data could offer insights into the progression of diabetes. Track- ing changes in patient data over time may contribute to a more dynamic understanding of the disease [3]. Adapting the model for real-time diag- nostics in a clinical setting is a crucial next step. Implementing the model within a healthcare infrastructure, considering factors like data security and processing speed, could facilitate practical applications.

REFERENCES

[1] Ratna Patil, Sharvari Tamane, Shitalkumar Adhar Rawandale, and Kanishk Patil. A modified mayfly-svm approach for early detection of type 2 diabetes mellitus. Int. J. Electr. Comput. Eng, 12(1):524–533, 2022.

[2] American Diabetes Association. 2. classification and diagnosis of diabetes: standards of medical care in diabetes—2018. Diabetes care, 41(Supplement 1):S13–S27, 2018.

[3] Ratna Nitin Patil. A survey paper on evolving techniques for the prediction oftype 2 diabetes. International Journal of Computer Science and Information Security (IJCSIS), 14(10), 2016.

[4] Tom Michael Mitchell. Key ideas in machine learning. Machine learning, pages 1–11, 2017.

[5] G Swapna, R Vinayakumar, and KP Soman. Diabetes detection using deep learning algorithms. ICT express, 4(4):243–246, 2018.

[6] Usman Ahmad, Hong Song, Awais Bilal, Shahid Mahmood, Mamoun Alazab, Alireza Jolfaei, Asad Ullah, and Uzair Saeed. A novel deep learning model to secure internet of things in healthcare. Machine intelligence and big dataanalytics for cybersecurity applications, pages 341–353, 2021.

[7] Zeeshan Ahmed, Khalid Mohamed, Saman Zeeshan, and XinQi Dong. Artifi- cial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. Database, 2020:baaa010, 2020.

[8] Nesreen Samer El Jerjawi and Samy S Abu-Naser. Diabetes prediction using artificial neural network. 2018.

[9] Samrat Kumar Dey, Ashraf Hossain, and Md Mahbubur Rahman. Imple- mentation of a web application to predict diabetes disease: an approach using machine learning algorithm. In 2018 21st international conference ofcomputer and information technology (ICCIT), pages 1–5. IEEE, 2018.

[10] Iqbal H Sarker, Md Faisal Faruque, Hamed Alqahtani, and Asra Kalim. K-nearest neighbor learning based diabetes mellitus prediction and analysis for ehealth services. EAI Endorsed Transactions on Scalable Information Systems, 7(26):e4–e4, 2020.

[11] Enrique V Carrera, Andrés González, and Ricardo Carrera. Automated detection of diabetic retinopathy using svm. In 2017 IEEE XXIV international conference on electronics, electrical engineering and computing (INTERCON),pages 1–4. IEEE, 2017.

[12] Quan Zou, Kaiyang Qu, Yamei Luo, Dehui Yin, Ying Ju, and Hua Tang. Predicting diabetes mellitus with machine learning techniques. Frontiers in genetics, 9:515, 2018.

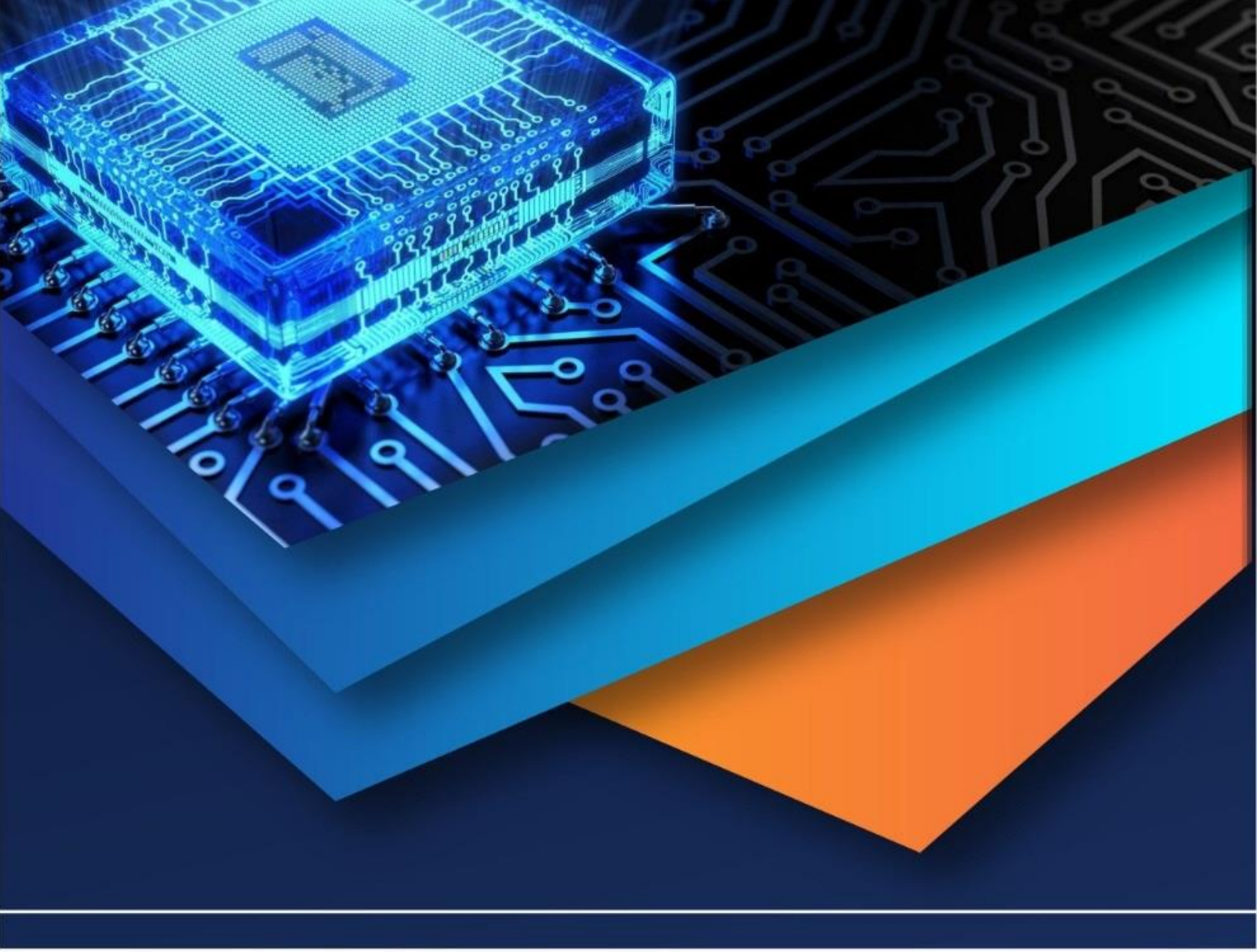
[13] Sinan Adnan Unknown Diwan Alalwan. Diabetic analytics: proposed concep-tual data mining approaches in type 2 diabetes dataset. Indonesian Journal of Electrical Engineering and Computer Science, 14(1), 2019.

[14] Leon Kopitar, Primoz Kocbek, Leona Cilar, Aziz Sheikh, and Gregor Stiglic. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. Scientific reports, 10(1):11981, 2020.

[15] Teja Kattenborn, Jens Leitloff, Felix Schiefer, and Stefan Hinz. Review on convolutional neural networks (cnn) in vegetation remote sensing. ISPRSjournal of photogrammetry and remote sensing, 173:24–49, 2021.

[16] Patrick Schratz, Jannes Muenchow, Eugenia Iturritxa, Jakob Richter, and Alexander Brenning. Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. Ecological Modelling, 406:109–120, 2019.

[17] R Vaishali, R Sasikala, S Ramasubbareddy, S Remya, and Sravani Nalluri. Genetic algorithm based feature selection and moe fuzzy classification algo-rithm on pima indians diabetes dataset. In 2017 international conference oncomputing networking and informatics (ICCNI), pages 1–5. IEEE, 2017.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)