iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** III **Month of publication:** March 2022

DOI: https://doi.org/10.22214/ijraset.2022.40780

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 10 Issue III Mar 2022- Available at www.ijraset.com

Diabesta Faction Security in Machine Learning

Unnati K Patel[1], Feon Jaison[2]
[1]*MCA Scholar,* [2]*Assistant Professor, School of CS & IT, Dept. of MCA Jain (Deemed-to-be) University, Bangalore*

Abstract: ***Diabetes is a disease that could impact high levels of glucose in the human body. It should not be ignored until proper treatment is administered with proper precautions, sometimes due to irresponsibility assumed by patients, leading to heart problems, kidney problems, blood pressure, lesions eyes and may affect other organs of the human body. If precautions are taken from the beginning, it can be cured. In the proposed work machine learning classification and defined techniques on a dataset to predict diabetes is being done. Such as Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM) and Random Forest (RF). The end result shows that Random Forest achieved higher accuracy than other machine learning techniques. The importance of privacy in deep learning applications is directly related to the emergence of distributed and multi-party models.***

Keywords: Machine learning, prediction, security, differential privacy, random forest, logistic regression, super vector machine.

I. INTRODUCTION

Existing system is based on algorithms, where Machine Learning (ML) / Deep Learning (DL) systems have transformed various industries such as manufacturing, transportation, and governance. In recent years, DL has brought excellence in various fields, such as computer vision, text analysis and language processing, etc. Due to the wide use of ML/DL algorithms in various fields (e.g. social networks); this technology has become an integral part of our daily lives. The ML/DL algorithms are now beginning to affect healthcare as well, an area traditionally immune to large-scale technological disruption. Diabetes caused by obesity or high blood sugar, etc. It impacts the hormone insulin, causing abnormal metabolism in crabs and improving blood sugar. Diabetes occurs when the body does not produce enough insulin. According to the World Health Organization (WHO), an estimated 422 million people have diabetes, particularly in low-income or inactive countries. And that could rise to 490 billion by 2030. Yet the prevalence of diabetes is found in several countries like Canada, China and India etc. India's population is now over 100 million, so the actual number of diabetics in India is 40 million. It is the first cause of death in the world. Early prediction of diseases such as diabetes can be controlled and save lives. To achieve this, this thesis examines the prediction of diabetes using many attributes related to diabetic diseases. For this purpose, we use kaggle's diabetes dataset; apply different classification techniques and machine learning set to predict diabetes. Machine learning is a method of training computers or machines in a targeted manner. Different machine learning techniques provide effective results in knowledge gathering by creating different models and classification sets from the collected data set. This collected data can be useful in predicting diabetes. Various machine learning techniques can make predictions, but choosing the best technique is difficult. Therefore, for this purpose, we apply common methods of collection and classification to datasets for forecasting. The problem statement indicates that physicians rely on general knowledge for treatment. When general knowledge is lacking, the studies are summarized after examining a number of cases. However, this process takes time, while machine learning can identify patterns early. To use machine learning, large amounts of data are required. Depending on the disease, only very limited amounts of data are available. Also, the number of samples that show no disease is very high compared to the number of samples that actually show disease.

Stakeholder Category	Examples
Knowledge Experts	Clinical Experts, e.g., radiologist and dermatologists etc.
	Health information and technology experts
	ML researchers. E.g., ML engineers and data scientists, etc.
Decision Makers	Institutional leadership
	Hospital administrators
	State and Federal Governments
	Regulatory agencies
Users	Physicians
	Nurses
	Laboratory technicians
	Patients
	Care takers, e.g., friends and family

Table I: Interdisciplinary teams with different stakeholders from multiple domains

Adnan Qayyum et al. [2] have suggested that these LM/DL techniques have great potential for clinical applications (e.g. triage of diabetic patients to radiologically detect pneumonia, etc.), but the limited acceptance in real clinical settings shows that these methods are not yet optimal and not ready for clinical use. In a recent study, Wains et al. provided a roadmap for safe, meaningful, and responsible machine learning in healthcare, arguing that implementing machine learning in any field should be done by an interdisciplinary team, bringing together different stakeholders from several disciplines, IH Knowledge can include experts, decision-makers and users.

Table I provide examples of an interdisciplinary team involving multiple stakeholders in the healthcare ecosystem.

