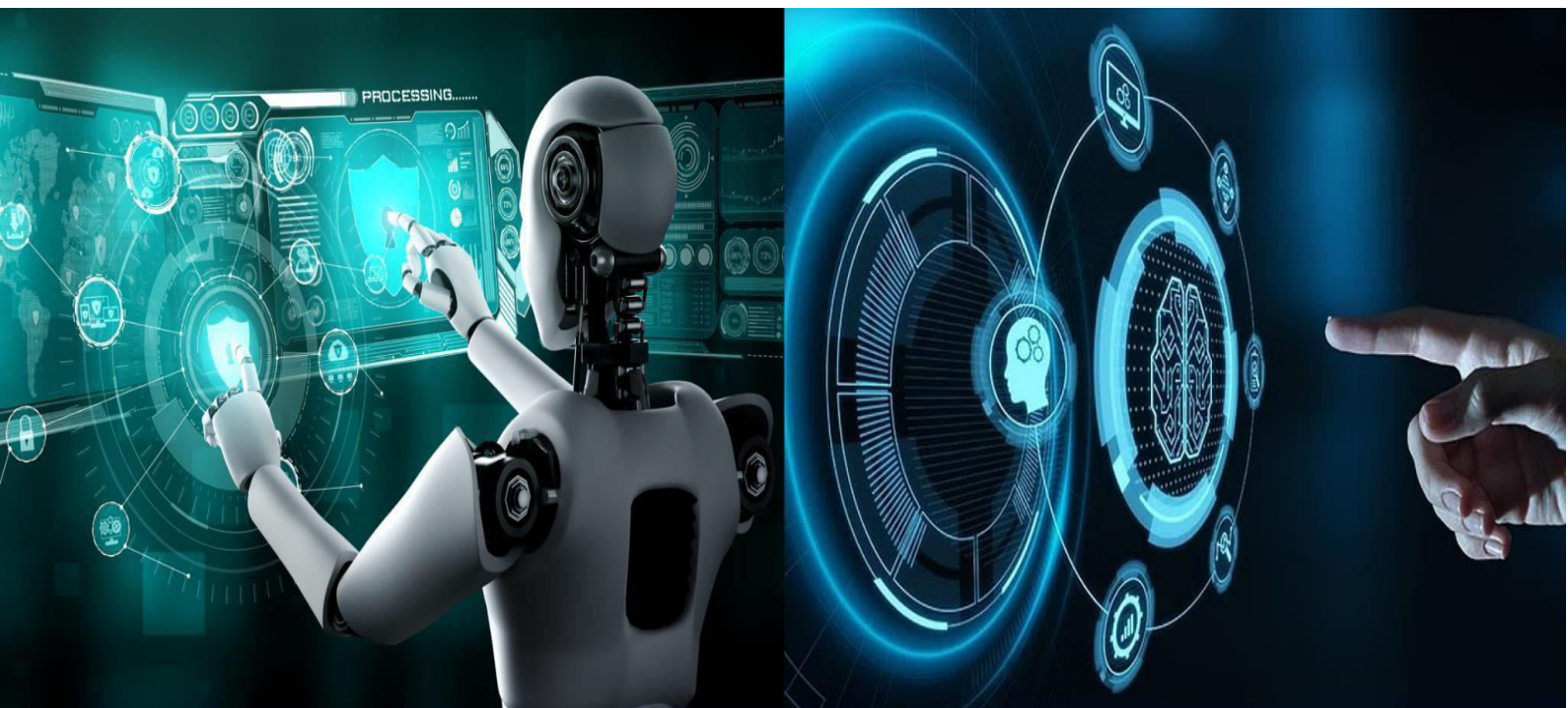




# International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)





## International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

# Machine Learning-Based Approach for Diabetes

Nishutha M S<sup>1</sup>, Chinnaswamy C N<sup>2</sup>, Rampur Srinath<sup>3</sup>

Department of ISE, National Institute of Engineering, Mysuru, Karnataka, India<sup>1</sup>

