



e-ISSN: 2319-8753 | p-ISSN: 2347-6710

# IJIRSET

International Journal of Innovative Research in  
**SCIENCE | ENGINEERING | TECHNOLOGY**



# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 3, March 2024

**ISSN** INTERNATIONAL  
STANDARD  
SERIAL  
NUMBER  
INDIA

Impact Factor: 8.423



9940 572 462



6381 907 438



ijirset@gmail.com



www.ijirset.com

# Early Stage Diabetic Prediction Using Machine Learning

Evelyn Tabitha. E <sup>1</sup>, Mary Nisha. D <sup>2</sup>, Nagadevi. B <sup>3</sup>

Assistant Professor, Department of CSE, PET Engineering College, Vallioor, Tirunelveli, India <sup>1,2</sup>

PG Student, Department of CSE, PET Engineering College, Vallioor, Tirunelveli, India <sup>3</sup>

**ABSTRACT:** Diabetes Mellitus is a chronic disease characterized by hyperglycemia. It is a common disease for human body caused by metabolic disorder when the sugar level is high. It can cause many complications. According to growing morbidity by the year 2050, the world's diabetic patients will reach 740 million, which means that one of ten adults or children may suffer diabetes. Globally, diabetes affects 537 million people, making it the deadliest and the most common non-communicable disease. Many factors can cause a person to get affected by diabetes, like excessive body weight, abnormal cholesterol level, family history, physical inactivity, bad food habit etc. Increased urination is one of the most common symptoms of diabetes. Early prediction of such disease can save human life. In an automatic diabetes prediction system has been developed using a private dataset of female patients in Bangladesh and various machine learning techniques. Used the Pima Indian diabetes dataset to achieve the goal. With rapid development of machine learning, machine learning has been applied in many aspects of medical health. It using some popular machine learning algorithm namely, K-Nearest Neighbor (KNN) to predict diabetes mellitus. Further the work can be extended to find how likely non-diabetic people can have diabetes in the next few years and also, the predicted model can be used for machine learning in the future to find diabetes for the prediction of diabetic or non-diabetic.

**KEYWORDS:** Hyperglycemia, Metabolic disorder, Pima Indian Diabetes Dataset, K-Nearest Neighbor, Prediction

## I. INTRODUCTION

Diabetes isn't a hereditary disorder however heterogeneous group of disorder which could ultimately result in a boom of glucose within the blood and lack of glucose inside the urine. Diabetes is typically resulting from genetics, way of life and surroundings. Eating a dangerous weight loss plan, being overweight play role in developing the diabetes [1]. High blood sugar tiers can also result in kidney diseases, coronary heart illnesses. The excess of sugar in the blood can harm the tiny blood vessels in your frame. Signs of diabetes are blurry imaginative and prescient, extreme hunger, unusual weight reduction, common urination and thirsty. Parameters used within the facts set to locate the diabetes are Glucose, Blood pressure, pores and skin thickness, Insulin, Age. Huge volumes of statistics units are generated by health care industries. Those facts sets are a collection of patient information about the diabetes from the hospitals. Big records analytics is the processing which it examines the information units and exhibits the hidden information. Pima Indians Diabetes Database (PIDD), dataset is taken from the national Institute of Diabetes and Digestive diseases [2]. The objective of the dataset is to predict whether or not the patient has diabetes or not, primarily based on diagnostic measurements in the dataset. Several constraints were taken from the massive database.

Nowadays, machine learning and deep learning technologies are a blessing that are used in a variety of industries, including finance, fraud detection, crime detection, predictive modelling, and more. One of the most evolved fields and common use cases which are a boon is in Medical applications to understand any disease better. Diabetes is one of the most common diseases where patient's body ability to produce and respond to insulin is impaired which may result in increased glucose level in blood. Long-term complications of diabetes include vision problems, more thirst, frequent urination while on critical stages it can lead up to death. Hence, early detection of these diseases alongside with proper treatment can help the patient to revive to a better condition. Coronary Arteries are the major blood vessels that supply our heart with oxygen, blood and nutrients. When these arteries become damaged or impaired, the patient develops Coronary artery disease (CAD). Oily plaque builds up in these arteries which results in narrowing them, Hence, dropping the blood flow to patient's heart [3]. They silently kill patient without giving any major symptoms to them until acute circumstances are not created such as Heart Attack which many time results in patient's death. Hence, dependency of diseases such as Coronary Kidney Disease (CKD), chronic obstructive pulmonary disease (COPD), Hypertension (HTN), Hypothyroidism can help in early detection and help in management of the disease [4]. For

instance, used the K-Nearest Neighbor (KNN) algorithm to design a system that can predict diabetes quickly and accurately.