II. LITERATURE REVIEW

Random Forest algorithmic rule will perform early prediction of illness for a patient with the next accuracy in machine learning technique. The projected model offers the simplest results for diabetic prediction and therefore the result showed that the prediction analysis was effective and efficient.

Nonso Nnamoko et al. [14] Conferred predicting diabetes onset: an ensemble supervised learning approach they used five wide classifiers for the ensembles and a meta-classifier is employed to mixture their outputs.

The outcome shows how with Kaggle dataset the prediction analysis for accuracy for all DL Algorithm and security is predicted. Polygenic disease Prediction mistreatment Machine Learning Techniques aims to predict diabetes via four completely different supervised machine learning ways including:

1) Random Forest
2) Decision Tree
3) Super Vector Machine
4) Logistic Regression

The prediction using data processing assembles intelligent diabetics' prediction system that provides analysis of diabetes patient database. To secure the information we are able to even work with AWS S3 Bucket Console to convey security to the access management list. Even with security frameworks like Pysyft, we are able to use to secure data in 3 completely different ways:

a) Secured Multi-Party Computations
b) Federated Learning
c) Differential Privacy

PySyft combines federated learning, secured multiple-party computations and differential privacy during a single programming model integrated into different deep learning frameworks akin to PyTorch, Keras or Tensor Flow. Secured Multi-Party Computations provides a protocol wherever no individual can see the opposite parties data whereas distributing the information across multi parties. It permits the data scientists and analysts to work out in camera on the distributed data while not exposing it. DL models are generally trained beneath the principle of empirical risk reduction (ERM) that provides sensible learning bounds and guarantees if its assumptions are satisfied.

Authors	Goal	Method	ML Model(s)	Medical Dataset(s)
David et al. [109]	Privacy	Commodity based cryptography.	Hyperplane decision and Naïve Bayes classifiers.	N/A
Zhu et al.[60]		Polynomial aggregation and multi-party random making.	SVM with nonlinear kernel.	N/A
Jagielski et al.[110]		Proposed an algorithm names as TRIM to defend poisoning attacks.	Linear Regression	Anticoagulant drug Warfarin
Liu et al. [111]		XMPP server and serval mobile device	Proposed a DL framework	Human Activity Recognition
Malathi et al.[112]		Paillier homomorphic encryption.	NaveBayesia, SVM, Neural Network, and FKNN-CBR	Indian Liver Patient
Takabi et al. [113]		Homomorphic Encryption.	DNN	15 datasets from UCI Repository
Kim et al.[114]	Security	Homomorphic encryption based secure logistic regression.	Logistic Regression	Five Medical datasets having binary classes.

Table II: Summary of the state of the art data security and privacy preserving methods in healthcare settings.

A summary of articles centered on the subject of "secure and privacy-preserving DL for healthcare" is conferred in Table II and varied approaches for secure, private, and strong DL [2]. For instance, one amongst the foremost and robust assumptions is that each the training and testing datasets are derived from the same domain (i.e., data distributions).

However, this assumption isn't valid in practice, and models trained beneath such an assumption fail to generalize to alternative domains. In contrast, the life-critical nature of clinical applications demands a sleek and safe operation of ML/DL techniques. Polygenic disease Prediction is changing into the realm of interest for researchers so as to coach the program to spot the patient are diabetic or not by applying correct classifier on the dataset. Thus a system is needed as polygenic disease Prediction is vital space in computers, to handle the problems identified.

III. PROPOSED METHODOLOGY

The aim of the article is to examine a model to more accurately predict diabetes. We experimented with different classification algorithms and ensembles to predict diabetes. Next, we briefly analyze the phase.

A. Dataset Description

The data is collected from the UCI repository called the Kaggle dataset. The record has many attributes.

Dataset Description Attributes:

1) Pregnancy
2) Glucose
3) Blood Pressure
4) Skin thickness
5) Insulin
6) BMI (Body Mass Index)
7) Diabetes Pedigree Function
8) Age

The ninth attribute is the class variable of each data point. This class variable returns the result 0 and 1 for diabetics, indicating positive or negative for diabetics.