Associate Professor, Department of ISE, National Institute of Engineering, Mysuru, Karnataka, India<sup>2</sup>

Associate Professor, Department of ISE, National Institute of Engineering, Mysuru, Karnataka, India<sup>3</sup>

**ABSTRACT:** Diabetes which is the on-going illness that may impacts large number of people over the world ,and sadly the large number of people living with it and also it keeps on increasing day by day. Basically the diabetes impacts how the body may controls the blood sugar, commonly resulting in high blood sugar. One of the ways diabetes can first show itself is through simple daily changes constant thirst, unusual hunger or frequent bathroom visits. Poorly managed diabetes can take a heavy toll on health, causing complications such as a kidney damage, heart disease and vision problems and in worst cases even amputation. The body usually relies on insulin a hormone that works like a key to let glucose move from the blood into the cells where its converted into energy. In diabetes the process breaks down the pancreases may not produce enough insulin ,or the body may no longer respond to it properly. There are several types of diabetes. Type1 usually shows up earlier in life type is more common in adults and gestational diabetes can develop while a woman in pregnant. Manage diabetes isn't always easy but technology is bringing new hope. Today Artificial Intelligence can scan through large sets of health data and pick up on patterns that might escape doctors, leading to quicker diagnosis and more personalized care.

**KEYWORDS:** Diabetes Prediction, SVM, ML

## I. INTRODUCTION

Diabetes is one of the most common long-term health conditions in the world, and it can affect people of any age –even children and teenagers. To understand how the body normally keeps blood sugar in balance. Many of the foods we eat everyday like bread, rice, cereals, and even daily are full of carbohydrates. After we eat them, our body quickly turns those carbs into glucose which is the main source of energy that keeps us going. After we eat, glucose from the food enters our bloodstream. Some of it is used right away to fuel vital organs like the brain, while the rest is stored in the liver or sent to other cells, ready to be used later when the body needs more energy. On a larger scale, tackling the impact of diabetes requires strong public health efforts like spreading awareness in communities, promoting early check-ups and supporting people to lead healthier lifestyles. Although diabetes lasts a lifetime it doesn't mean people can't live well. With the right information, tools, and support, people living with diabetes can manage their health and keep enjoying an active, fulfilling life . Type 1 diabetes happens when the body's own immune system mistakenly attacks the pancreas and destroys the cells that make insulin. Without enough insulin, the body struggles to manage blood sugar levels, which to make daily care and attention so important. There is a type 2 diabetes which is about to 90% of the cases, which can be develops and produces the enough amount of insulin for resistant of the body. Genetics plays an important role of a person lifestyle about the not having the proper diet, no exercise for the body and also the over weight of the body which can be increases the risk.

## II. LITERATURE REVIEW

1)Early detection of diseases like diabetes is crucial, especially as modern diets high in sugar and fat increase the risk. Timely diagnosis allows for better management, and analysing patient data using computer-based methods can aid in early prediction. However, achieving high accuracy is essential for these systems to be reliable. Methods are widely used, the irrelevant or unstructured data can reduce their effectiveness. To address this, optimization techniques can enhance feature selection and improve classification performance. This study has the SVM with the Aquila Optimization (AO) algorithm, referred to as AO-SVM. The SVM is used for classification, while AO helps fine-tune its parameters. With this setup, the model reached an accuracy of 96% in just 0.42 seconds, along with 92.5% precision, 92.1% recall, and an F1-score of 0.91.

2) Non-Insulin Dependent Diabetes Mellitus, another name for type 2 diabetes, is a dangerous illness that affects millions of people worldwide and is estimated to be the cause of 2 to 5 million deaths annually. Early detection can



## International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

prevent serious consequences such as heart disease, renal failure, and other related issues. As machine learning (ML) advances, the use of predictive models in healthcare is increasing. This might make use of the UCI repository's Pima Indian Diabetes dataset. After thorough data preprocessing and feature selection using methods like Random Forest Importance and RFE, a various models, including KNN, Logistic Regression, SVM, Random Forest, Light GBM, and XGBoost, were employed with different train-test splits (60–40, 70–30, and 80–20). The LightGBM model achieved the highest accuracy of 91.47% while using the 80–20 split.

