JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

MACHINE LEARNING BASED WEB APPLICATION *FOR DIABETES PREDICTION*

[1] Kiran D. Yesugade, [2] Harshada V. Ankam, [3]Anushka A. Urunkar, [4]Poonam D. Dede, [5]Sonal S. Kale

[2,3,4,5]Student, Dept. of Computer Engineering, Bharati Vidyapeeth's College of Engineering for Women, Pune, Maharashtra, India

[1]Assistant Professor, Dept. of Computer Engineering, Bharati Vidyapeeth's College of Engineering for Women, Pune, Maharashtra, India

Abstract:

Diabetes is a global health problem that affects patients for the rest of their lives. People of all ages groups are affected by diabetes. Technology development is a novel technique to predicting diabetes and providing more accurate and efficient results. The majority of the studies were conducted to predict diabetes, and the majority of the researchers used the Pima Indian dataset. The authors of this research develop a framework for estimating the chance of diabetes in patients with the greatest accuracy. The authors of this research develop a framework for estimating the chance of diabetes in patients with the greatest accuracy. To diagnose diabetes in its early stages, the developers of this proposed system employ Machine Learning techniques such as Random Forest, SVM, Logistics Regression, and KNN. The Kaggle and Pima Indian Datasets are used in this proposed system. This proposed system includes three different diabetes prediction modules. Diabetes Prediction Using Blood Reports, Gestational Diabetes Prediction, and Diabetes Prediction Using Retina Images Here are three alternative modules for Diabetes Prediction based on users' preferences: the first is a medical report check, the second is just for pregnant women, and the third is retinal diabetes Prediction.

Keywords: Machine learning, XG Boost, Random Forest, KNN (K-Nearest Neighbors), SVM (Support Vector Machine), Logistic Regression, CNN (Convolutional Neural Network).

I. INTRODUCTION

Diabetes is one of the most dangerous metabolic diseases in the world. Diabetes is a chronic health issue. Diabetes affects around 400 million people worldwide. India has an estimated 77 million diabetics, second only to China in terms of diabetic population. In the supplied base article, they used Logistic Regression with 96 percent accuracy and AdaBoost with 98.8% accuracy to compare both Machine Learning Algorithms. As a result, we look at the accuracy of the KNN, SVM, Logistic Regression, CNN, and Random Forest algorithms to see which one has the best accuracy %. As a result, this is where you should implement the best algorithm for this project. As the number of diabetic patients worldwide rises, not all test labs are reliable or accurate. As a result, the Diabetes Prediction Web Application is proposed, which would assist users in analyzing and cross-checking test lab reports with actual data. If a patient has test lab reports, he or she can use this proposed application to self-test them. The user can also check himself using the retina scan image. This proposed system also includes a gestational module, which is specifically developed for pregnant women. On the basis of Diabetes Probability, users are recommended home cures, pharmaceuticals, and doctor's advice.

II Data and Sources of Data

Dataset obtained from the Kaggle website for this proposed system. Kaggle hosts enormous amounts of open-source public data from a variety of sources. Users can use Kaggle to search and publish data sets, explore and construct models in a web-based data science environment, collaborate with other data scientists and machine learning experts, and compete in data science challenges.

III RESEARCH METHODOLOGY

Machine Learning is used in the research for this proposed system. CNN (Convolutional Neural Network), KNN (K-Nearest Neighbors), SVM (Support Vector Machine), Logistic Regression, and Random Forest are only some of the machine learning algorithms employed.

1) Random Forest:

Random Forest is a popular machine learning technique for supervised learning. It can be used to handle both scheduling and retrieval challenges in machine learning. It is based on the concept of integrated learning, which is the process of merging multiple components to solve a complex problem and improve model performance quickly. The Random Forest is a sub-discipline that contains a number of decision trees for various datasets and collects measurements to enhance the dataset's prediction accuracy, as the name implies. Rather than relying on a single decision tree, the random forest forecasts the final result by generating projections on each tree based on a large number of expected votes.