B. Data Preprocessing

Data pre-processing is the most important process. Most health-related data contains missing values and other contaminants that can lead to data invalidity. In order to improve the quality and effectiveness achieved after the extraction process, data pre-processing is performed. To effectively apply machine learning techniques to the data set, this process is essential for an accurate result and a successful prediction. For the Kaggle diabetes dataset, we need to do the pre-processing in two steps:

1) *Remove Missing Values:* Remove all instances that have zero (0) as the null value. It is not possible to have a null value. Therefore this instance is deleted. By removing irrelevant features/instances we create a subset of features. This process is called feature subset selection, which reduces data diametrically and helps you work faster.
2) *Splitting of Data:* After cleaning the data, the data is normalized when training and testing the model. When the data is split, we train the algorithm on the training dataset and reserve the test data. This training process will generate the training model based on logic and algorithms and feature values on training data. Basically, the goal of normalization is to bring all attributes to the same scale.

C. Apply Machine Learning

When the data is ready, we apply the machine learning technique. We tend to use completely different classification and ensemble techniques to predict diabetes. The methods used in the Kaggle Diabetes data set. The main goal is to apply machine learning techniques to analyze the performance of these methods and find their accuracy, and also to discover the responsible/important feature that plays an important role in the prediction. The techniques are as follows:

1) *Decision Tree:* The decision tree is a basic classification method. It is a supervised learning method. The decision tree is used when the response variable is categorical. The decision tree has a model based on a tree-like structure that describes the classification process based on the input function. Input variables can be of any type, such as B. graphic, text, discrete, continuous, etc.

Steps for Decision Tree Algorithm-

a) Create a tree using nodes as the input feature.
b) Select the feature to predict the output of the input feature that has the most information gain.
c) The highest information gain is calculated for each attribute in each node of the tree.
d) Repeat step 2 to form a subtree with the function not used in the previous node.

2) *Logistic Regression:* Logistic regression is also a supervised learning classification algorithm. Its accustomed estimate the likelihood of a binary response supported one or a lot of predictors. They can be continuous or discrete. Logistic regression used when we want to classify or distinguish some data items into categories.

It classify the data in binary form means only in 0 and 1 which refer case to classify patient that is positive or negative for diabetes.
Main aim of logistic regression is to best match that is liable for describing the link between target and predictor variable. Logistic regression is a based on Linear regression model. Logistic regression model uses sigmoid operate to predict likelihood of positive and negative class.
Sigmoid function:
P = 1/1+e – (a+bx)
Here P = probability, a and b = parameter of Model.

Ensembling- Ensembling is a machine learning technique. It means using several learning algorithms together for certain activities. It provides higher prediction than the other individual model that's why it's used. The main cause of errors is the distortion and variance of the noise. Ensemble methods help reduce or minimize these errors. There are two popular assembly methods such as Bagging, Boosting, Ada Boosting, Gradient Boosting, Voting, Averaging, etc. Here in this article, we have used bagging ensemble (random forest) and gradient methods to predict diabetes.

3) *Random Forest:* It is style of ensemble learning technique and additionally used for classification and regression tasks. The accuracy it provides is greater than compared to different models. This method will simply handle massive datasets. Random Forest is developed by Leo Bremen. It is popular ensemble Learning Method. Random Forest Improve decision tree performance by reducing variance. It operates by constructing a large number of decision trees at training time and outputs the class that is the mode of the classes or classification or mean prediction (regression) of the individual trees.

Algorithm-

a) The first step is to select the R features from the total features m where R<<M.
b) Among the R features, the node using the best split point.
c) Split the node into sub nodes using the best split.
d) Repeat a to c steps until l number of nodes has been reached.
e) Built random forest by repeating steps a to d for a number of times to create n number of trees.
f) The random forest finds the best split using the GinIndex Cost Function which is given by:

- The first step is to need the take a glance at choices and use the foundations of each indiscriminately created decision tree to predict the result and stores the anticipated outcome at intervals the target place.
- Secondly, calculate the votes for each predicted target and ultimately, admit the high voted predicted target as a result of the ultimate prediction from the random forest formula. Some of the options of Random Forest does correct predictions result for a spread of applications are offered.