3) Machine learning is widely applied in fields like healthcare, banking, and education for extracting insights and making predictions. In healthcare, improving diagnostic accuracy and potentially saving lives. This study focuses on analyzing prediction performance from various research works and developing an effective ML-based model for diabetes prediction. Algorithms such as Decision Tree, Random Forest, SVM, K-NN, Naïve Bayes, and the MLP classifier were evaluated based on accuracy and other performance metrics.

4) Because machine learning makes early illness prediction possible, it significantly lessens the strain of medical workers. The cardiovascular mortality rate in India is greater than the global norm, at about 272 per 10,000. According to recent Ministry of Health and Family Welfare surveys, 11.5% of those 45 and older in both rural and urban regions have been diagnosed with diabetes. Diabetes and heart disease together continue to be the nation's top causes of mortality, even with the availability of therapies. Important health variables that are easily assessed at primary care facilities, including age, gender, blood pressure, glucose, skin thickness, and insulin levels, are crucial in predicting these disorders.

5) Diabetes is a chronic health condition characterized by consistently high blood sugar levels. Some of its common signs include constant hunger, excessive thirst, frequent urination, and unintentional weight loss. If not managed properly, it can lead to severe complications such as heart disease and nerve damage. Early detection is therefore vital to reduce its impact. In this study, machine learning methods—particularly the Random Forest and K-Nearest Neighbors (K-NN) classifiers—are applied.