## II. REVIEW OF LITERATURE

Diabetes is a long-term illness caused by the inefficient use of insulin generated by the pancreas. If diabetes is detected at an early stage, patients can live their lives healthier. Unlike previously used analytical approaches, deep learning does not need feature extraction. In order to support this viewpoint, we developed a real-time monitoring hybrid deep learning-based model to detect and predict Type 2 diabetes mellitus using the publicly available PIMA Indian diabetes database. This study contributes in four ways. First, we perform a comparative study of different deep learning models. Based on experimental findings, we next suggested merging two models, CNN-Bi-LSTM, to detect (and predict) Type 2 diabetes [5]. These findings demonstrate that CNN-Bi-LSTM surpasses the other deep learning methods in terms of accuracy (98%), sensitivity (97%), and specificity (98%), and it is 1.1% better compared to other existing state-of-the-art algorithms.

S. Abhari, et. al Proposed the incidence of type 2 diabetes mellitus has increased significantly in recent years. With the development of artificial intelligence applications in healthcare, they are used for diagnosis, therapeutic decision making, and outcome prediction, especially in type 2 diabetes mellitus. Artificial intelligence (AI) has been applied in various medical fields for many purposes [6]. Actually, AI is defined as “a field of science and engineering concerned with the computational understanding of what is commonly called intelligent behavior with the creation of artifacts that exhibit such behavior”. Considering the importance of type 2 diabetes mellitus (T2DM) care as well as assuming that AI applications for diabetes care are effective tools.

Johnson, et. al studied diabetes is an enduring disease with likely to cause a global health care crisis. In entire world 452 million people are living with diabetes according to International Diabetes Federation. By 2040, this will be folded as 693 million. Diabetes is caused due to the growth level of blood glucose. Machine learning is a developing scientific field in data science dealing with the ways in which machines learn from experience. The project aims to develop an early diabetes prediction system that merges the results of different machine learning techniques to predict diabetes more accurately for patients at an earlier stage of their disease. The algorithms used in this project are K nearest neighbor, Logistic Regression, Random forest, Support vector machine and Decision tree. The correctness of the model using each of the algorithms is calculated. Then the model which have highest correctness is taken for predicting the diabetes. The Pima Indians Diabetes (PID) Data Set is used in experiment [7]. This data set is obtained from the UCI machine learning repository. Testing positive for diabetes means 1 and testing negative for diabetes means 0.269 (34.9%) cases are present in class 1 for positive test and 500 (65.1%) cases in class 0 for negative test.

Davis, et. al studied diabetes mellitus is one of the most common human diseases worldwide and may cause several health-related complications. It is responsible for considerable morbidity, mortality, and economic loss. A timely diagnosis and prediction of this disease could provide patients with an opportunity to take the appropriate preventive and treatment strategies. To improve the understanding of risk factors, we predict type 2 diabetes for Pima Indian women utilizing a logistic regression model and decision tree—a machine learning algorithm [8]. Our analysis finds five main predictors of type 2 diabetes: glucose, pregnancy, body mass index (BMI), diabetes pedigree function, and age. We further explore a classification tree to complement and validate our analysis. The six-fold classification tree indicates glucose, BMI, and age are important factors, while the ten-node tree implies glucose, BMI, pregnancy, diabetes pedigree function, and age as the significant predictors. Being by looking at the predictors of the prevalence of diabetes using the logistic regression model. Using goodness of fit tests and model selection criteria, compared each model with all possible combinations of predictors [9]. At 5% significance level, skin thickness, blood pressure and insulin do not predict diabetes, but interactions between age and pregnancy, glucose and insulin with pedigree provide more meaningful results.