4) *Support Vector Machine:* Support Vector Machine additionally referred to as SVM could be a supervised machine learning algorithm. It is one of the popular classification techniques in machine learning. SVM produces a hyperplane that separate 2 categories. It will produce a hyperplane or set of hyperplane in high dimensional space. This hyper plane may be used for classification or regression also. SVM differentiates instances in specific classes and might also classify the entities that don't seem to be supported by information. Separation is done by through hyperplane performs the separation to the closest training point of any class.

Algorithm-

a) Select the hyper plane which divides the class better.
b) To find the better hyper plane you have to calculate the distance between the planes and the data which is called Margin.
c) If the distance between the classes is low then the chance of miss conception is high and vice versa. So we need to.
d) Select the class which has the high margin.
e) Margin = distance to positive point + Distance to negative point.

IV. MODEL BUILDING

This is most important phase which includes model building for prediction of diabetes. In this we have implemented various machine learning algorithms which are discussed above for diabetes prediction.

Procedure of Proposed Methodology-

1) Step 1: Import required libraries, Import diabetes dataset.
2) Step 2: Pre-process data to remove missing data.
3) Step 3: Perform percentage split of 80% to divide dataset as Training set and 20% to Test set.
4) Step 4: Select the machine learning algorithm i.e. Support Vector Machine, Decision Tree, Logistic regression and Random Forest.
5) Step 5: Build the classifier model for the mentioned machine learning algorithm based on training set.
6) Step 6: Test the Classifier model for the mentioned machine learning algorithm based on test set.
7) Step 7: Perform Comparison Evaluation of the experimental performance results obtained for each classifier.
8) Step 8: After analyzing based on various measures conclude the best performing algorithm.
9) Once model is analyzed and concluded we move with Security implementation part.
10) Step 9: Upload the csv file data to AWS S3 Cloud storage, create the Bucket and assign the suitable Access Control List (ACL) privileges.
11) Step 10: Once privilege is assigned we need to copy the secret key ID, and access key ID.
12) Step 11: Then run the code.
13) Step 12: Even Security framework is implemented like PySyft, which is like how data is being protected from intruders.
14) Step 13: It even gives the accuracy result for train dataset and test dataset.

V. DESIGN

In designing and developing the architecture for the diabetes management system, the clinical requirements and system design analysis were based on discussions with Kaggle dataset.
The following functionalities mentioned are:

1) Schedule and remind diabetics to take their medication and check their blood glucose levels.
2) Recommend healthy meals to diabetics to keep their blood glucose levels under control.
3) Encourage and track diabetics' activities.

A. Providing a Visual Interface for Diabetics

The system architecture for the diabetes management system shown in Figure I below is the conceptual model that defines the structure, behavioural interactions, and multiple views of the system that support the development of the system. Presents the formal descriptions of the graphically captured systems supporting the argument and the developed sub-modules, as well as the data flows between the developed modules.

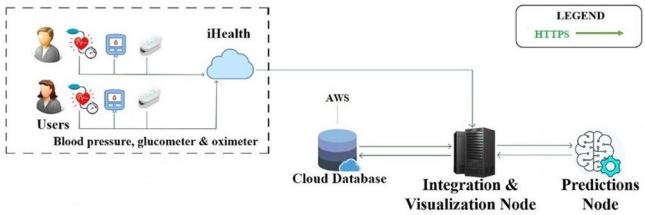


Fig. I: System Architecture

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue III Mar 2022- Available at www.ijraset.com

The following diagram represents the data flow of the ML task training process, including supervised learning, unsupervised learning and active learning, shown in the following figure II. When the training data is labelled, the training process is represented as a supervised training process, while otherwise referred to as an unsupervised training process. In contrast, when both labelled and unlabeled data are used for training processing, this training process is said to be semi-supervised.