### III. METHODOLOGY

#### A) ARCHITECTURAL DIAGRAM

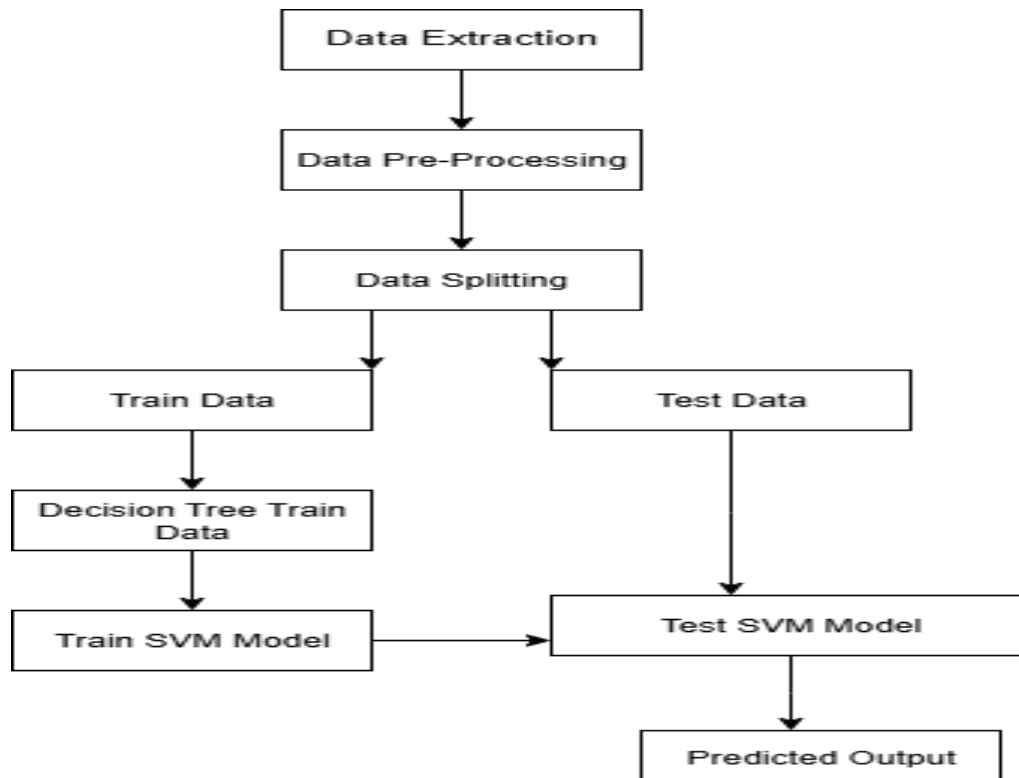


Fig.1. Architectural diagram



## International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

The architecture diagram begins with data extraction and moves through a set of stages, including data pre-processing and splitting the dataset into model fitting and evaluation portions. Two models are trained using the training data. Although both models undergo training, the diagram specifically highlights that only the SVM model is tested with the test data. The final output is derived from the SVM's predictions. This setup indicates a comparative or hybrid modeling strategy, where multiple models are trained, but one is ultimately chosen for testing and prediction based on its performance.

B) **Data extraction** is the foundational and essential step in any machine learning pipeline. When evaluating the performance of both models, it becomes possible to directly compare how effective each one is. Unlike SVM, which is often seen as a “black box” because it doesn't clearly show how decisions are made, decision trees are much easier to interpret. They provide a simple, visual pathway of how conclusions are reached, making it easier not only to understand the model's reasoning but also to explain it to others in a straightforward way.

C) **Data Reading:** Python is employed to develop the model. The loaded data using the `pread_csv` function from the pandas library. It contains medical records of 768 individuals, which are used to classify each person as diabetic or non-diabetic. Where 8 columns represent the independent features and the final column, labeled Outcome, indicates the class — specifying whether the individual is diabetic or not.

D) **Data Correlation:** Data correlation which measures the strength of the relationship between the different attributes in the correlation. Correlation measures the how changes in one feature are related to the changes in another. If altering the one of variable influences the level of another, then two are said to be correlated. Which can be examined relationships between all attributes in the dataset. The outcome variable showed the strongest correlation with patients' glucose levels, with a coefficient of 0.46. This was more than the correlation with a BMI (0.29), age (0.23), and the count of pregnancies (0.22). These will results that suggest the glucose level significant factor which is predicting whether a patient has diabetes. After glucose, BMI, age, and pregnancy count appear as the next most influential factors.

Outcome:1.000000  
 Pregnancies:17.000000  
 Glucose:199.000000  
 Blood Pressure:122.000000  
 SkinThickness:99.000000  
 Insulin:846.000000  
 BMI:67.100000  
 Diabetes Pedegree Function:2.420000  
 Age:81.000000

E) **Data Pre-processing:** This crucial step in the machine learning workflow takes place right after data. Here, the raw data is cleaned, formatted, and made ready for modeling. Typical pre-processing activities include addressing missing values, eliminating duplicate entries, encoding categorical data, normalizing or scaling numerical variables, and adjusting data types where needed. These processes ensure the data is clean, consistent, algorithms can easily understand and learn from. Effective pre-processing significantly enhances the quality of the data, which in turn improves the model's accuracy and overall performance by maintaining consistency across all features used in training and testing.

F) Elimination of NaN (Not a Number) values NaN values with the mean and median of the relevant characteristics are replaced.

G) **Data Scaling:** The each input has an equal influence on the model, data scaling is done as part of pre-processing to scale all feature values to a same scale. This is particularly crucial for algorithms like SVM, which perform better when the data is appropriately scaled and are sensitive to variations in feature magnitudes.

H) **Data splitting:** Data splitting, which comes after pre-processing but before model training starts, is a essential stage in the machine learning process. one for model testing and one for training. Decision Trees and Support Vector Machines (SVM) employ the training set to discover relationships and correlations among the features. In contrast, the testing set that can be used to evaluate the SVM's ability to generalise to new, unknown data, which the model hasn't seen during training. This divide promotes the model to discover significant patterns that may be applied to real-world



## International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

situations rather than just memorising the training data. Based on variables like dataset size and complexity, common split ratios are 80:20 or 70:30. Based on the predictions made from the test set, this method allows the model's performance to be evaluated through metrics such as accuracy, precision, and recall.

I) **The Support Vector Machine (SVM):** Fundamental categorisation method in the framework. The SVM model is trained using a subset of the training data after the data has been pre-processed and divided. In this stage, the algorithm finds the ideal hyper plane that, by maximizing the gap between them, maximally splits the dataset into discrete groups. following the training procedure. Using the patterns discovered during training, the goal is to correctly identify data that has never been seen before. When combined with kernel functions like polynomial kernels or the radial basis function (RBF), SVM is well-known for its robust performance in high-dimensional feature spaces and its capacity to reduce over fitting.

J) **Decision Tree algorithm:** Decision Tree Algorithm is incorporated as a straightforward, interpretable, and efficient classification method alongside the SVM model. It serves multiple purposes. Identifying the key knowledge of decision trees it can make easy to see the important in the prediction of outcomes.

### IV. RESULT

It helps us understand how accurately the model identifies diabetic and non-diabetic individuals, making its performance easier to evaluate. Four core values include,

True Positives(TP): The percentage of positive numbers correcting the diagnosed and as well with the diabetes.

True Negative(TN): This is the percent of individuals who are appropriately categorised as not being diabetic and its True Negatives (TN).

At the other end, False positives (FP) is the number of people without diabete who are

given a false diagnosis.False Negatives (FN): The diabetics who were incorrectly diagnosed as not having the condition.

The accuracy was considered good given the higher true values as compared to the false values for the matrix. As indicated by the high rate of false negatives, the model which many actual cases of diabetes. This directly reduces the model's recall, as it becomes less capable of identifying all positive cases. In any case, the model performs reasonably well, as the confusion matrix indicates, especially in predicting non-diabetic cases, which shows that the system is quite reliable.

Table1. Confusion Matrix Overview

		Predicted: Non-Diabetic (0)	Predicted: Diabetic (1)
Actual: Non-Diabetic (0)		True Negative (TN) 78	False Positive (FP)21
Actual: Diabetic (1)	False Negative (FN)18	True Positive (TP)37	



## International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

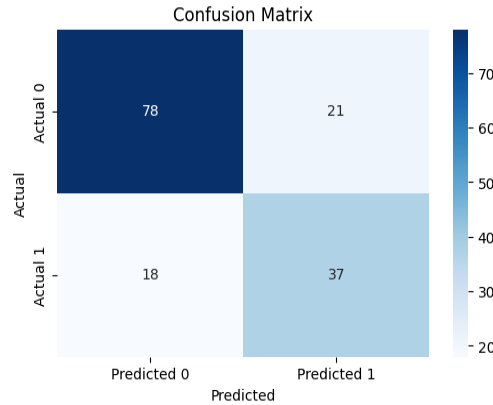


Fig.1 Confusion Matrix

**Accuracy:** Accuracy: The accuracy score of testing data 74.68%, and the accuracy score of training data 77.04% indicating it correctly classified most cases. This means it made the right prediction for about 77 out of every100 individuals.

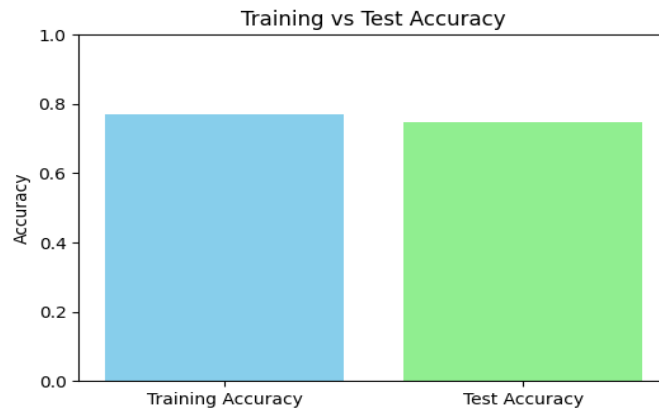


Fig.2 Training and Testing Data of the Accuracy Score

$$\text{Accuracy} = \frac{TP+TN}{FP+FNP+TN}$$

Accuracy=74.68% (Testing data)  
 Accuracy=77.04%(Training data)

**Precision:** The model's precision is 62.22%, meaning that most individuals predicted as diabetic were correctly identified.