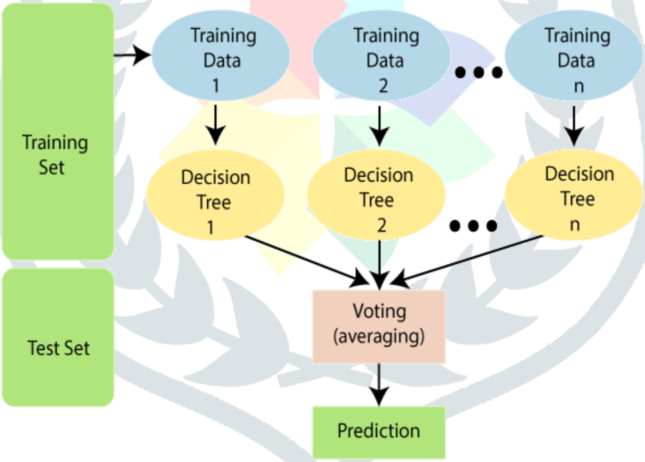
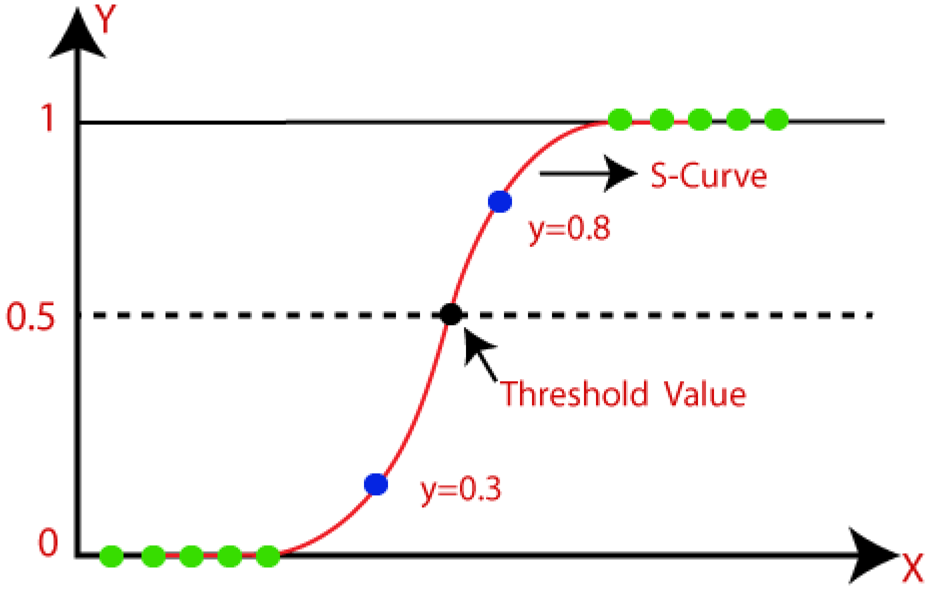


Fig 1. Random Forest

2) Logistic Regression:

- Under the Supervised Learning approach, one of the most prominent Machine Learning algorithms is logistic regression. It's used to predict a categorical dependent variable from a group of independent variables.
- A categorical dependent variable's output is predicted using logistic regression. As a result, the result must be a discrete or categorical value. It can be Yes or No, 0 or 1, true or false, and so on, but instead of giving exact values like 0 and 1, it delivers probabilistic values that fall between 0 and 1.



○

Fig 2: Logistic Regression

3) Convolutional Neural Network:

Day by day, computer vision evolves rapidly. Deep Learning is one of the reasons for this. Convolutional Neural Network is a term used in computer vision. Because CNN is heavily used here, it comes to mind. Face Recognition, Image Classification, and other applications of CNN in Computer Vision are examples. It resembles a simple Neural Network.