Thompson, et. al applied due to the ever-increasing incidence of diabetes, effective screening strategies are needed for early diagnosis and intervention. This study proposes a novel approach that harnesses the power of artificial intelligence (AI) to predict diabetes risk [10]. Using machine learning techniques and a database with demographic, clinical and lifestyle variables, the proposed model achieves the best accuracy in predicting the probability of developing diabetes.



### III. PROPOSED SYSTEM

This consists of various modules of the proposed model. It consists of five modules, namely, **Module 1:** Data Collection, **Module 2:** Dataset Description, **Module 3:** Data Preprocessing, **Module 4:** Data Preparation, **Module 5:** Model Selection, **Module 6:** Apply Machine Learning.

**Module 1: Data Collection:** This is the first real step towards the real development of a machine learning model, collecting data. This is a critical step that will cascade in how good the model will be, the more and better data that we get, the better our model will perform [11]. There are several techniques to collect the data, like web scraping, manual interventions and etc.

**Module 2: Dataset Description:** The Pima Indian dataset is an open-source dataset that is publicly available for machine learning classification, which has been used in this work along with a private dataset. It contains 768 patients' data, and 268 of them have developed diabetes [12]. Dataset can be seen in following Table 3.1.

Pregnancies	Skin thickness	Diabetes pedigree function
Glucose	Insulin	Age
Blood pressure	BMI	

**Table 3.1 Features of the Pima Indian Dataset**

**Module 3: Data Preprocessing:** Data preprocessing is most important process. Mostly healthcare related data contains missing value and other impurities that can cause effectiveness of data. To improve quality and effectiveness obtained after mining process, Data preprocessing is done. To use Machine Learning Techniques on the dataset effectively this process is essential for accurate result and successful prediction. Missing Values Removal and Data splitting are used.

**Module 4: Data Preparation:** We will transform the data. By getting rid of missing data and removing some columns. First we will create a list of column names that we want to keep or retain. Next we drop or remove all columns except for the columns that we want to retain. Finally, we drop or remove the rows that have missing values from the data set. Split into training and evaluation sets.

**Module 5: Model Selection:** The principal component analysis is the technique that is used, especially for the reduction of the dimension of the given dataset [13]. The principal component analysis is one of the most efficient and an accurate method for reducing the dimensions of data, and it provides the desired results. This method reduces the aspects of the given dataset into a desired number of attributes called principal components. This method takes all the input as the dataset which is having a high number of attributes so as the dimension of the dataset is very high. This method reduces the size of the dataset by taking the data points on the same axis.

**Module 6: Apply Machine Learning:** When data has been ready we apply Machine Learning Technique. We use KNN Classifier technique, to predict diabetes. KNN classifier: A discrete-valued function can be approximated by  $K$  number of nearest classifiers. To categorize, it creates a plane with the available training points and calculates the distance between the query and trained points [14]. It determines the  $K$  number of neighbors (depending on the dataset) and classifies them using majority voting. In our research, we used  $K = 5$  for the binary classification.

### IV. METHODOLOGY

Disease Prediction using Machine Learning is the system that is used to predict the diseases from the symptoms which are given by the patients or any user. Parameters used within the facts set to locate the diabetes are Glucose, Blood pressure, pores and skin thickness, Insulin, Age. Huge volumes of statistics units are generated by health care industries. Pima Indians Diabetes Database (PIDD), dataset is taken from the national Institute of Diabetes and Digestive diseases. For instance, used the K-Nearest Neighbor (KNN) algorithm to design a system that can predict diabetes quickly and accurately.

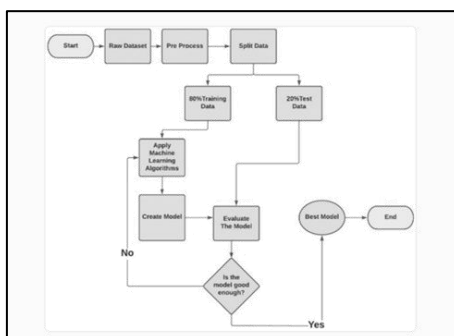


Fig 4.1 Proposed Diabetes Prediction System

In the statistical context, Machine Learning is defined as an application of artificial intelligence where available information is used through algorithms to process or assist the processing of statistical data. While Machine Learning involves concepts of automation, it requires human guidance. Machine Learning involves a high level of generalization in order to get a system that performs well on yet unseen data instances.