## International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

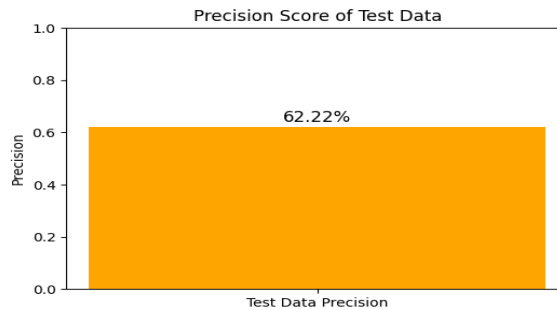


Fig.3 Precision Score

$$\text{Precision} = \frac{TP}{TP+PF}$$

$$\text{Precision} = 62.22\%$$

**Recall:** The recall of the model is 51.85%, which means it can identify over half of the actual diabetic cases. This indicates that the model missed a significant number of true positives.

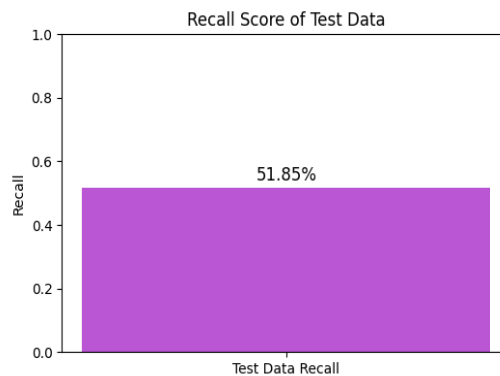


Fig.4 Recall Score

$$\text{Recall} = \frac{TP}{TP+NF}$$

$$\text{Recall} = 51.85\%$$

**F1 Score:** The F1 score of 56.57% It indicates that the model performs moderately well in identifying diabetic cases while minimizing false alarms.



## International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

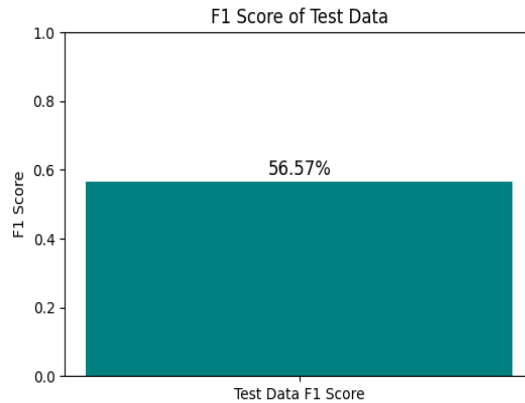


Fig.5 F1 Score

$$F1\ Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

F1 Score=56.57%

Metric	Value(%)	Interpretation
Accuracy	74.68(testing data)77.04(training data)	Overall correctness of predictions
Precision	62.22	How many predicted diabetics were correct
Recall	51.85	How many actual diabetics were correctly predicted
F1 Score	56.57	Trade off between precision and recall

Table2. Summary Table

### V. CONCLUSION

This project implemented a machine learning method for anticipating diabetes risk using the Pima Indians dataset. Support Vector Machine (SVM) with a linear kernel which used to perform binary classification, distinguishing between diabetic and non-diabetic individuals. To ensure unbiased training and evaluation, the data was standardized and split into model Prediction fitting and evaluation. The model attained an accuracy of nearly 77%, indicating on overall performance. It showed a precision of 62%, suggesting that when it predicted patient as diabetic, it was often correct. However, the recall was lower, around 52%, meaning the model failed to identify some diabetic cases. The F1 score stood at 56%, reflecting a moderate trade-off between precision and recall. The confusion matrix revealed a higher number of false negatives than desired. Future enhancements could aim at improving recall by adjusting model parameters or exploring alternative algorithms. Overall, this project highlights the usefulness of aiding early detection of diabetes.