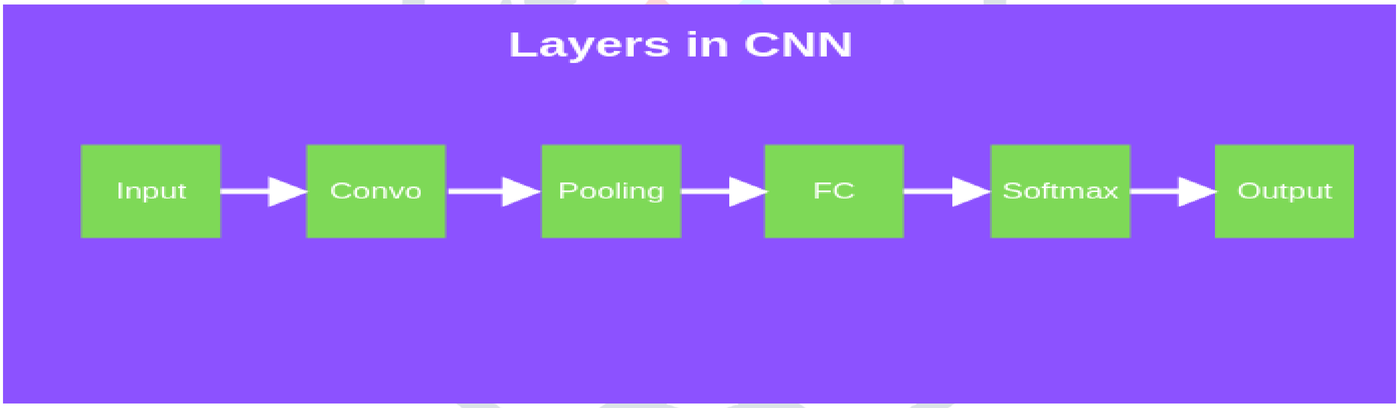


Fig 5: Layers in CNN

IV PROPOSED SYSTEM ARCHITECTURE

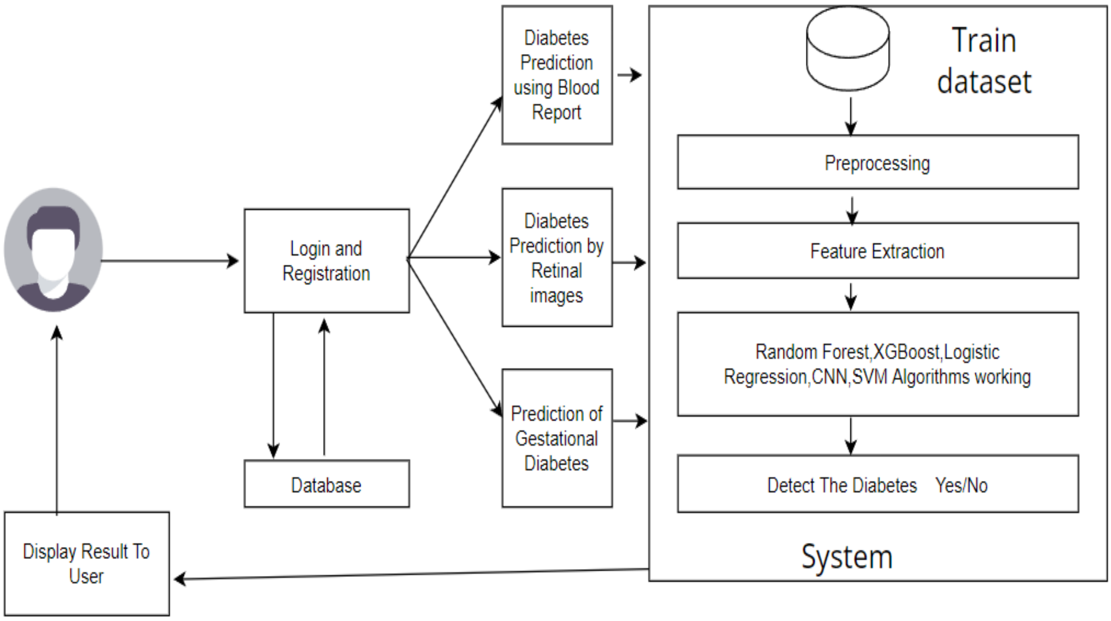

Login and Registration
Database
Display Result To User
Diabetes Prediction using Blood Report
Diabetes Prediction by Retinal images
Prediction of Gestational Diabetes
Train dataset
Preprocessing
Feature Extraction
Random Forest,XGBoost,Logistic Regression,CNN,SVM Algorithms working
Detect The Diabetes Yes/No
System

FIG 6: SYSTEM ARCHITECTURE

Diabetes is detected by this proposed system. Different Machine Learning Algorithms such as CNN (Convolutional Neural Network), KNN (K-Nearest Neighbors), Random Forest, Logistic Regression, and SVM were used to develop this system (Support vector Machine). This proposed system consists of three modules that allow users to register and login to the system.

Following registration, customers are given three alternative options from which to choose for Prediction.

This system has three modules for diabetes prediction.

1) Predicting Diabetes using Blood Report.

2) Predicting Gestational Diabetes.

3) Predicting Diabetes using Retina Images.

Predicting Diabetes using Blood Report

This module is used to predict Diabetes, therefore users must enter all required information from the blood report, such as glucose, BMI, age, skin thickness, and so on. Based on the attributes, the system will predict Diabetes Probability.

Predicting Gestational Diabetes

This module is only for females who want to predict gestational diabetes. Users must fill out all needed information in this module, including the number of pregnancies, BMI, Glucose, Age, and Pedigree Function. The system will predict diabetes based on the data.

Predicting Diabetes using Retina Images

This module is used to predict diabetes using retina images, and users must upload a retina image to do so.

V. RESULTS AND ACCURACY:

Various Machine Learning Algorithms are used to implement the proposed system. The accuracy of different machine learning algorithms varies. Different algorithms such as Random Forest, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), CNN (Convolutional Neural Network), Logistics Regression, and others are implemented in this proposed system.
These modules' accuracies are being compared here. The Accuracies of Algorithms are shown in the table below.