Most statistical techniques follow the paradigm of determining a particular probabilistic model that best describes observed data among a class of related models [15]. Similarly, most machine learning techniques are designed to find models that best fit data (i.e. they solve certain optimization problems), except that these machine learning models are no longer restricted to probabilistic ones.

Machine learning might be able to provide a broader class of more flexible alternative analysis methods better suited to modern sources of data. It is imperative for statistical agencies to explore the possible use of machine learning techniques to determine whether their future needs might be better met with such techniques than with traditional ones.

The KNN algorithm means **K-NEAREST NEIGHBOURS**. This algorithm often used in classification when we have some classified data and we have new data item, but we not sure which is the class of that new data, then we use KNN machine learning algorithm. KNN is supervised and pattern classification learning algorithm. KNN can be used in classification as well as regression. The KNN algorithm is the most accurate model because it makes highly accurate prediction, so we can use the KNN algorithm where we want highly accuracy. This algorithm has some drawback which is the outcomes accuracy is depend on the quality of the available data. So, if we have good quality of data then outcomes accuracy is higher else, we won't get the higher accuracy. The KNN algorithm is easy to implement there are two parameters is required to implement. First is the value of the K, and second one is the distance function. KNN is one of the most use data mining and classification algorithms. And KNN is used in cancer diagnosis, pattern recognition, text classification, email spam detection, fraud detection, and in regression it used to risk assessment, score prediction etc. in this paper we will see how KNN machine learning algorithm actually work and its challenges as well as advantages. KNN algorithm store all the available cases and classify new data of K in similarity measure. It suggests that if you are similar to your neighbor then you are one of them [16]. For example, apple is more similar to orange, banana, and mango rather than dog, monkey, lion then most likely apple is belonging to group of fruits. KNN used is search application when you want to similar items then you called the search as a KNN search. K is the number of neighbors near to the new object which we have to assign. If  $k=3$  then the most common three nearest neighbors are checked and the most common neighbors class are assigning to the testing data item. So, this is the K in KNN algorithm. The biggest use of KNN algorithm is the recommendation system. The recommended system is like the shop counter when you asking for a product it not shows only that product, it displayed you and also suggest your relevant sets of products and related to the item you are already too interested to buying it. The KNN algorithm is used in recommended product like in amazon and recommended media in case of NETFLIX. More than 35% of amazon revenue is generated by its recommendation engine. For this kind of purpose, we use KNN algorithm and more advanced example may like handwriting detection, image recognition or even video recognition, and it is used to get missing value, used in pattern recognition, and it used in gene expression this is also the example of KNN machine learning algorithm.

The k nearest neighbor algorithm find the nearest neighbor of new data item, if  $k=3$ , (k is the nearest neighbor) then 3 closest neighbors has been checked and most common of cases data item class has been assign to new data item. This is about KNN algorithm. So, the question arises how we calculate the distance between k and new data

item? We can measure distance between  $k$  and new data point through Euclidean distance. we can also calculate distance through hamming distance, Manhattan distance formula for KNN algorithm. The Euclidean distance is based on the Pythagoras theorem if we want to calculate the distance of  $AC^2(AC \text{ SQUAR})$  then in Pythagoras we calculate the  $AB^2 + BC^2(AB \text{ SQUAR} + BC \text{ AQUAR})$  this method can be used in Pythagoras. Likely if we want to calculate the distance between two point  $(x_1, x_2)$  and  $(y_1, y_2)$  in two dimensional space then the Euclidean distance between two points is  $D = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$ . ( $^2$  is denoted the square) And in three-dimensional space, for points  $(x_1, y_1, z_1)$  and  $(x_2, y_2, z_2)$  then in this case the Euclidean distance of this point is  $D = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$ . This is how Euclidean distance is calculated. Now will see how dose KNN algorithm work. We see how the Euclidean distance was calculated [17]. Now we apply this distance calculating formula in KNN and see how its work. So, in KNN we have to have data set first, before applying KNN algorithm. KNN will study this data and then predict the new data item class that's why KNN called as a lazy algorithm because it does not have specialized training phase it uses all data for training while classification. We have some flower sepal length and sepal width and their spices. So, if new flower sample is found but we don't know the new flower spices. We apply KNN algorithm to find the classification in flower spices. KNN algorithm use facture similarities to predict the value so the first step is storing the data set during the first step of KNN algorithm.