## International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

### REFERENCES

- [1]Ajay V Associate Software Engineer PWC Acceleration Centre, Bengaluru ajayvdurga191999@gmail.com 2022 International conference on Artificial intelligence and Data Engineering(AIDE) An Aquila-optimized SVM Classifier for diabetes Prediction
- [2]Burla Praveen Kumar Reddy Associate Software Engineer Pwc Acceleration Centre, Bengaluru burlapraveenjee@gmail.com 2022 International conference on Artificial intelligence and Data Engineering(AIDE) An Aquila-optimized SVM Classifier for diabetes Prediction
- [3]Metun Telecommunication Engineering Bangalore institute of technology (VTU), Bengaluru, India metunnivn@gmail. 2022 International conference on Artificial intelligence and Data Engineering(AIDE) An Aquila-optimized SVM Classifier for diabetes Prediction
- [4]C. Charitha B.Tech Students in Electronics and Communication Engineering, SASTRA Deemed to be University, Thanjavur 613401,India 2022 International Conference On Computer Communication and Information(ICCCI) Type –II Diabetes Prediction Using Machine Learning Algorithms
- [5]Dr. C.Lakshmi Faculty in School of Electrical and Electronics Engineering, SASTRA Deemed to be University, Thanjavur 613401,India 2022 International Conference On Computer Communication and Information(ICCCI) Type –II Diabetes Prediction Using Machine Learning Algorithms
- [6]Amuluru DeviChaitrasree B.Tech Students in Electronics and Communication Engineering, SASTRA Deemed to be University, Thanjavur 613401,India 2022 International Conference On Computer Communication and Information(ICCCI) Type –II Diabetes Prediction Using Machine Learning Algorithms
- [7]Sathya Seelan K, Asst. professor Dept of Information Tech Sri Ramakrishna Institute of Tech Coimbatore, India subramaniaswami@gmail.com 2022 8th International Conference On Advanced Computing and Communication Systems(ICACCS) A Comparative Analysis of Diabetes Prediction Models Using Machine Learning Algorithms
- [8]Rakshith, UG Scholar Dept of Information Tech PSG Tech Coimbatore, India 2022 8th International Conference On Advanced Computing and Communication Systems(ICACCS) A Comparative Analysis of Diabetes Prediction Models Using Machine Learning Algorithms
- [9]Ram Varun, UG Scholar Dept of Information Tech PSG Tech Coimbatore, India 2022 8th International Conference On Advanced Computing and Communication Systems(ICACCS) A Comparative Analysis of Diabetes Prediction Models Using Machine Learning Algorithms
- [10]D. Sharathchandra PG student, Dept. of EIE, Kakatiya Institute of Technology & Science, Warangal, sharathdevaram11@gmail.com 2022 IEEE Delhi Section Conference(DELCON) ML Based Interactive Disease Prediction Model
- [11]M. Raghu Ram Assoc. Prof., Dept. of EIE, Kakatiya Institute of Technology & Science, Warangal, mrr.eie@kitsw.ac.in India 2022 IEEE Delhi Section Conference(DELCON) ML Based Interactive Disease Prediction Model
- [12]T Krishnaveni B.Tech(CSE) Vignan Institute of Technology and Science, Yadadri Bhuvanagiri, India.2021 Third International Conference On Inventive Research in Computing Applications(ICIRCA) Diabetes Prediction Using Different Machine Learning Algorithms
- [13]G.Nikitha B.Tech(CSE) Vignan Institute of Technology and Science, Yadadri Bhuvanagiri, India.2021 Third International Conference On Inventive Research in Computing Applications(ICIRCA) Diabetes Prediction Using Different Machine Learning Algorithms
- [14]S Kranthi Reddy Assistant Professor Department of Computer Science & Engineering Vignan Institute of Technology and Science, Yadadri Bhuvanagiri, India 2021 Third International Conference On Inventive Research in Computing Applications(ICIRCA) Diabetes Prediction Using Different Machine Learning Algorithms



INTERNATIONAL  
STANDARD  
SERIAL  
NUMBER  
INDIA



# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

 9940 572 462  6381 907 438  [ijircce@gmail.com](mailto:ijircce@gmail.com)



[www.ijircce.com](http://www.ijircce.com)

Scan to save the contact details