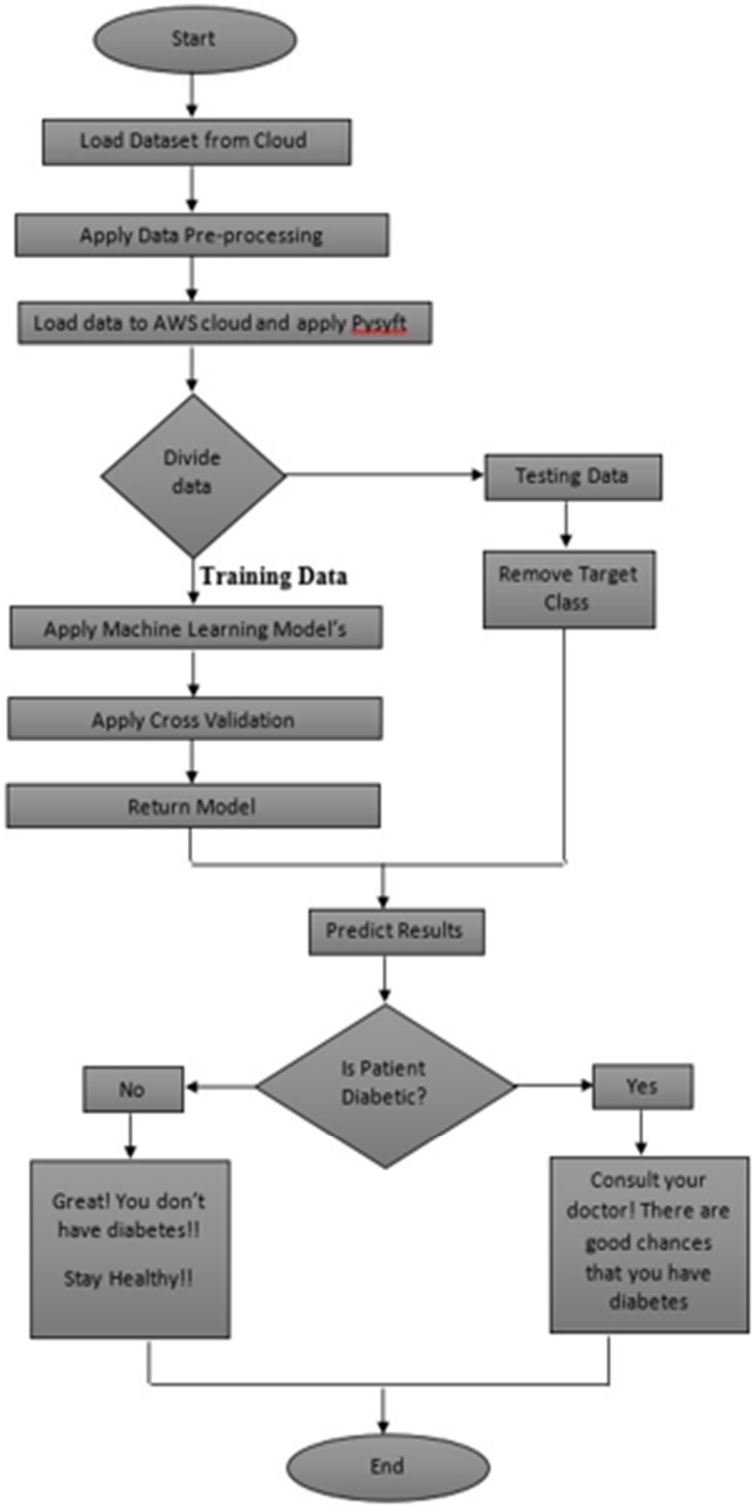


Fig. II: Dataflow Process

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue III Mar 2022- Available at www.ijraset.com

VI. RESULT AND CONCLUSION

The main goals of the project were to develop an algorithm that will be used to identify answers to questions submitted by users. Various steps were taken in this work. The proposed approach uses different sorting and set methods and is implemented with Python. These methods are standard machine learning methods used to achieve the best data accuracy compared to other. In general, we used the best machine learning techniques to predict and achieve high performance accuracy, as shown in Figure III. Here we present the feature that played an important role in the prediction for the random forest algorithm.

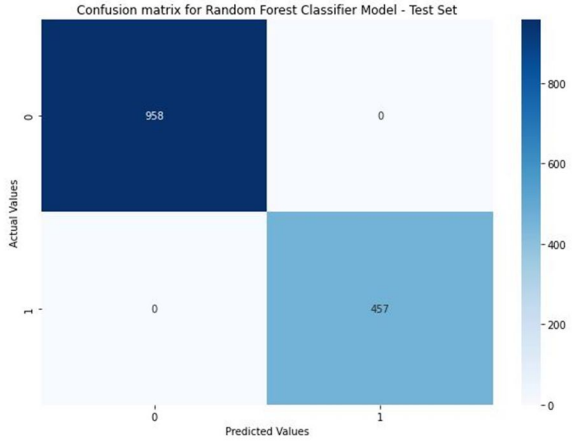


Fig. III: Test Prediction in Confusion matrix with regards to Random Forest (K-Fold Technique)