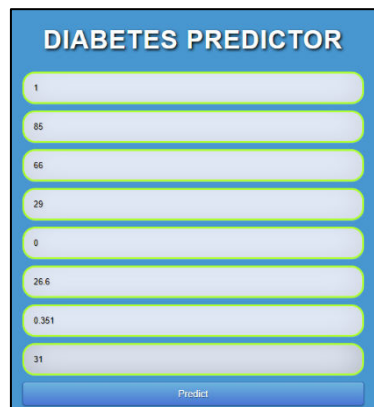
## V. RESULT AND DISCUSSION

First we have to enter the patient details in order to know the parameters. Here we have given the information's like Number of Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin Level, Body Mass Index, Diabetes pedigree Function and Age.



Fig 5.1 Home page

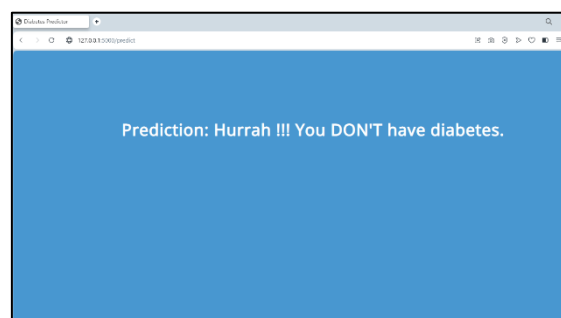
Here we have to enter the Patient Parameters like Number of Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin Level, Body Mass Index, Diabetes pedigree Function and Age.



The screenshot shows a web application titled "DIABETES PREDICTOR". It contains eight input fields with the following values: 1, 85, 66, 29, 0, 26.6, 0.351, and 31. A "Predict" button is located at the bottom of the form.

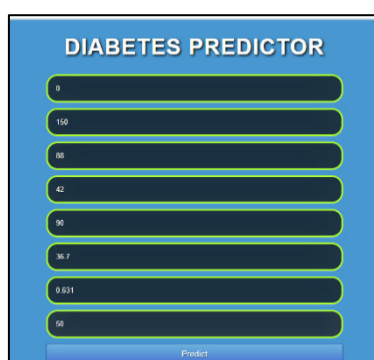
**Fig 5.2 Enter the Patient Parameters**

Based on given Number of Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin Level, Body Mass Index, Diabetes pedigree Function and Age the machine learning classifier will predict that the don't have diabetes.



**Fig 5.3 Diabetes has been predicted here**

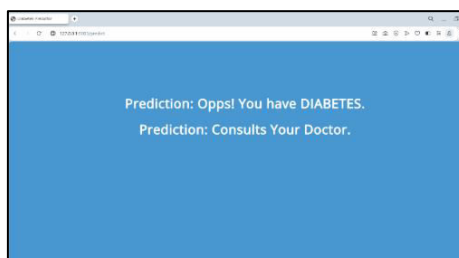
Here we have to enter the Another Patient Parameters like Number of Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin Level, Body Mass Index, Diabetes pedigree Function and Age.



The screenshot shows the same web application with input fields for another patient's parameters. The values entered are: 0, 158, 80, 42, 90, 36.7, 0.631, and 58. A "Predict" button is at the bottom.

**Fig 5.4 Enter the Another Patient Parameters**

Based on given Number of Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin Level, Body Mass Index, Diabetes pedigree function and Age the machine learning classifier will predict that the you have diabetes.

**Fig 5.5 Diabetes has been predicted here**

## VI. CONCLUSION AND FUTURE ENHANCEMENT

Diabetes prediction is one of the prevalent medical problems in real-world. The persistence of it in the human body for a long time leads to the microvascular complications of diabetes. Machine learning classification algorithm K-Nearest Neighbor (KNN) for predicting diabetes at an early stage. In Future work of creating an early stage diabetes risk prediction application may be done by using various machine learning model and compare the accuracy. There are some future scopes of this work, for example, we recommend getting additional private data with a larger cohort of patients to get better results.

## ACKNOWLEDGMENT

I would like to express my sincere thanks to Mrs. B. Shanmuga Sundari for her valuable guidance and support in complete this article.

I would also like to express my gratitude towards our college.

## REFERENCES

- [1] Talha Mahboob Alama , Muhammad Atif Iqbala , Yasir Alia , Abdul Wahabb , Safdar Ijazb , Talha Imtiaz Baigb , Ayaz Hussainc , Muhammad Awais Malikb , Muhammad Mehdi Razab , Salman Ibrarb , Zunish Abbasd. A model for early prediction of diabetes. Informatics in Medicine Unlocked 16 ( ) 100204.
- [2] M. Lichman, "Pima Indians diabetes database," ed. Center for machine learning and intelligent systems.: UCI Machine Learning repository.
- [3] Falvo D, Holland BE. Medical and psychosocial aspects of chronic illness and disability. Jones & Bartlett Learning; 2017.
- [4] Skyler JS, Bakris GL, Bonifacio E, Darsow T, Eckel RH, Groop L, et al. Differentiation of diabetes by pathophysiology, natural history, and prognosis. Diabetes 2017;66:241–55.
- [5] Tao Z, Shi A, Zhao J. Epidemiological perspectives of diabetes. Cell Biochem Biophys 2015;73:181–5.
- [6] Organization WH. World health statistics 2016: monitoring health for the SDGs sustainable development goals. World Health Organization; 2016.
- [7] Cho N, Shaw J, Karuranga S, Huang Y, da Rocha Fernandes J, Ohlrogge A, et al. IDF Diabetes Atlas: global estimates of diabetes prevalence for 2017 and projections for 2045. Diabetes Res Clin Pract 2018;138:271–81.
- [8] Diwani S, Mishol S, Kayange DS, Machuve D, Sam A. Overview applications of data mining in health care: the case study of Arusha regionl. Int J Comput Eng Res 2013;3:73–7.
- [9] Alam TM, Awan MJ. Domain analysis of information ExtractionTechniques. Int J Multidiscip Sci Eng 2018;9:1–9.
- [10] Alam TM, Khan MMA, Iqbal MA, Wahab A, Mushtaq M. Cervical cancer prediction through different screening methods using data mining. Int J Adv Comput Sci Appl 2019;10:388–96.
- [11] Cobos L. Unreliable hemoglobin A1C (HBA1C) in a patient with new onset diabetes after transplant (nodat). Endocr Pract 2018;24:43–4.
- [12] [Dorcely B, Katz K, Jagannathan R, Chiang SS, Oluwadare B, Goldberg IJ, et al. Novel biomarkers for prediabetes, diabetes, and associated complications. Diabetes, Metab Syndrome Obes Targets Ther 2017;10:345.
- [13] Singh PP, Prasad S, Das B, Poddar U, Choudhury DR. Classification of diabetic patient data using machine learning techniques. Ambient communications and computer systems. Springer; 2018. p. 427–36.
- [14] Negi A, Jaiswal V. A first attempt to develop a diabetes prediction method based on different global datasets. 2016 fourth international conference on parallel, distributed and grid computing. PDGC); 2016. p. 237–41.





- [15] Murat N, Dünder E, Cengiz MA, Onger ME. The use of several information criteria for logistic regression model to investigate the effects of diabetic drugs on HbA1c levels. Biomed Res 2018;29:1370–5.
- [16] Radin MS. Pitfalls in hemoglobin A1c measurement: when results may be misleading. J Gen Intern Med 2014;29:388–94.
- [17] Merad-boudia HN, Dali-Sahi M, Kachekouche Y, Dennouni-Medjati N. Hematologic disorders during essential hypertension," diabetes & metabolic syndrome. Clinical Research & Reviews; 2019.



INTERNATIONAL  
STANDARD  
SERIAL  
NUMBER  
INDIA



# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

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



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

Scan to save the contact details