BLOOD REPORT MODEL ACCURACY TABLE

Algorithm	Accuracy
Random Forest	0.909091
SVM	0.763636
KNN	0.807273
Logistic Regression	0.763636

GESTATIONAL MODEL ACCURACY TABLE

Algorithm	Accuracy
Random Forest	0.831169
SVM	0.811688
KNN	0.779221
Logistic Regression	0.831169

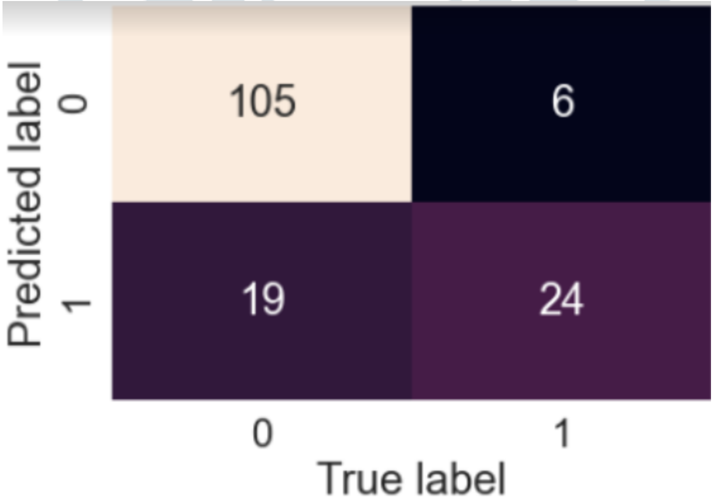
VI. ADVANTAGES

- Machine Learning is one of the most accurate techniques that can be applied in the system for predicting diseases.
- System can predict the diabetes automatically once it is modelled neatly and it will give the probability of having Diabetes.
- Machine Learning techniques are widely used for prediction of diseases at an early stage. Prediction of diabetes at an early stage will help to take accurate treatment and get remedy for diabetes.

VII. APPLICATIONS

- Diabetes prediction system is very useful system in the Healthcare and Medical Field.
- Common People can use this Web Application to verify their blood test reports.
- Since, there are many who struggle with diabetes, this Research of Detecting Diabetes using Machine Learning algorithms would be important to the modern society which will help them for their diabetes prediction.

VIII. CONFUSION MATRIX



VIII FUTURE SCOPE:

Diabetes prediction is the proposed system. We can improve this approach by applying machine learning to predict ailments and make drug recommendations. In the future scope of this project, more parameters and factors will be included. When the settings are increased, the accuracy will improve even more. We improve the accuracy by incorporating more machine learning approaches and algorithms. Retina images come in a variety of formats, including .png and .jpg.

VIX CONCLUSION:

As a result, we used machine learning algorithms such as Random Forest, KNN, SVM, logistic regression, CNN, and others to create a web application. We discovered the most accurate results by comparing these algorithms. Thus, we have examined and encountered that in Diabetes Prediction – Random Forest has 90% accuracy, Gestational Diabetes Prediction – Logistic Regression has 83% accuracy and CNN – Retinopathy. We've also developed a Retinopathy Detection Module to determine whether or not a person has diabetic retinal disease. Diabetes prediction has been achieved in this literature utilizing the proposed ensemble model from the dataset, where preprocessing is critical for robust and precise prediction. The proposed preprocessing technique increased the dataset's quality, with outlier rejection and missing value filling being top priorities.

REFERENCES

1. Aishwarya Mujumdar, Dr. Videhi V,"Diabetes Prediction using Machine Learning Algorithms", INTERNATIONAL CONFERENCE ON RECENT TRENDS IN ADVANCED COMPUTING 2019, ICRTAC 2019/

2. Gauri D. Kalyankar, Shivananda R. Poojara and Nagaraj V. Dharwadkar, " Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop", International Conference On I-SMAC,978-1-5090-3243-3,2017.

3. B. Nithya and Dr. V. Ilango," Predictive Analytics in Health Care Using Machine Learning Tools and Techniques", International Conference on Intelligent Computing and Control Systems, 978-1-5386-2745-7,2017.

4. Huma Naz1 & Sachin Ahuja "Deep learning approach for diabetes prediction using PIMA Indian dataset", 6 November 2019 /Accepted: 20 March 2020.

5. A.G Ramakrishnan, Kanika Verma, Prakash Deep "Detection and Classification of diabetic retinopathy using retinal images", December 2011.

6. [1] Kiran D. Yesugade, [2] Harshada V. Ankam, [3]Anushka A. Urunkar, [4]Poonam D. Dede, [5]Sonal S. Kale, "Diabetes Prediction using Machine Learning Algorithms", International Research Journal of Engineering and Technology (IRJET) Volume: 09 Issue: 04 | Apr 2022.