The sum of the importance of each characteristic playing an important role in diabetes was plotted, with the x-axis representing the importance of the actual values and the y-axis representing the names of the predicted values. 98% accuracy was achieved in the PySyft framework as shown in Figure IV below.

```
test(model,test_loader)

Test set: Average loss: -16.9124, Accuracy: 9782/10000 (98%)
```

Fig. IV: Test Prediction in Pysyft

VII. FUTURE ENHACEMENT

Several useful features can be added to this project in the future such as:

Proposed system uses “Random Forest algorithm” to find the diabetes disease, in data science we have many algorithms for classification such as Naive Bayes, KNN, ID3 etc. in future we can add more algorithms to find outputs and algorithms can be compared to find the efficient algorithm. We can add visitor query module, where visitors can post queries to administrator and admin can send reply to those queries.

We can add treatment module, where doctors upload treatment details for patients and patient can view those treatment details. And they are different security framework like Crypten, Syfer Text etc. can be implemented.

REFERENCES

[1] Author: Thenappan, S.; Valan Rajkumar, M.; Manoharan, P. S. “Predicting Diabetes Mellitus Using Modified Support Vector Machine with Cloud Security”, vol. (1-11), IETE Journal Research, 2020

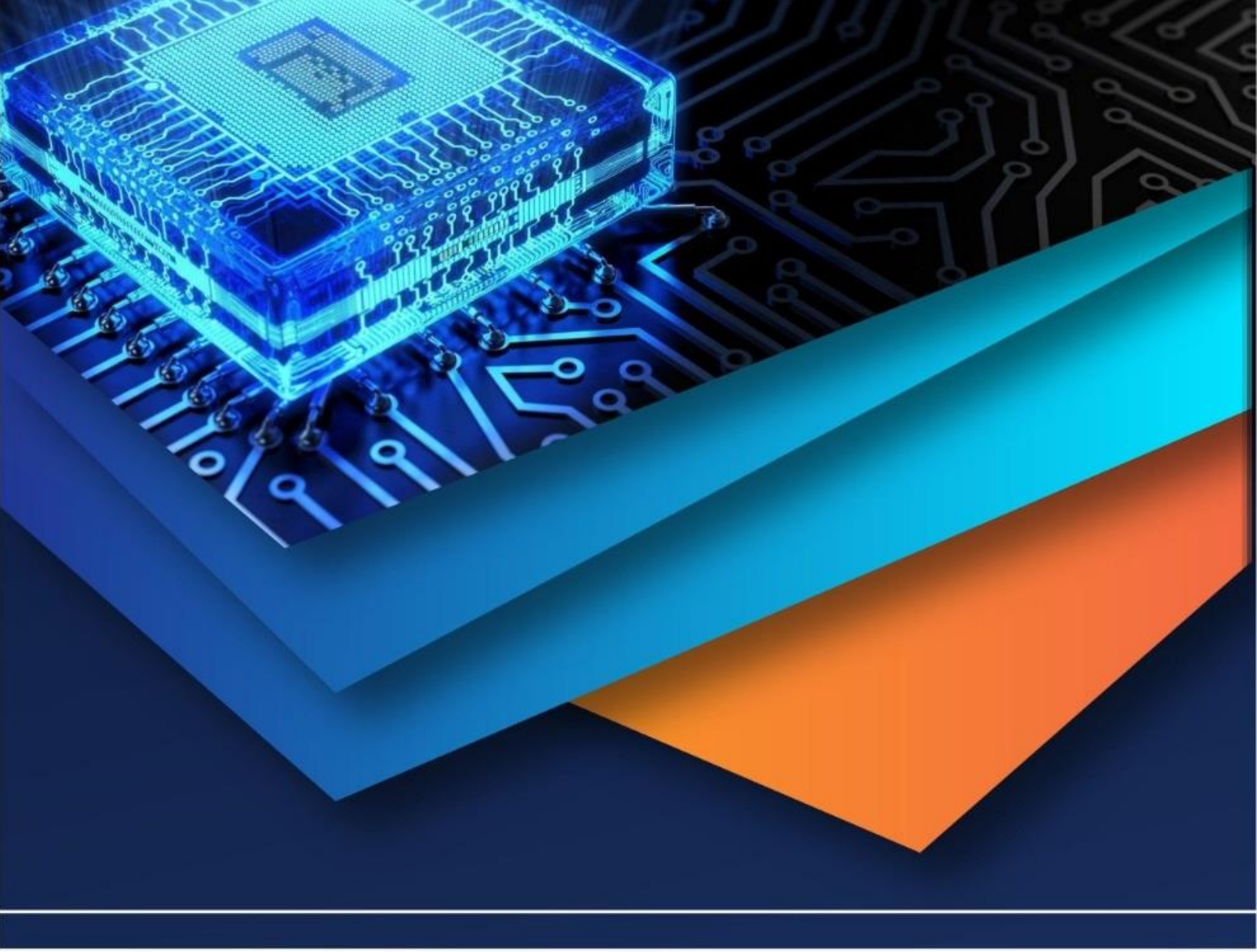
[2] Author: Adnan Qayyum , Junaid Qadir , Muhammad Bilal , and Ala Al-Fuqaha Information Technology University (ITU), Punjab Lahore, Pakistan University of the West England (UWE), Bristol, United Kingdom Hamad Bin Khalifa University (HBKU), Doha, Qatar “Secure and Robust Machine Learning for Healthcare: A Survey”, arxiv, 2020.

[3] Author: March, Jinying Chen, Author Orcid Image ; John Lalor, Author Orcid Image ; Weisong Liu, Author Orcid Image ; Emily Druhl Author Orcid Image ; Edgard Granillo, Author Orcid Image ; Varsha G Vimalananda, Author Orcid Image ; Hong Yu “Detecting Hypoglycemia Incidents Reported in Patients’ Secure Messages: Using Cost-Sensitive Learning and Oversampling to Reduce Data Imbalance”, vol. (21), JMR Publication, 2019.

[4] Author: Kumar, P. Suresh; Pranavi, S. “Performance analysis of machine learning algorithms on diabetes dataset using big data analytics”, vol. (508-513), IEEE, 2017.

[5] Author: Gadekallu, T. R., Khare, N., Bhattacharya, S. “Early Detection of Diabetic Retinopathy using PCA-Firefly based Deep Learning Model. Electronics”, vol. (9[2]), Research gate, 2019.

[6] Author: Singh, A., Dhillon, A., Kumar, N., Hossain, M. S., Muhammad, G., & Kumar, M. "eDiaPredict: An Ensemble-based Framework for Diabetes Prediction. ACM Transactions on Multimedia Computing, Communications, and Applications", vol. (17 [2s]), ACM Journals, 2021.
[7] Author: Kiratsata, H. J., & Panchal, M. "A Comparative Analysis of Machine Learning Models developed from Homomorphic Encryption based RSA and Paillier algorithm", ICICCS, 2021.
[8] Author: Mujumdar, Aishwarya; Vaidehi, V "Diabetes Prediction using Machine Learning Algorithms", vol. (165), Science Direct, 2019.
[9] Author: Sarmah, Simanta Shekhar "An efficient IoT based patient monitoring and heart disease prediction system using Deep learning modified neural network", vol. (1-1), IEEE, 2020.
[10] Author: Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh "Analyzing Feature Importances for Diabetes Prediction using Machine Learning", vol. (924-928), IEEE, 2018.
[11] Author: K.VijiyaKumar, B.Lavanya, I.Nirmala, S.Sofia Caroline "Random Forest Algorithm for the Prediction of Diabetes", vol. (1-5), IEEE, 2019.
[12] Author: Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus", IEEE, 2019.
[13] Author: Tejas N. Joshi, Prof. Pramila M. Chawan "Diabetes Prediction Using Machine Learning Techniques", vol. (8[1]), IJERA, 2018.
[14] Author: Nonso Nnamoko, Abir Hussain, David England. "Predicting Diabetes Onset: an Ensemble Supervised Learning Approach", vol. (1-7), IEEE, 2018.
[15] Author: Deeraj Shetty, Kishor Rit, Sohail Shaikh, Nikita Patil "Diabetes Disease Prediction Using Data Mining", ICIIECS, 2017.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY