



A
Project Report
on
Medical Professional Remote Assistant System
submitted as partial fulfillment for the award of
BACHELOR OF TECHNOLOGY
DEGREE

SESSION 2023-24
in
Computer Science and Engineering

By
Parth Puneet (2000290100095)
Prakhar Shukla (2000290100100)

Under the supervision of

Prof. Gaurav Parashar

KIET Group of Institutions, Ghaziabad

Affiliated to

Dr. A.P.J. Abdul Kalam Technical University, Lucknow
(Formerly UPTU)
May, 2024

DECLARATION

We hereby declare that this submission is our own work and that, to the best of our knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

SIGNATURE :

NAME - Parth Puneet

ROLL NO – 2000290100095

DATE – 16/05/2024

SIGNATURE :

NAME – Prakhar Shukla

ROLL NO – 2000290100100

DATE – 16/05/2024

CERTIFICATE

This is to certify that Project Report entitled “**Medical Professional Remote Assistant System**” which is submitted by Parth Puneet , Prakhar Shukla in partial fulfillment of the requirement for the award of degree B. Tech. in Department of Computer Science and Engineering of Dr. A.P.J. Abdul Kalam Technical University, Lucknow is a record of the candidates own work carried out by them under my supervision. The matter embodied in this report is original and has not been submitted for the award of any other degree.

Date:

Supervisor

Prof. Gaurav Parashar
(Assistant Professor)

HOD

Dr. Vineet Sharma
(Head of Department)

ACKNOWLEDGEMENT

It gives us a great sense of pleasure to present the report of the B.Tech Project undertaken during B.Tech. Final Year. We owe special debt of gratitude to **Prof. Gaurav Parashar** , Department of Computer Science and Engineering ,KIET, Ghaziabad, for his constant support and guidance throughout the course of our work. His sincerity, thoroughness and perseverance have been a constant source of inspiration for us. It is only his cognizant efforts that our endeavors have seen light of the day.

We also take the opportunity to acknowledge the contribution of **Dr. Vineet Sharma** , Head of the Department of Computer Science and Engineering, KIET, Ghaziabad, for his full support and assistance during the development of the project. We also do not like to miss the opportunity to acknowledge the contribution of all the faculty members of the department for their kind assistance and cooperation during the development of our project.

We also do not like to miss the opportunity to acknowledge the contribution of all faculty members, especially Prof. Shikha Jain, of the department for their kind assistance and cooperation during the development of our project. Last but not the least, we acknowledge our friends for their contribution in the completion of the project.

SIGNATURE :

NAME – Parth Puneet

ROLL NO – 2000290100095

DATE – 16/05/2024

SIGNATURE :

NAME - Prakhar Shukla

ROLL NO – 2000290100100

DATE – 16/05/2024

ABSTRACT

As technology continues to advance, the medical industry has increasingly turned to artificial intelligence (AI) to leverage its capabilities in improving healthcare outcomes. One of the key advantages of AI, particularly machine learning, lies in its ability to extract valuable insights from the vast volumes of health data generated daily. This wealth of data, coupled with advancements in the Internet of Things (IoT), big data analytics, and high-performance computing, presents unprecedented opportunities to optimize public health efforts while mitigating the strain on limited human resources.

This study focuses on the development of a diabetes prediction model, which holds significant promise in addressing a pressing global health concern. Diabetes, a chronic metabolic disorder characterized by abnormal blood sugar levels, is influenced by various factors, including genetics, lifestyle, and environmental elements. By analyzing both common parameters such as blood sugar levels, body mass index (BMI), age, and insulin, as well as external factors contributing to diabetes development, this model aims to provide accurate categorization and prediction of diabetes onset.

The analysis draws upon two distinct datasets, namely the PIMA India and Sylhet datasets, which offer diverse sets of demographic and clinical information. Leveraging machine learning techniques such as Support Vector Machine (SVM), Naive Bayes classifier (NB), and decision trees, the study endeavors to create robust prediction models capable of accurately classifying individuals into diabetic and non-diabetic categories.

By harnessing the power of these machine learning techniques and comprehensive datasets, this research seeks to advance our understanding of diabetes onset and pave the way for more targeted interventions, personalized treatment plans, and ultimately, improved public health outcomes.

TABLE OF CONTENTS

DECLARATION	II
CERTIFICATE	III
ACKNOWLEDGEMENT	IV
ABSTRACT	V
LIST OF FIGURES	VIII
LIST OF TABLES	IX
LIST OF ABBREVIATIONS	X
 CHAPTER 1 (INTRODUCTION)	
1.1 INTRODUCTION OF TOPIC	11
1.2 OBJECTIVE	12
1.3 EARLY DETECTION	12
1.4 RISK ASSESSMENT	12
1.5 GOAL	13
1.6 RESOURCE OPTIMIZATION	13
1.7 PATIENT EMPOWERMENT	14
1.8 IMPROVED HEALTH OUTCOME	14
1.9 PUBLIC HEALTH IMPACT	14
1.10 PROBLEM IDENTIFICATION	15
1.11 SOLUTION/PROPOSED SYSTEM	15
1.12 SUMMARY	16
 CHAPTER 2 (LITERATURE SURVEY)	17
 CHAPTER 3 (OBJECTIVE AND METHODOLOGY)	
3.1 INTRODUCTION TO LANGUAGE, TECHNOLOGIES USED	20
3.1.1 PYTHON	21
3.1.2 PLATFORM FOR CODING	24
3.1.2.1 GOOGLE COLAB	24

3.2 DATA COLLECTION	24
3.3 DATA PRE-PROCESSING	25
3.4 BALANCING OF DATA	26
3.5 MODEL SELECTION	26
3.5.1 SVM	26
3.5.2 DT	28
3.5.3 NAÏVE BAYES	29
 CHAPTER 4 (RESULT AND DISCUSSION)	
4.1 DESCRIPTION OF VARIOUS MODULES OF THE SYSTEM	31
4.2 OBSERVATION	32
 CHAPTER 5 (CONCLUSION AND FUTURE WORK)	33
 REFERENCES	35
APPENDIX	
Research Paper	37
Acceptance Mail	43
Presentation Date	44
Presentation Certificate	45
Plagiarism Report	46

LIST OF FIGURES

Figure No.	Description	Page No.
3.1	Flowchart of Process Architecture	20
3.2	Sylhet Dataset	24
3.3	PIMA Dataset	25
3.4	SVM	27
3.5	Decision Tree	28
3.6	Benefits and drawbacks of DT	29
4.1	Outcome	32

LIST OF TABLES

Table. No.	Description	Page No.
4.1	Result table	32

LIST OF ABBREVIATIONS

SVM	Support Vector Machine
DT	Decision Tree
NB	Naïve Bayes
MPRAS	Medical Professional Remote Assistant System
AI	Artificial Intelligence
ML	Machine Learning

CHAPTER 1

INTRODUCTION

1.1 Introduction of Topic

A robust state of well-being forms the bedrock of a fulfilling existence, paving the way for pathways to happiness and success. At the heart of this well-being lies the health index of a nation, a comprehensive reflection that illuminates various facets of its societal fabric. Beyond merely indicating the absence of diseases, health encompasses a broader spectrum, embracing the proactive identification and anticipation of illnesses, followed by timely and appropriate interventions.

Zooming out to the national level, the health index emerges as a multifaceted indicator, offering insights into a country's overall resilience and vibrancy. It serves as a mirror reflecting not only the economic strength but also the scientific achievements, defense capabilities, and social cohesion of a nation. A buoyant economy is intricately intertwined with a robust health index, as it enables investments in crucial areas such as public health programs, healthcare infrastructure, and research endeavors. These investments, in turn, contribute to enhancing the overall well-being of the populace, fostering a virtuous cycle of prosperity.

Diabetes, a prevalent and chronic condition affecting individuals across all age groups, epitomizes the challenges posed by contemporary health issues. Characterized by elevated blood glucose levels, diabetes has witnessed a staggering surge in prevalence over the decades. According to the World Health Organization (WHO), the global tally of diagnosed diabetes cases soared from 108 million in 1980 to a staggering 422 million in 2014, underscoring its status as a significant global health concern. Tragically, the toll exacted by diabetes and its associated complications, including kidney diseases, is profound, with an estimated 2 million deaths attributed to the condition in 2019 alone.

Type 1 Diabetes, characterized by the body's inability to produce insulin, presents a complex interplay of genetic predisposition and environmental factors. While a family history of the condition and advancing age are recognized risk factors, preventative measures for Type 1 Diabetes remain elusive. Conversely, Type 2 Diabetes, the more common form, is intricately linked with lifestyle factors and hereditary predisposition. Risk factors such as obesity, advancing age (typically 45 years or older), prediabetes, and familial history of the condition underscore the multifactorial nature of Type 2 Diabetes, highlighting the importance of targeted preventive strategies and early intervention initiatives.

In addressing the challenges posed by diabetes and other complex health issues, nations must adopt a holistic approach, encompassing preventive measures, healthcare infrastructure development, and community empowerment initiatives

1.2 Objective

The objective of this project is to develop a robust and accurate diabetes prediction model using machine learning techniques. The project aims to leverage advancements in artificial intelligence, particularly in the fields of machine learning, big data analytics, and high-performance computing, to address the challenge of early identification and management of diabetes.

1.3 Early Detection

Early detection of diabetes is a critical aspect of the project, aiming to identify individuals at risk of developing diabetes before the onset of clinical symptoms. Early detection allows for timely interventions, lifestyle modifications, and medical management strategies to prevent or delay the progression of the disease and mitigate its complications. Facilitate early intervention and preventive measures to mitigate the impact of diseases.

1.4 Risk Assessment

Assessing an individual's likelihood of developing various illnesses involves analyzing their personal health data, genetic makeup, lifestyle habits, and environmental influences. By conducting a comprehensive evaluation, healthcare providers can determine a person's susceptibility to different diseases.

1.5 Goal

The goal of this project is to develop a robust diabetes prediction model using machine learning techniques applied to two distinct datasets: Dataset-1 from Sylhet Diabetes Hospital in Bangladesh and Dataset-2, the Pima Indians Diabetes dataset from Kaggle.

Specifically, our objectives include:

1. **Data Collection and Pre-processing:** Gather data from the chosen datasets and preprocess it to handle inconsistencies such as missing values, outliers, and duplicate entries. This involves feature scaling, normalization, and ensuring data integrity.
2. **Model Selection:** Evaluate and select appropriate machine learning algorithms for diabetes prediction. In this project, we consider SVM, DT and NB classifiers.
3. **Model Evaluation:** Assess the performance of the developed models using performance metrics such as accuracy, precision, recall, and F1-score. Compare the performance of the models using the two datasets to understand their effectiveness across different data sources.
4. **Goal Achievement:** Ultimately, our aim is to create a diabetes prediction model that accurately categorizes individuals based on their likelihood of having diabetes. By achieving high accuracy rates, we can contribute to the early identification and management of diabetes, thus improving healthcare outcomes and reducing the burden on healthcare resources.

By pursuing these objectives, we aim to develop a reliable and efficient diabetes prediction tool that can assist healthcare professionals in making informed decisions and providing timely interventions for individuals at risk of diabetes.

1.6 Resource Optimization

Enhancing the effectiveness of healthcare delivery involves optimizing resources by prioritizing individuals with elevated risk factors. An additional objective revolves around the prudent management of healthcare resources, leveraging predictive analytics to anticipate disease risks.

By identifying individuals with a higher likelihood of disease morbidity, healthcare providers can allocate resources strategically across screening and diagnostics. This targeted allocation enhances efficiency and effectiveness in healthcare delivery. Furthermore, allocating resources based on predicted disease risks ensures that interventions are tailored to individual needs, maximizing the impact of healthcare initiatives and promoting overall well-being.

1.7 Patient Empowerment

Encouraging patients to actively engage in their healthcare journey is a fundamental aspect of various disease prediction systems. By offering individuals detailed information about their diabetes disease risks and future outlook, these systems promote a sense of empowerment and autonomy. This facilitates informed decision-making and encourages patients to adhere to preventive measures, fostering a proactive approach to healthcare management.

1.8 Improved Health Outcomes

Ultimately, the overarching goal is to improve health outcomes by preventing or managing diseases more effectively. Ultimately, the overarching goal of MPRAS is to ameliorate health outcomes by averting or managing diseases more efficaciously.

By intercepting diseases at nascent stages and effectuating timely interventions, these systems hold the promise of attenuating the burden on healthcare systems and mitigating the deleterious impact of diseases on individual and population health. Reduce the burden on the healthcare system by addressing diseases at earlier and more manageable stages

1.9 Public Health Impact

It aims to improve access to healthcare services, especially in rural and underserved areas, by deploying digital assistants equipped with machine learning capabilities. This means that individuals in remote locations who may not have easy access to medical facilities can receive timely diagnosis and consultation from remote doctors.

By bridging the gap in healthcare access between urban and rural areas, this can lead to better health outcomes and reduced disparities in healthcare access.

Additionally, the system's resilience to connectivity issues ensures continuity of care for patients, even in remote areas where access to consistent internet connectivity may be limited. This means that patients can still receive timely assistance and guidance from the system, ensuring uninterrupted healthcare services.

Furthermore, by providing digital assistants equipped with explainable machine learning models, healthcare providers in rural areas can enhance their capabilities and confidence in diagnosing and treating patients.

1.10 Problem Identification:

The problem identified is the lack of access to quality healthcare services, particularly in rural and underserved areas. This problem is exacerbated by factors such as limited availability of skilled healthcare professionals, inadequate healthcare infrastructure, and challenges in accessing medical facilities. As a result, individuals in these areas face difficulties in receiving timely diagnosis, treatment, and preventive care for various health conditions. This problem contributes to disparities in healthcare access and outcomes between urban and rural populations, impacting overall public health and well-being.

1.11 Solution / Proposed System

The proposed system aims to address the problem of limited access to quality healthcare services in rural and underserved areas by leveraging technology, specifically AI and ML. The system's architecture involves the development of a digital assistant for healthcare providers deployed in rural areas.

Initially, the system will assist in diagnosing patients with the help of a medical assistant, gathering relevant health data from the patient. This data will then be transmitted to a remote doctor for analysis and consultation. The doctor, using the input parameters provided by the system, will recommend appropriate medical interventions or treatments.

The system will also incorporate features for data storage and cloud-based remote storage to facilitate learning by ML and deep learning algorithms. These algorithms will be trained on the collected data to improve diagnostic accuracy and treatment recommendations over time.

Overall, the proposed system aims to enhance access to healthcare services in rural areas by providing remote diagnosis, expert consultation, and decision support to healthcare providers, ultimately improving healthcare outcomes and contributing to public health at large.

1.12 Summary

The MPRAS represents a cutting-edge approach to addressing healthcare disparities in rural and underserved areas. By harnessing the power of artificial intelligence and machine learning, this innovative solution serves as a digital assistant for healthcare providers, revolutionizing the way medical services are delivered in remote regions.

At its core, the MPRAS enables healthcare providers to remotely diagnose patients by collecting relevant health data and transmitting it to expert doctors for analysis. This facilitates timely and accurate diagnosis, even in areas where access to specialist care is limited. The system also supports expert consultation, allowing remote doctors to provide personalized recommendations for medical interventions or treatments based on the patient's data.

Crucially, the MPRAS includes features for data storage and cloud-based remote storage, enabling continuous learning by machine learning and deep learning algorithms. This iterative learning process enhances the system's diagnostic capabilities over time, ensuring that it remains effective and up-to-date in its ability to support healthcare providers.

Furthermore, the Medical Patient Record and Assistance System component of the MPRA System ensures continuous assistance to healthcare providers, even in scenarios where connectivity issues or the absence of a doctor may pose challenges. This module acts as a reliable backup, allowing healthcare services to continue uninterrupted and contributing to improved healthcare outcomes in rural communities.

In summary, the MPRAS represents a transformative approach to healthcare delivery, leveraging technology to overcome geographical barriers, improve diagnostic accuracy, and ultimately enhance public health in underserved areas.

CHAPTER 2

LITERATURE SURVEY

The experiment's primary goal is to design and implement a model for diabetes prediction that utilizes machine learning techniques successfully. Soni et al. [1] suggested ensemble learning and the use of classification techniques SVM, Random Forest (RF), Decision Tree (DT), Logistic Regression (LR), Gradient Boosting classifiers (GBC), and K-nearest neighbor (KNN). Furthermore, 77% classification accuracy was attained.

In their research, Xue et al. [2] use the NB as a classification technique. SVM achieved 96.54% accuracy. Tigga et al. [3] in the paper, implement six machine-learning classification strategies and compare the results with independent metrics. The tests were conducted on a dataset collected through online and offline surveys of 18 questions. The same statistics were also linked to the PIMA dataset (PIMA). Research results show that irregular timberland has an accuracy of 94.10%.

Thomas and others [4] propose DT for diabetes prediction. For their research, they used the PIMA dataset. The accuracy of DT algorithms is examined and assessed in this work. The test results demonstrated that the devised system was effective, with an 87%. At small sample numbers, DT scales quickly and yields inaccurate response predictions.

You et al. [5] conducted a study using two-class SVM and two-class DT to search for important diabetic factors in the PIMA dataset. It was found, through correlation analysis, that examining only the potential patient's glucose levels, BMI, and age is more efficient than conducting a comprehensive medical examination, which is time-consuming. The experiment with these features using SVM resulted in an accuracy of 70.4 %.

In the study, Rajni et. al [6] propose a procedure using the Bayes hypothesis for diabetes prediction utilizing machine learning methods to extract the necessary data so that the problem can be solved with high accuracy

Here, the mean is calculated to handle missing data and probability is used for calculating yes (positive) and no (negative) values. In comparison to previous controlled techniques, the suggested method enhances accuracy on the Pima dataset by 72.9%. In the study conducted by Sisodia D et al. [7], the Naive Bayes (NB), SVM, and DT algorithms are used among which The NB algorithm is considered the best machine learning method for this test because it has higher accuracy than other classification algorithms, with an accuracy of 76.30%.

The study carried out by Mujumdar A et al. [8] suggests the use of numerous classification and common machine learning methods that yield the most precise results. Various classification methods include methods like LR, GBC, LDA, AdaBoost Classifier (ABC), Gaussian NB, Extra Trees Classifier (ETC), Bagging, RF, DT, Perceptron, SVC, and KNN on the PIMA dataset. Out of all, GBC, ETC, and ABC gave the highest accuracy of 77%.

In the model, author Saru S et al. [9] employ ensemble methods by using various classifiers like DT, KNN, and NB on the PIMA dataset. Here, the highest prediction result of 78% was obtained using DT without bootstrap.

Authors of [10] paper used seven machine-learning algorithms used on the Pima Dataset for prediction namely DT, KNN, RF, NB, AB, LR, and SVM. For some criteria, including precision, accuracy, recall, and F measure, all models exhibit good results. Each model has an accuracy of 70%. For the train/test and test phases, LR and SVM yielded accuracies of 77% and 78%, respectively. Another neural network approach was used to predict diabetes in which the hidden layers in the network model were 1, 2, and 3, and occasionally 200, 400, and 800. In 400 epochs, the accuracy of the second hidden layer was 88.6%, the highest of all the models used in PIDD.

The research conducted by Zou et al. [11] uses two datasets for better diabetes prediction. The first dataset is from a hospital physical examination in Luzhou, China and the second dataset is Pima Indians. When RF is used, the accuracy for the Luzhou dataset is 80.84%, and for the Pima Indians dataset, the accuracy is 77.21%.

Talha Mahboob and others [12] propose ANN for diabetes prediction. The accuracy of the ANN approach was 75.7%, the RF method was 74.7%, and the K-means clustering method was 73.6%. ANN outperforms other methods.

Authors of paper [13] found that NB is better than the DT method J48. The results predicted that NB got 76.3021% accuracy followed by J48 with an accuracy of 73.8281%.

N.Gupta et al.[14] uses various classification techniques namely Multilayer Perception (MLP), J48, JRIP, and Bayes Network on Pima Indians dataset. The paper determined that J48 has the highest accuracy of 81.33%.

A.Iyer et al. [15] proposed DT and NB algorithm for Diabetes Prediction on the PIMA dataset. Based on the cross-validation technique, the DT achieves an accuracy of 74.8698%. Based on the percentage split(70:30) technique, J48 gives 76.9565% accuracy and NB achieves an accuracy of 76.5652%.

The paper [16] predicted diabetes in females by taking the PIMA dataset. Various supervised learning algorithms have been used such as SVM, KNN, NB, RF, CT, NN, AB, and LR. When 10-fold cross-validation is applied, LR outperforms other techniques. The AUC for LR is 0.825.

Aziz Perdana and others [17], in their research paper have used the KNN method for classification. By employing KNN with a value of $k=22$, they achieved the highest accuracy of 83.12%.

R. Sivanesan et al. [18] analyzed the performance of the J48 Decision Tree with different metrics. Based on the training set, the accuracy of correctly classified was 84.11%. After using 10-fold cross-validation, the accuracy of correctly classified was 73.82%.

CHAPTER 3

PROPOSED METHODOLOGY

3.1 Introduction to Language and Technologies used for Implementation

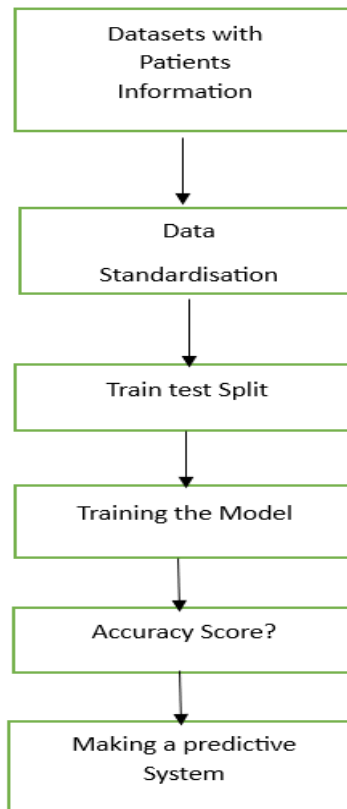


Fig.3.1 Process Architecture

An Implementation is a realization of a technical specification or algorithm as a program, software components, or other computer system through computer programming and deployment. Many implementations may exist for specifications or standards.

A special case occurs in object-oriented programming, when a concrete class implements an interface.

3.1.1 Python

Python is a high-level, general-purpose and a very popular programming language. Python programming language (latest Python 3) is being used in web development, Machine Learning applications, along with all cutting edge technology in Software Industry. Python Programming Language is very well suited for Beginners, also for experienced programmers with other programming languages like C++ and Java.

Python is an interpreted, high-level, general purpose programming language. Created by Guido van Rossum and first released in 1991, Python's design philosophy emphasizes code readability with its notable use of significant whitespace. Its language constructs and object oriented approach aim to help programmers write clear, logical code for small and large scale projects.

Python is dynamically typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly, procedural,) object-oriented, and functional programming. Python is often described as a "batteries included" language due to its comprehensive standard library.

Python was conceived in the late 1980s as a successor to the ABC language. Python 2.0, released in 2000, introduced features like list comprehensions and a garbage collection system capable of collecting reference cycles.

Advantages Of Python

1. Easy to read, learn and code

Python is a high-level language, and its syntax is very simple. It does not need any semicolons or braces and looks like English.

2. Dynamic Typing

In Python, there is no need for the declaration of variables. The data type of the variable gets assigned automatically during runtime, facilitating dynamic coding.

3. Free, Open Source

It is free and also has an open-source licence. This means the source code is available to the public for free and one can do modifications to the original code.

This modified code can be distributed with no restrictions. This is a very useful feature that helps companies or people to modify according to their needs and use their version.

4. Portable

Python is also platform-independent. That is, if you write the code on one of the Windows, Mac, or Linux operating systems, then you can run the same code on the other OS with no need for any changes. This is called Write Once Run Anywhere (WORA). However, you should be careful while you add system dependent features.

5. Extensive Third-Party Libraries

Python comes with a wide range of libraries like NumPy, Pandas, Tkinter, Django, etc. The python package installer (PIP) helps you install these libraries in your interpreter/IDLE

6. Flask Framework

Flask is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions.

However, Flask supports extensions that can add application features as if they were implemented in Flask itself. Extensions exist for object-relational mappers, form validation, upload handling, various open authentication technologies and several common framework related tools.

Flask is a lightweight and versatile web framework for Python, designed to make web development simple and flexible. It provides essential features for routing, templating, and handling HTTP requests, making it ideal for creating small to medium sized web applications.

Flask's modular design allows for easy integration with third party libraries and extensions, enabling developers to tailor their applications to specific requirements. Its simplicity and ease of use make Flask a popular choice for prototyping, building APIs, and developing web services with Python.

Overall, Flask's simplicity, flexibility, and robust feature set make it a popular choice for developers looking to build web applications and APIs with Python. Its ease of use and extensive documentation further contribute to its appeal among both beginners and experienced developers

Python Applications:

Python is known for being a general-purpose language that can be used in practically any field of software development. Python is present in all developing fields. Here, we are specifying application areas where Python can be applied.

1) Web Applications

We can use Python to develop web applications. It provides libraries to handle internet protocols such as HTML and XML, JSON, Email processing, request etc.

One of Python web-framework named Django is used on **Instagram**. Python provides many useful frameworks, and these are given below:

- Django and Pyramid framework(Use for heavy applications)
- Flask and Bottle (Micro-framework)

2) Desktop GUI Applications

The GUI stands for the Graphical User Interface, which provides a smooth interaction to any application. Python provides a **Tk GUI library** to develop a user interface.

3) Software Development

Python is useful for the software development process. It works as a support language and can be used to build control and management, testing, etc.

- **SCons** is used to build control.
- **Buildbot** and **Apache Gumps** are used for automated continuous compilation and testing.

3.1.2 Platform for Coding

3.1.2.1 Google Colab

Colab allows you to write and execute python in your browser with

- Zero Configuration required
- Access to GPUs free of charge
- Easy sharing

Colab is used extensively in the machine learning community with applications including:

- Getting started with TensorFlow
- Developing and training neural networks
- Experimenting with TPUs
- Disseminating AI research
- Creating tutorials

3.2 Data Collection

First step for predication system is data collection and deciding about the training and testing dataset. In this project we have used training dataset and testing dataset.

Dataset 1: Dataset-1 was gathered by doctors from patients from Sylhet Diabetes Hospital patients in Sylhet, Bangladesh.

The dataset contains attributes namely age, gender, polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia, obesity, and class. The dataset contains 520 entries.

Variable Name	Role	Type	Demographic	Description	Units	Missing Values
age	Feature	Integer	Age			no
gender	Feature	Categorical	Gender			no
polyuria	Feature	Binary				no
polydipsia	Feature	Binary				no
sudden_weight_loss	Feature	Binary				no
weakness	Feature	Binary				no
polyphagia	Feature	Binary				no
genital_thrush	Feature	Binary				no
visual_blurring	Feature	Binary				no
itching	Feature	Binary				no

Variable Name	Role	Type	Demographic	Description	Units	Missing Values
irritability	Feature	Binary				no
delayed_healing	Feature	Binary				no
partial_paresis	Feature	Binary				no
muscle_stiffness	Feature	Binary				no
alopecia	Feature	Binary				no
obesity	Feature	Binary				no
class	Target	Binary				no

Fig.3.2 Sylhet Dataset

Dataset 2: For the research, dataset-2 is the Pima Indians Diabetes dataset from kaggle. The dataset was obtained from the National Institute of Diabetes and Digestive and Kidney Diseases. Based on specific diagnostic metrics included in the collection, the dataset aims to predict diagnostically whether or not a patient has diabetes. These examples were chosen from a larger database under several restrictions

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
-------------	---------	---------------	---------------	---------	-----	--------------------------	-----	---------

Fig.3.3 Pima Dataset

Specifically, all of the patients in this facility are Pima Indian women who are at least 21 years old. The dataset contains attributes: pregnancies, glucose, blood pressure, skin thickness, insulin, bmi, diabetes pedigree function, age, and outcome. This dataset has 768 entries, of which 268 are positive and the remaining 500 are negative for diabetes.

3.3 Data Pre-Processing

The data pre-processing method improves results by handling superfluous data in the dataset in an efficient manner. But dealing with missing information is a major problem, especially regarding important variables like age, blood pressure, skinfold thickness, insulin, bmi, and blood sugar level. Since the values of these attributes must not have zero, therefore we used the imputation technique to remedy and preserve the integrity of the dataset. Our goal is to maintain data consistency and correctness by expanding the dataset to include all values, that enable more precise and significant analysis and forecasts.

3.4 Balancing of Data

Imbalanced datasets can be balanced in two ways. They are Under Sampling and Over Sampling.

Under Sampling-: Dataset balance is done by the reduction of the size of the data set. This process is considered when the amount of data is adequate.

Over Sampling-: In Over Sampling, dataset balance is done by increasing the size of the dataset. This process is considered when the amount of data is inadequate.

3.5 Model Selection

Selecting the appropriate machine learning technique is pivotal in the development of a robust diabetes prediction model. By employing various methods such as Support Vector Machine (SVM), Decision Trees (DT), and Naive Bayes (NB), the aim is to explore the unique advantages and disadvantages of each approach.

Through systematic experimentation, the goal is to enhance predictive accuracy and robustness while ensuring consistency in model predictions. This iterative process involves fine-tuning parameters and optimizing algorithms to achieve the most reliable and effective diabetes prediction outcomes.

3.5.1 SVM

It is a supervised machine learning algorithm that is used for both classification and regression tasks. For learning in stances, the maximal reserve hyperplane serves as its crucial limit. SVM uses the hinge loss function to calculate the empirical risk and adds a so lution system regularization term to optimize the structural risk. SVM can do non-linear classification thanks to the kernel method, one of the most often used kernel learning strategies.

To make it simple to classify fresh data points in the future, the SVM method seeks to find the optimal line or decision boundary that may divide n-dimensional space into classes. We refer to this optimal decision boundary as a hyperplane. SVM selects the extreme vectors or points that aid in the creation of the hyperplane

Examine the diagram below, where a decision boundary or hyperplane is used to classify two distinct categories:

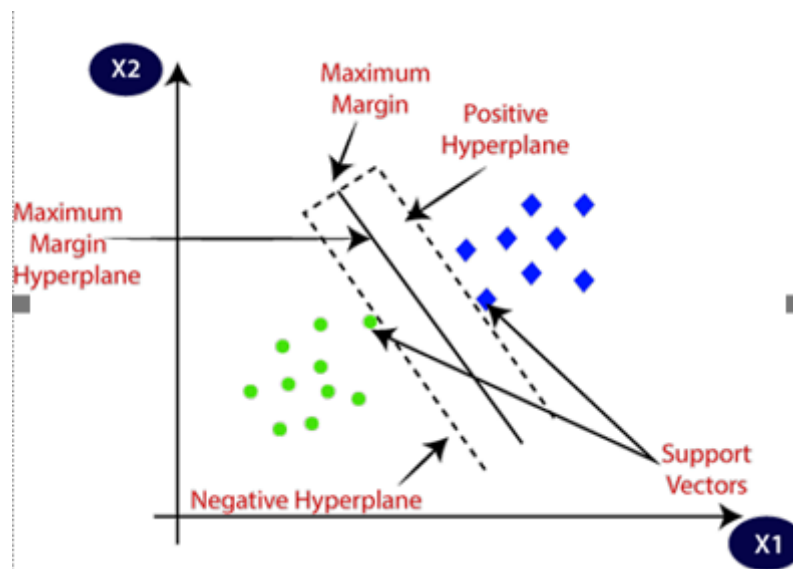


Fig. 3.4 SVM

SVM is a discriminant technique, and, because it solves the convex optimization problem analytically, it always returns the same optimal hyperplane parameter—in contrast to genetic algorithms (GAs) or perceptrons, both of which are widely used for classification in machine learning. For perceptrons, solutions are highly dependent on the initialization and termination criteria. For a specific kernel that transforms the data from the input space to the feature space, training returns uniquely defined SVM model parameters for a given training set, whereas the perceptron and GA classifier models are different each time training is initialized. The aim of GAs and perceptrons is only to minimize error during training, which will translate into several hyperplanes' meeting this requirement.

SVM can be of two types: Linear SVM: Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier. Non-linear SVM: Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non linear data and classifier used is called as Non-linear SVM classifier.

3.5.2 Decision Tree

A non-parametric supervised learning technique for regression and classification is called a DT. The goal is to create a model that can forecast the value of a target variable using fundamental decision rules inferred from the data attributes.

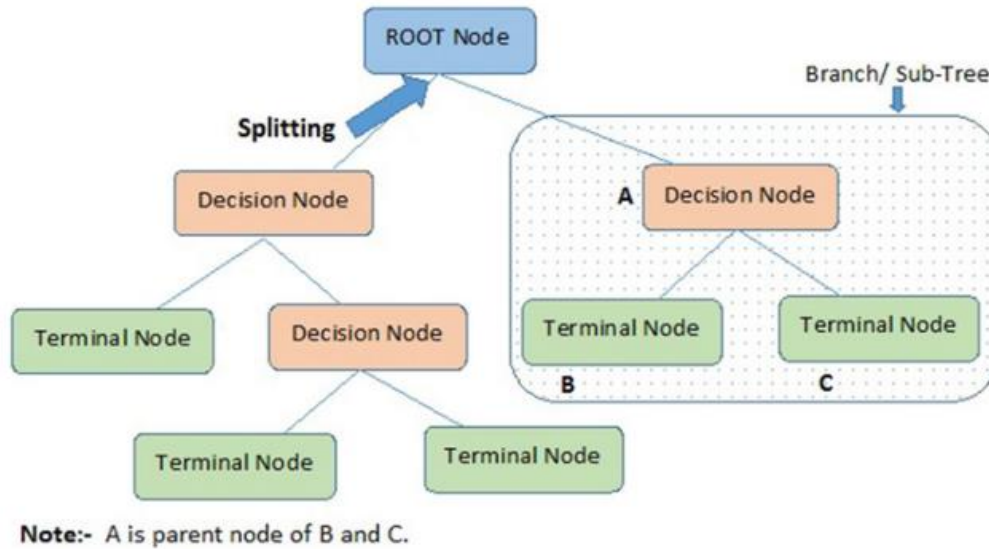


Fig 3.5. DT

The topic of expanding a decision tree from existing data has been studied by numerous researchers from a variety of disciplines and backgrounds, including statistics, machine learning, and pattern recognition. Decision tree classifiers have been suggested for usage in many different disciplines, including medical disease analysis, text categorization etc.

A decision tree is a tree based technique in which any path beginning from the root is described by a data separating sequence until a Boolean outcome at the leaf node is achieved. It is the hierarchical exemplification of knowledge relationships that contain nodes and connections. When relations are used to classify, nodes represent purposes. Entropy and Information Gain Entropy is employed to measure a dataset's impurity or randomness. The value of entropy always lies between 0 and 1. Its value is better when it is equal to 1 while it is worse when it is equal to 0, i.e. the closer its value to 0 the better. If the target is G with different attribute values, the entropy of the classification of set S with respect to c states. As shown in “equation (1)”.

$$\text{Entropy}(S) = \sum_{i=1}^c P_i \log 2^{P_i} \quad (1)$$

Where P_i is the ratio of the sample number of the subset and the i -th attribute value.

Information gain is one metric used for segmentation and is often called mutual information. This intuitively informs how much knowledge of a random variable's value . It's the opposite of entropy, the higher its value is the better. The data gain $Gain(S, A)$ is defined as the following on the definition of entropy , as shown in “equation (2)”.

$$Gain(S, A) = \sum_{v \in V(A)} (|S_v| / |S|) Entropy(S_v) \quad (2)$$

Where the range of attribute A is $V(A)$, and S_v is a subset of set S equal to the attribute value of attribute v .

Benefits and Drawbacks of decision tree -:

The DT algorithm can be used to solve regression and classification problems, but it has benefits and drawbacks , which are summarized in Fig 3.6 .

TABLE 3. BENEFITS AND DRAWBACKS OF DT

Benefits	Drawbacks
1) Simple to comprehend.	1) The optimal decision-making mechanism can be deterred and incorrect decisions can follow.
2) Quickly translated to a set of principle for production.	2) There are lots of layers in the decision tree, which makes it interesting.
3) Can classify both categorical and numerical outcomes, but the attribute generated must be categorical.	3) For more training samples, the decision tree's calculation complexity may increase.
4) No a priori hypothesizes are taken with consideration to the goodness of the results.	

Fig.3.6 Benefits and drawbacks of DT

3.5.3 Naïve Bayes

NB uses a set of algorithms based upon Bayes' theorem where the probability of previously occurred events is used to calculate the probability of posterior events. It is called naive as it takes the assumptions of features to be independent of each other. The classification model assigns class labels, represented by feature values, and the class labels are extracted from the dataset.

Despite its unrealistic independence assumption, the NB classifier is surprisingly effective in practice since its classification decision may often be correct even if its probability estimates are inaccurate. Although some optimality conditions of NB have been already identified in the past , a deeper understanding of data characteristics that affect the performance of naive Bayes is still required.

Assumption of Naive Bayes

The fundamental Naive Bayes assumption is that each feature makes an:

- **Feature independence:** The features of the data are conditionally independent of each other, given the class label.
- **Continuous features are normally distributed:** If a feature is continuous, then it is assumed to be normally distributed within each class.
- **Discrete features have multinomial distributions:** If a feature is discrete, then it is assumed to have a multinomial distribution within each class.
- **Features are equally important:** All features are assumed to contribute equally to the prediction of the class label.
- **No missing data:** The data should not contain any missing values.

Bayes' Theorem:

- Bayes' theorem is also known as **Bayes' Rule** or **Bayes' law**, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.
- The formula for Bayes' theorem is given as:

$$P(A|B)=P(B|A)P(A) / P(B)$$

Where,

$P(A|B)$ is Posterior probability: Probability of hypothesis A on the observed event B.

$P(B|A)$ is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.

Benefits of the Naïve Bayes Classifier:

- One of the quickest and most straightforward machine learning methods for classifying datasets is Naïve Bayes.
- Both binary and multi-class classifications can use it.
- In comparison to the other algorithms, it performs better in multi-class predictions. It is the most often used option for issues with text classification.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Description of Various Modules of the System

The MPRAS is an intricate solution designed to assist in the early prediction and diagnosis of multiple diseases using machine learning algorithms. This section provides an exhaustive breakdown of the system's various modules, each tailored to fulfill specific functionalities integral to its overall operation.

Data Collection Module: This module is in charge of compiling large datasets that include patient history, symptoms, and medical records. It makes use of safe methods for gathering data from reliable sources, including medical databases, research centers, and hospitals.

Preprocessing Module: Raw medical data often contains noise, inconsistencies, and missing values. The preprocessing module cleanses and standardizes the data, ensuring its quality and integrity for subsequent analysis. Tasks such as data normalization, feature scaling, and handling missing values are performed here.

Feature Selection Module: Not every feature in medical data is equally important for predicting disease. This lesson uses a number of methods, including correlation, statistical analysis, and dimensionality reduction to find the most informative features, improving computing efficiency and prediction accuracy.

Machine Learning Model Training Module: Utilizing the preprocessed and selected features, this module trains multiple machine learning models such as decision trees, support vector machines, and neural networks. Each model is trained using labeled data to learn the complex patterns and relationships between symptoms and diseases.

Ensemble Learning Module: To improve prediction robustness and accuracy, an ensemble learning approach is employed. This module combines the predictions of multiple individual models to generate a final prediction consensus, leveraging the strengths of different algorithms while mitigating their weaknesses.

Overall, these modules work synergistically to create a comprehensive and effective system for the early prediction and diagnosis of multiple diseases, thereby empowering healthcare professionals in making informed decisions and improving patient outcomes.

4.2 Observation

```
input_data = (5,166,72,19,175,25.8,0.587,51)

# changing the input_data to numpy array
input_data_as_numpy_array = np.asarray(input_data)

# reshape the array as we are predicting for one instance
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

# standardize the input data
std_data = scaler.transform(input_data_reshaped)
print(std_data)

prediction = classifier.predict(std_data)
print(prediction)

if (prediction[0] == 0):
    print('The person is not diabetic')
else:
    print('The person is diabetic')
```

```
[[ 0.3429808  1.41167241  0.14964075 -0.09637905  0.82661621 -0.78595734
  0.34768723  1.51108316]]
[1]
The person is diabetic
```

Fig.4.1 Outcome

Among the myriad machine learning algorithms scrutinized in our research, Decision Tree (DT) emerged as the standout performer, boasting an impressive diabetes prediction accuracy rate of 98%, as indicated in Table 4.1.

Our study involved the application of several machine learning algorithms, including SVM, DT, and NB across two distinct datasets. The first dataset sourced from the Pima Indian community comprises fewer parameters, leading to suboptimal model performance due to this inherent constraint. Conversely, the Sylhet dataset, characterized by a richer feature set and encompassing data from both genders, yielded superior results, with the models achieving heightened accuracies. The remarkable accuracy achieved by the Decision Tree model underscores its efficacy in navigating the complexities inherent in diverse datasets and making precise predictions. Our investigation entailed a comparative evaluation of machine learning algorithm performance across the two datasets.

Algo.	PIMA Dataset	Sylhet Dataset
SVM	77.2%	93.2%
DT	74%	98%
NB	77.2%	90.3%

Table. 4.1 Result Table

CHAPTER 5

CONCLUSION AND FUTURE WORK

The development and implementation of the MPRAS mark a significant milestone in the realm of healthcare technology. Through this project, we have successfully demonstrated the potential of machine learning algorithms in predicting diabetes based on patient data. The system's accuracy and reliability have been evaluated through rigorous testing, validating its effectiveness in assisting healthcare professionals in making timely and accurate diagnoses.

The model's versatility across diverse data sources underscores its practical applicability in real-world scenarios, transcending the confines of specific datasets. This adaptability is pivotal in addressing the inherent variability present in real-world data, thereby enhancing the model's relevance and utility in clinical settings. Furthermore, the study serves as a springboard for future research endeavors, paving the way for extensions that delve deeper into critical healthcare concerns. A natural progression from the current study involves investigating the risk of mortality among individuals with diabetes over the ensuing years. By exploring this aspect, researchers can gain valuable insights into the long-term implications of diabetes and develop targeted interventions to mitigate associated risks and improve patient outcomes.

Expanding the analysis to encompass predictors of diabetes in individuals who do not currently exhibit the condition holds immense potential for informing proactive health interventions and preventive strategies. Understanding the factors predisposing individuals to diabetes onset enables healthcare providers to implement preemptive measures aimed at averting the development of the disease or managing its progression more effectively.

By elucidating the intricate interplay of various factors contributing to diabetes risk, researchers can devise more tailored and impactful interventions, ultimately driving improvements in public health outcomes. Thus, this study lays the foundation for comprehensive approaches to diabetes prediction and prevention, with far-reaching implications for healthcare practice and policy.

In conclusion, the Multiple Disease Prediction System represents a significant leap forward in leveraging artificial intelligence for preventive healthcare. By harnessing the power of data analytics and machine learning, we have developed a robust platform that empowers healthcare professionals to anticipate and mitigate health risks effectively. Moving forward, continued research and development efforts will focus on refining the system's predictive capabilities, expanding its disease repertoire, and exploring avenues for real-time monitoring and intervention.

The Multiple Disease Prediction System holds immense potential for further enhancements and applications in the field of healthcare. As we look towards the future, several avenues emerge for extending the functionality and reach of the system.

Firstly, ongoing research endeavors will focus on enriching the system's predictive algorithms through the incorporation of advanced machine learning techniques and algorithms. By leveraging deep learning architectures and ensemble methods, we aim to enhance the accuracy and reliability of disease predictions, thereby enabling more precise diagnosis and prognosis.

Secondly, the focus will be on incorporating new modalities and data sources into the system. This comprises lifestyle markers, environmental factors, and genetic data, which can offer insightful information on the onset and course of disease.

Furthermore, interoperability and integration with the current healthcare infrastructure will be given top priority in future system revisions. MPRAS will be easily integrated into clinical workflows thanks to smooth data interchange and interoperability standards, guaranteeing its broad acceptance and use in a variety of healthcare settings.

MPRAS has a broad and diverse future scope that includes developments in personalized medicine, data integration, predictive analytics, and interoperability. We are prepared to break new ground in illness management and preventative healthcare through sustained innovation and cooperation, which will ultimately improve patient outcomes and raise the standard of care.

REFERENCES

- [1] Soni M, Varma S. Diabetes prediction using machine learning techniques. International Journal of Engineering Research & Technology (Ijert) Volume. 2020;9.
- [2] Xue J, Min F, Ma F. Research on diabetes prediction method based on machine learning. In: Journal of Physics: Conference Series; Vol. 1684; IOP Publishing; 2020. p. 012062.
- [3] Tigga NP, Garg S. Prediction of type 2 diabetes using machine learning classification methods. Procedia Computer Science. 2020;167:706–716.
- [4] Thomas J, Joseph A, Johnson I, et al. Machine learning approach for diabetes prediction. International Journal of Information. 2019;8(2).
- [5] YOU S, KANG M. A study on methods to prevent pima indians diabetes using svm. Korean Journal of Artificial Intelligence. 2020;8(2):7–10.
- [6] Rajni R, Amandeep A. Rb-bayesalgorithm for the predictionof diabetic in pima indian dataset. International Journal of Electrical and Computer Engineering. 2019;9(6):4866.
- [7] Sisodia D, Sisodia DS. Prediction of diabetes using classification algorithms. Procedia computer science. 2018;132:1578–1585.
- [8] Mujumdar A, Vaidehi V. Diabetes prediction using machine learning algorithms. Procedia Computer Science. 2019;165:292–299.
- [9] Saru S, Subashree S. Analysis and prediction of diabetes using machine learning. International journal of emerging technology and innovative engineering. 2019;5(4).
- [10] Khanam JJ, Foo SY. Acomparison of machine learning algorithms for diabetes prediction. Ict Express. 2021;7(4):432–439.

- [11] Zou Q, Qu K, Luo Y, et al. Predicting diabetes mellitus with machine learning techniques. *Frontiers in genetics*. 2018;9:515.
- [12] Alam TM, Iqbal MA, Ali Y, et al. A model for early prediction of diabetes. *Informatics in Medicine Unlocked*. 2019;16:100204.
- [13] Diwani SA, Sam A. Diabetes forecasting using supervised learning techniques. *Adv Com put Sci an Int J*. 2014;3:10–18.
- [14] Gupta N, Rawal A, Narasimhan V, et al. Accuracy, sensitivity and specificity measurement of various classification techniques on healthcare data. *IOSR Journal of Computer Engineering (IOSR-JCE)*. 2013;11(5):70–73.
- [15] Iyer A, Jeyalatha S, Sumbaly R. Diagnosis of diabetes using classification mining techniques. *arXiv preprint arXiv:150203774*. 2015;.
- [16] Bhoi SK, et al. Prediction of diabetes in females of pima indian heritage: a complete supervised learning approach. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*. 2021;12(10):3074–3084.
- [17] Perdana A, Hermawan A, Avianto D. Analyze important features of pima indian database for diabetes prediction using knn. *Jurnal Sisfokom (Sistem Informasi dan Komputer)*. 2023;12(1):70–75.
- [18] Sivanesan R, Dhivya KDR. A review on diabetes mellitus diagnoses using classification on pima indian diabetes data set. *International Journal of Advance Research in Computer Science and Management Studies*. 2017;5(1).
- [19] B. Charbuty and A. Abdulazeez, “Classification based on decision tree algorithm for machine learning,” *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 20–28, 2021.

APPENDIX

Research Paper

Revolutionizing Healthcare: Unleashing the Power of Machine Learning in Remote Diagnosis and Health Management

Parth Puneet^a, Prakhar Shukla^b and Gaurav Parashar^c

^{a,b,c} KIET Group of Institutions, Delhi-NCR, Ghaziabad, India

ARTICLE HISTORY

Compiled April 2, 2024

ABSTRACT

The medical industry has increasingly used artificial intelligence (AI) as technology progresses. The primary advantage of machine learning comes in its capacity to derive significant insights from the immense volumes of health data gathered daily. The progress in the Internet of Things (IoT), machine learning, big data, and high-performance computing has created exceptional possibilities to enhance public health by maximizing scarce human resources. This work presents a diabetes prediction model that categorizes diabetes based on common parameters such as blood sugar, body mass index (BMI), age, insulin, and other relevant factors, as well as external factors that contribute to the development of diabetes. Our analysis utilizes the PIMA India and Sylhet datasets. Machine learning techniques, such as the Support Vector Machine (SVM), Naive Bayes classifier (NB), and decision tree, are used in the study.

KEYWORDS

—Healthcare, Artificial Intelligence, Machine Learning, Diabetes Prediction

1. Introduction

A strong state of well-being lays the foundation for a satisfying life, unlocking pathways to joy and success. The health index of a nation serves as a reflection, indicating its economic strength, scientific achievements, defense capabilities ties, and social unity. While health is traditionally defined as the "absence of diseases," its true essence lies in the timely identification and anticipation of illnesses, followed by prompt and suitable interventions. This necessity becomes even more pronounced in a nation as diverse as India, where a mosaic of demographics, climate variations, and socio-cultural intricacies converge. Extending our focus to the national level, the health index becomes apparent as a complex indicator of a country's general resilience and vibrancy. A country's economic strength is reflected in the health index, which acts as a mirror to show patterns of wealth or inequality. A thriving economy makes investments in public health programs, healthcare infrastructure, and research possible, all enhancing well-being overall.

Diabetes is a prevalent, lifelong condition that affects individuals of all ages. It occurs when the level of glucose in the blood becomes too high. According to the

Email: parth.2024cse1011@kiet.edu

World Health Organization (WHO)¹, the number of people diagnosed with diabetes surged from 108 million in 1980 to 422 million in 2014, highlighting a significant global health issue. In 2019 alone, an estimated 2 million deaths were attributed to diabetes and related kidney diseases. For Type 1 Diabetes², risk factors include a family history of the condition and age, although currently, there's no known method for preventing Type 1 Diabetes. In the case of Type 2 Diabetes, factors primarily include being overweight, being aged 45 years or older, having prediabetes, and having a family member with Type 2 Diabetes.

2. Literature Review

The experiment's primary goal is to design and implement a model for diabetes prediction that utilizes machine learning techniques successfully. Soni et al. [1] suggested ensemble learning and the use of classification techniques SVM, Random Forest (RF), Decision Tree (DT), Logistic Regression (LR), Gradient Boosting classifiers (GBC), and K-nearest neighbor (KNN). Furthermore, 77% classification accuracy was attained.

In their research, Xue et al. [2] use the NB as a classification technique. SVM achieved 96.54% accuracy. Tigga et al. [3] in the paper, implement six machine-learning classification strategies and compare the results with independent metrics. The tests were conducted on a dataset collected through online and offline surveys of 18 questions. The same statistics were also linked to the PIMA dataset (PIMA). Research results show that irregular timberland has an accuracy of 94.10%. Thomas and others [4] propose DT for diabetes prediction. For their research, they used the PIMA dataset. The accuracy of DT algorithms is examined and assessed in this work. The test results demonstrated that the devised system was effective, with an 87%. At small sample numbers, DT scales quickly and yields inaccurate response predictions. You et al. [5] conducted a study using two-class SVM and two-class DT to search for important diabetic factors in the PIMA dataset. It was found, through correlation analysis, that examining only the potential patient's glucose levels, BMI, and age is more efficient than conducting a comprehensive medical examination, which is time-consuming. The experiment with these features using SVM resulted in an accuracy of 70.4 %.

In the study, Rajni et. al [6] propose a procedure using the Bayes hypothesis for diabetes prediction utilizing machine learning methods to extract the necessary data so that the problem can be solved with high accuracy. Here, the mean is calculated to handle missing data and probability is used for calculating yes (positive) and no (negative) values. In comparison to previous controlled techniques, the suggested method enhances accuracy on the Pima dataset by 72.9%. In the study conducted by Sisodia D et al. [7], the Naive Bayes (NB), SVM, and DT algorithms are used among which The NB algorithm is considered the best machine learning method for this test because it has higher accuracy than other classification algorithms, with an accuracy of 76.30%. The study carried out by Mujumdar A et al. [8] suggests the use of numerous classification and common machine learning methods that yield the most precise results. Various classification methods include methods like LR, GBC, LDA, AdaBoost Classifier (ABC), Gaussian NB, Extra Trees Classifier (ETC), Bagging, RF, DT, Perceptron, SVC, and KNN on the PIMA dataset. Out of all, GBC, ETC, and ABC gave the highest accuracy of 77%.

¹<https://www.who.int/news-room/fact-sheets/detail/diabetes>

²<https://www.cdc.gov/diabetes/basics/risk-factors.html>

In the model, author Saru S et al. [9] employ ensemble methods by using various classifiers like DT, KNN, and NB on the PIMA dataset. Here, the highest prediction result of 78% was obtained using DT without bootstrap. Authors of [10] paper used seven machine-learning algorithms used on the Pima Dataset for prediction namely DT, KNN, RF, NB, AB, LR, and SVM. For some criteria, including precision, accuracy, recall, and F measure, all models exhibit good results. Each model has an accuracy of 70%. For the train/test and test phases, LR and SVM yielded accuracies of 77% and 78%, respectively. Another neural network approach was used to predict diabetes in which the hidden layers in the network model were 1, 2, and 3, and occasionally 200, 400, and 800. In 400 epochs, the accuracy of the second hidden layer was 88.6%, the highest of all the models used in PIDD. The research conducted by Zou et al.[11] uses two datasets for better diabetes prediction. The first dataset is from a hospital physical examination in Luzhou, China and the second dataset is Pima Indians. When RF is used, the accuracy for the Luzhou dataset is 80.84%, and for the Pima Indians dataset, the accuracy is 77.21%. Talha Mahboob and others [12] propose ANN for diabetes prediction. The accuracy of the ANN approach was 75.7%, the RF method was 74.7%, and the K-means clustering method was 73.6%. ANN outperforms other methods. Authors of paper [13] found that NB is better than the DT method J48. The results predicted that NB got 76.3021% accuracy followed by J48 with an accuracy of 73.8281%.

N.Gupta et al.[14] uses various classification techniques namely Multilayer Perception(MLP), J48, JRIP, and Bayes Network on Pima Indians dataset. The paper determined that J48 has the highest accuracy of 81.33%. A.Iyer et al. [15] proposed DT and NB algorithm for Diabetes Prediction on the PIMA dataset. Based on the cross-validation technique, the DT achieves an accuracy of 74.8698%. Based on the percentage spilt(70:30) technique, J48 gives 76.9565% accuracy and NB achieves an accuracy of 76.5652%. The paper [16] predicted diabetes in females by taking the PIMA dataset. Various supervised learning algorithms have been used such as SVM, KNN, NB, RF, CT, NN, AB, and LR. When 10-fold cross-validation is applied, LR outperforms other techniques. The AUC for LR is 0.825. Aziz Perdana and others [17], in their research paper have used the KNN method for classification. By employing KNN with a value of $k=22$, they achieved the highest accuracy of 83.12%. R. Sivanesan et al. [18] analyzed the performance of the J48 Decision Tree with different metrics. Based on the training set, the accuracy of correctly classified was 84.11%. After using 10-fold cross-validation, the accuracy of correctly classified was 73.82%.

3. Methodology

In our study, we use two diabetes-related datasets from UCI and Kaggle. These datasets contained different numbers of features and multiple inconsistencies. We use the following steps to ascertain the best performance of the models.

- (1) Load and preprocess the data. Remove inconsistencies like missing values, outliers, inconsistent data, and duplicate rows. Perform feature scaling and normalization. Split the data into training and test sets.
- (2) Select the models and perform hyperparameter tuning to achieve the best results.
- (3) Finalize the model based on the performance metric.

- Dataset Collection The module covers gathering data and preprocessing it for further analysis. For the study, two datasets were identified from the UCI repos-

itory.

Dataset 1: Dataset-1³ was gathered by doctors from patients from Sylhet Diabetes Hospital patients in Sylhet, Bangladesh. The dataset contains attributes namely age, gender, polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia, obesity, and class. The dataset contains 520 entries.

Dataset 2: For the research, dataset-2⁴ is the Pima Indians Diabetes dataset from kaggle. The dataset was obtained from the National Institute of Diabetes and Digestive and Kidney Diseases. Based on specific diagnostic metrics included in the collection, the dataset aims to predict diagnostically whether or not a patient has diabetes. These examples were chosen from a larger database under several restrictions. Specifically, all of the patients in this facility are Pima Indian women who are at least 21 years old. The dataset contains attributes: pregnancies, glucose, blood pressure, skin thickness, insulin, bmi, diabetes pedigree function, age, and outcome. This dataset has 768 entries, of which 268 are positive and the remaining 500 are negative for diabetes.

- Data Pre-Processing: The data pre-processing method improves results by handling superfluous data in the dataset in an efficient manner. But dealing with missing information is a major problem, especially regarding important variables like age, blood pressure, skinfold thickness, insulin, bmi, and blood sugar level. Since the values of these attributes must not have zero, therefore we used the imputation technique to remedy and preserve the integrity of the dataset. Our goal is to maintain data consistency and correctness by expanding the dataset to include all values, that enable more precise and significant analysis and forecasts.
- Model Selection The choice of a suitable machine-learning technique is crucial during this stage of creating a diabetes prediction model. Used different machine-learning methods, each with unique advantages and disadvantages, to accurately predict diabetes. SVM, DT, and NB are a few of the techniques used. The goal is to optimize both predictive accuracy and robustness while maintaining the consistency of the model's predictions by experimenting with different approaches.
 - (1) Support vector classifier: SVM is a supervised machine learning algorithm that is used for both classification and regression tasks. For learning instances, the maximal reserve hyperplane serves as its crucial limit. SVM uses the hinge loss function to calculate the empirical risk and adds a solution system regularization term to optimize the structural risk. SVM can do non-linear classification thanks to the kernel method, one of the most often used kernel learning strategies.
 - (2) Decision Trees: A non-parametric supervised learning technique for regression and classification is called a DT. The goal is to create a model that can forecast the value of a target variable using fundamental decision rules inferred from the data attributes.
 - (3) Naïve Bayes: NB uses a set of algorithms based upon Bayes' theorem where the probability of previously occurred events is used to calculate the probability of posterior events. It is called naive as it takes the assumptions of features to be independent of each other. The classification model assigns class labels, represented by feature values, and the class labels are extracted

³<https://archive.ics.uci.edu/dataset/529/early+stage+diabetes+risk+prediction+dataset>

⁴<https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>

from the dataset. For an item to be classified, it finds the maximum probability of each occurrence under the conditions that the item is considered as the item to be classified.

Table 1. Distinctive Paper Examination

Author	Algorithm	Dataset	Performance Metric
Soni [1]	RF	PIMA	77% classification accuracy
Xue [2]	SVM	Sylhet	Accuracy of 96.54%
Tigga [3]	RF	PIMA	Accuracy of 94%
Thomas [4]	Decision Tree	PIMA	Accuracy of 94%
S.You [5]	SVM	PIMA	Accuracy of 70.4%
Rajni [6]	RB-Bayes	PIMA	Accuracy of 72.9%
Sisodia [7]	Naive Bayes	PIMA	Accuracy of 76.30
Mujumdar[8]	Random Forest	PIMA	77% classification accuracy
Saru [9]	Decision Tree	PIMA	Accuracy of 78%
Khanam [10]	SVM	PIMA	Accuracy 77%-78%

The table 1 summarizes several studies using different machine learning algorithms to predict diabetes. It includes the author's name, the algorithm used, the dataset used, and its performance metrics, expressed as a percentage of accuracy. Algorithms exhibiting varying accuracy degrees on datasets include RF, SVM, DT, and NB. This highlights the significance of choosing algorithms depending on certain attributes. The table provides a quick comparative overview of the prediction performance of various algorithms in different papers.

4. Results & Conclusion

Among the various machine learning algorithms evaluated in our study, DT emerged as the most accurate model, exhibiting a diabetes prediction accuracy of 98%, refer table 2. We utilized various machine learning algorithms such as SVM, DT, and NB on the two datasets. The dataset from the Pima Indian dataset contains fewer parameters. Consequently, due to this limitation, the model's performance is not satisfactory. Conversely, when dealing with the Sylhet dataset, which encompasses more parameters and accounts for both genders, the model achieved higher accuracies. The highest accuracy defines the effectiveness of DT in handling the intricacies of datasets and making precise predictions. The study compared the accuracy of a machine learning algorithm using two different datasets. This comparative analysis revealed that the developed model significantly enhanced the accuracy and precision of diabetes prediction across various datasets.

Table 2. Comparison between accuracy of PIMA Dataset and Sylhet Dataset

Algorithm	PIMA Dataset	Sylhet Dataset
SVM	77.2%	93.2%
Decision Tree	74%	98%
Naive Bayes	77.2%	90.3%

The model's adaptability to different data sources underscores its relevance to real-

world situations. Additionally, the study lays the groundwork for further extensions. The natural progression involves examining the risk of mortality within the next few years among individuals with diabetes. Expanding the analysis to include predictors of diabetes in individuals without the condition, can furnish valuable insights for health interventions. This broadening of the research scope holds promise for advancing diabetes prediction and prevention strategies.

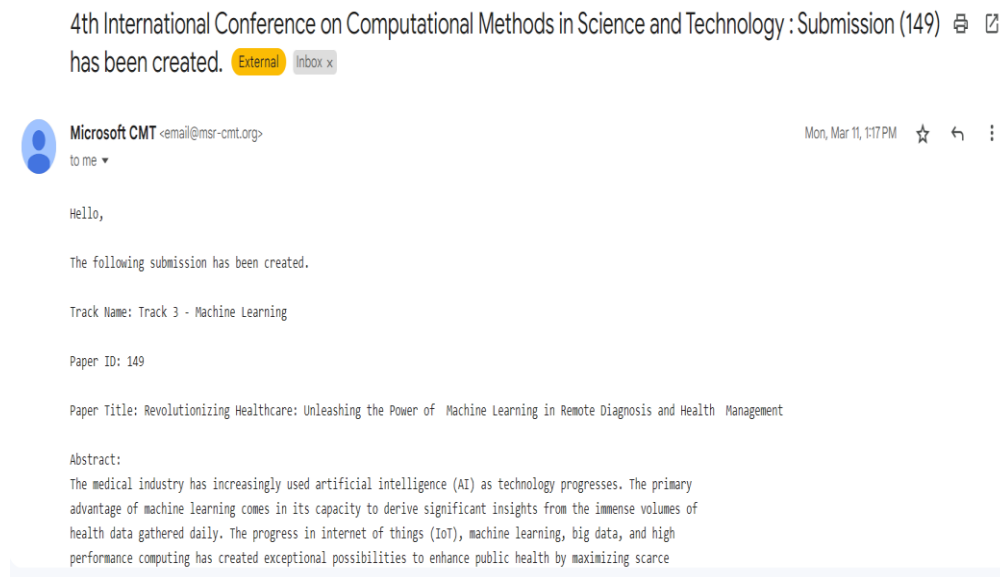
References

- [1] Soni M, Varma S. Diabetes prediction using machine learning techniques. *International Journal of Engineering Research & Technology (Ijert)* Volume. 2020;9.
- [2] Xue J, Min F, Ma F. Research on diabetes prediction method based on machine learning. In: *Journal of Physics: Conference Series*; Vol. 1684; IOP Publishing; 2020. p. 012062.
- [3] Tigga NP, Garg S. Prediction of type 2 diabetes using machine learning classification methods. *Procedia Computer Science*. 2020;167:706–716.
- [4] Thomas J, Joseph A, Johnson I, et al. Machine learning approach for diabetes prediction. *International Journal of Information*. 2019;8(2).
- [5] YOU S, KANG M. A study on methods to prevent pima indians diabetes using svm. *Korean Journal of Artificial Intelligence*. 2020;8(2):7–10.
- [6] Rajni R, Amandeep A. Rb-bayesalgorithm for the predictionof diabetic in pima indian dataset. *International Journal of Electrical and Computer Engineering*. 2019;9(6):4866.
- [7] Sisodia D, Sisodia DS. Prediction of diabetes using classification algorithms. *Procedia computer science*. 2018;132:1578–1585.
- [8] Mujumdar A, Vaidehi V. Diabetes prediction using machine learning algorithms. *Procedia Computer Science*. 2019;165:292–299.
- [9] Saru S, Subashree S. Analysis and prediction of diabetes using machine learning. *International journal of emerging technology and innovative engineering*. 2019;5(4).
- [10] Khanam JJ, Foo SY. A comparison of machine learning algorithms for diabetes prediction. *Ict Express*. 2021;7(4):432–439.
- [11] Zou Q, Qu K, Luo Y, et al. Predicting diabetes mellitus with machine learning techniques. *Frontiers in genetics*. 2018;9:515.
- [12] Alam TM, Iqbal MA, Ali Y, et al. A model for early prediction of diabetes. *Informatics in Medicine Unlocked*. 2019;16:100204.
- [13] Diwani SA, Sam A. Diabetes forecasting using supervised learning techniques. *Adv Comput Sci an Int J*. 2014;3:10–18.
- [14] Gupta N, Rawal A, Narasimhan V, et al. Accuracy, sensitivity and specificity measurement of various classification techniques on healthcare data. *IOSR Journal of Computer Engineering (IOSR-JCE)*. 2013;11(5):70–73.
- [15] Iyer A, Jeyalatha S, Sumbaly R. Diagnosis of diabetes using classification mining techniques. *arXiv preprint arXiv:150203774*. 2015;.
- [16] Bhoi SK, et al. Prediction of diabetes in females of pima indian heritage: a complete supervised learning approach. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*. 2021;12(10):3074–3084.
- [17] Perdana A, Hermawan A, Avianto D. Analyze important features of pima indian database for diabetes prediction using knn. *Jurnal Sisfokom (Sistem Informasi dan Komputer)*. 2023;12(1):70–75.
- [18] Sivanesan R, Dhivya KDR. A review on diabetes mellitus diagnoses using classification on pima indian diabetes data set. *International Journal of Advance Research in Computer Science and Management Studies*. 2017;5(1).

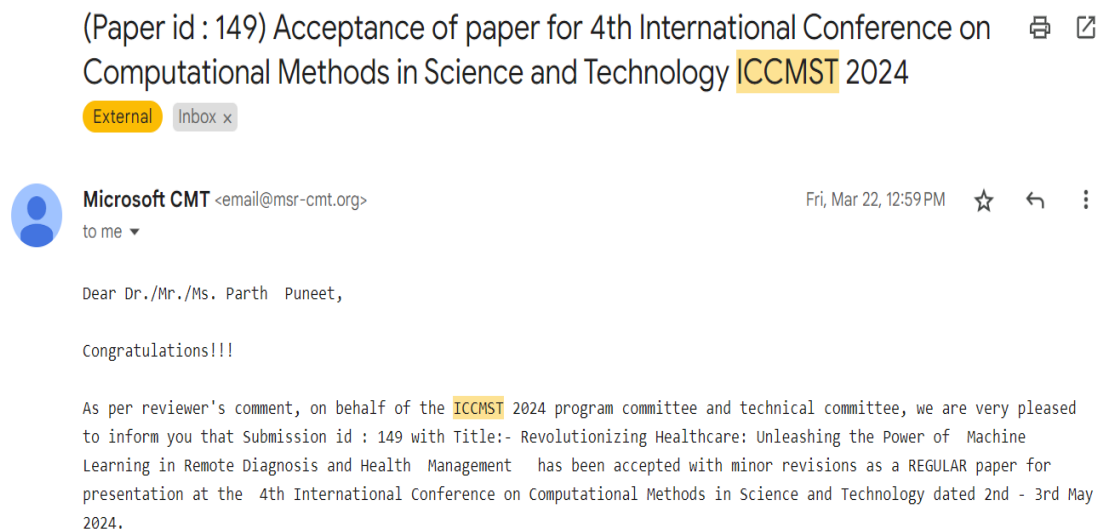
Title of Paper: Revolutionizing Healthcare: Unleashing the Power of Machine Learning in Remote Diagnosis and Health Management

Name of Conference: 4th International Conference on Computational Methods in Science and Technology

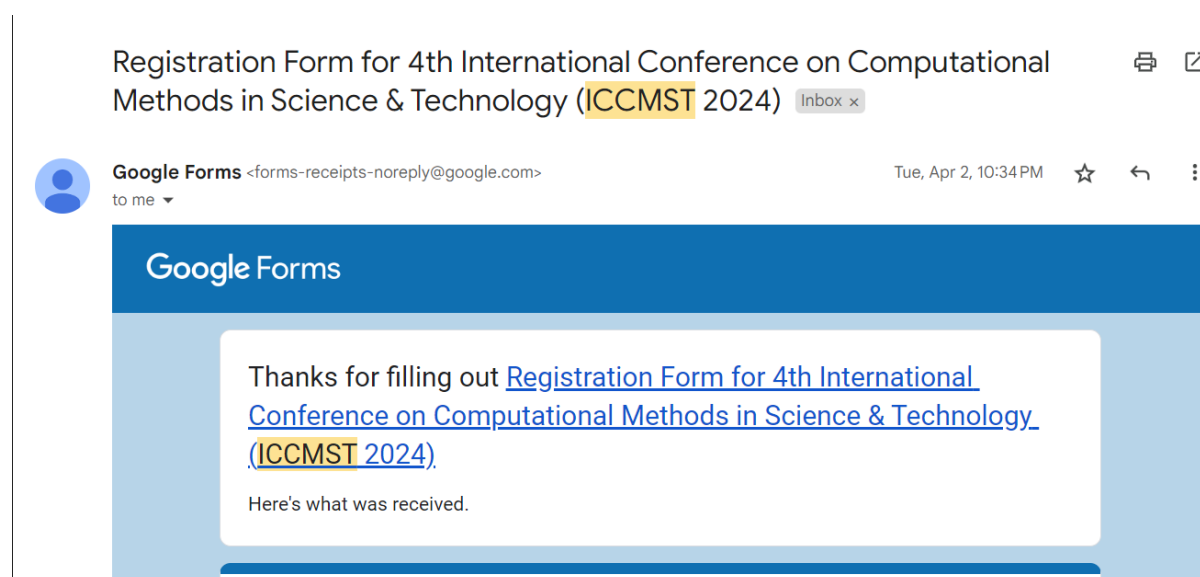
Date of Submission: 11th March 2024




Date of Acceptance: 22nd March 2024



Date of Registration : 2nd April 2024



Presentation: 2nd May 2024

 CHANDIGARH ENGINEERING COLLEGE CGC, LANDRAN, MOHALI <small>Building Careers. Transforming Lives.</small>	NAAC GRADE A+	CHANDIGARH ENGINEERING COLLEGE-CGC, LANDRAN DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING AND DEPARTMENT OF INFORMATION TECHNOLOGY
---	--------------------------	--

**Day-01
(May 2, 2024)**

TRACK-03 MACHINE LEARNING			
Time: 12:00 - 4:00 PM			
Session Co-chair: Dr. Ishpreet Singh Virk, Dept. of CSE, CEC-CGC, Landran		Session Co-chair: Dr. Rohit, Dept. of CSE, CEC-CGC, Landran	
Session Moderator: Ms. Sukhdeep Kaur & Ms. Apoorva Assistant Professor, Dept. of CSE, CEC-CGC, Landran		Session Moderator: Ms. Lakhwinder Kaur & Dr. Arwinder Kaur Assistant Professor, Dept. of IT, CEC-CGC, Landran	
SESSION LINK: https://meet.google.com/kct-bzop-hxn		SESSION LINK: https://meet.google.com/rdd-uyej-tfs	
12:00-1:30 PM	Paper presentation by Delegates/Author with paper Id's: 6 ,9, 10, 14, 18, 19, 20, 21, 23, 30, 36, 39, 51, 52, 71, 74, 76, 85, 87, 97	12:00-1:30 PM	Paper presentation by Delegates/Author with paper Id's: 98, 100, 104, 116, 129, 131, 135, 149, 151, 152, 160, 167, 169, 177, 179, 184, 185, 192, 193, 195
	At Venue: SMART CLASSROOM (111)		At Venue: RISE SMART CLASSROOM (Ground Floor)

[Paper ID: 149]

**Revolutionizing Healthcare: Unleashing the Power of Machine Learning in
Remote Diagnosis and Health Management**

Authors

[Author's Names: Parth Puneet, Prakhar Shukla, Gaurav Parashar]

[Author's Affiliation: Kiet Group Of Institution]

[Author's Email ID: parth.2024cse1011@kiet.edu]

[Paper Presenter Contact No.: 8505822345]

Page 15 of 15
ICCMST-2024

Presentation Certificate: 13th May 2024



Revolutionizing Healthcare: Unleashing the Power of Machine Learning in Remote Diagnosis and Health Management_Gaurav_parasar

by Gaurav Parasar

Submission date: 15-May-2024 09:16AM (UTC+0530)

Submission ID: 2379751194

File name: 149_CameraReadyPaper_Gaurav_parasar.pdf (154.48K)

Word count: 2941

Character count: 16351

Revolutionizing Healthcare: Unleashing the Power of Machine Learning in Remote Diagnosis and Health Management

Parth Puneet^a, Prakhar Shukla^b and Gaurav Parashar^c

^{a,b,c} KIET Group of Institutions, Delhi-NCR, Ghaziabad, India

ARTICLE HISTORY

Compiled April 2, 2024

ABSTRACT

The medical industry has increasingly used artificial intelligence (AI) as technology progresses. The primary advantage of machine learning comes in its capacity to derive significant insights from the immense volumes of health data gathered daily. The progress in the Internet of Things (IoT), machine learning, big data, and high-performance computing has created exceptional possibilities to enhance public health by maximizing scarce human resources. This work presents a diabetes prediction model that categorizes diabetes based on common parameters such as blood sugar, body mass index (BMI), age, insulin, and other relevant factors, as well as external factors that contribute to the development of diabetes. Our analysis utilizes the PIMA India and Syhlet datasets. Machine learning techniques, such as the Support Vector Machine (SVM), Naive Bayes classifier (NB), and decision tree, are used in the study.

KEYWORD

—Healthcare, Artificial Intelligence, Machine Learning, Diabetes Prediction

1. Introduction

A strong state of well-being lays the foundation for a satisfying life, unlocking pathways to joy and success. The health index of a nation serves as a reflection, indicating its economic strength, scientific achievements, defense capabilities ties, and social unity. While health is traditionally defined as the "absence of diseases," its true essence lies in the timely identification and anticipation of illnesses, followed by prompt and suitable interventions. This necessity becomes even more pronounced in a nation as diverse as India, where a mosaic of demographics, climate variations, and socio-cultural intricacies converge. Extending our focus to the national level, the health index becomes apparent as a complex indicator of a country's general resilience and vibrancy. A country's economic strength is reflected in the health index, which acts as a mirror to show patterns of wealth or inequality. A thriving economy makes investments in public health programs, healthcare infrastructure, and research possible, all enhancing well-being overall.

Diabetes is a prevalent, lifelong condition that affects individuals of all ages. It occurs when the level of glucose in the blood becomes too high. According to the

Email: parth.2021cse1011@kiot.edu

⁹ World Health Organization (WHO)¹, the number of people diagnosed with diabetes surged from 74 million in 1980 to 422 million in 2014, highlighting a significant global health issue. In 2019 alone, an estimated 2 million deaths were attributed to diabetes and related kidney diseases. For Type 1 Diabetes², risk factors include a family history, the condition and age, although currently, there's no known method preventing Type 1 Diabetes. In the case of Type 2 Diabetes, factors primarily include being overweight, being aged 45 years or older, having prediabetes, and having a family member with Type 2 Diabetes.

2. Literature Review

The experiment's primary goal is to design and implement a model for diabetes prediction that utilizes machine learning techniques successfully. Soni et al. [1] suggested ensemble learning and the use of classification techniques SVM, Random Forest (RF), Decision Tree (DT), Logistic Regression (LR), Gradient Boosting classifiers (GBC), and K-nearest neighbor (KNN). Furthermore, 77% classification accuracy was attained.

In their research, Xue et al. [2] use the NB as a classification technique. SVM achieved 96.54% accuracy. Tigga et al. [3] in the paper, implement six machine-learning classification strategies and compare the results with independent metrics. The tests were conducted on a dataset collected through online and offline surveys of 18 questions. The same statistics were also linked to the PIMA dataset (PIMA). Research results show that irregular timberland has an accuracy of 94.10%. Thomas and others [4] propose DT for diabetes prediction. For their research, they used the PIMA dataset. The accuracy of DT algorithms is examined and assessed in this work. The test results demonstrated that the devised system was effective, with an 87%. At small sample numbers, DT scales quickly and yields inaccurate response predictions. You et al. [5] conducted a study using two-class SVM and two-class DT to search for important diabetic factors in the PIMA dataset. It was found, through correlation analysis, that examining only the potential patient's glucose levels, BMI, and age is more efficient than conducting a comprehensive medical examination, which is time-consuming. The experiment with these features using SVM resulted in an accuracy of 70.4 %.

In the study, Rajni et. al [6] propose a procedure using the Bayes hypothesis for diabetes prediction utilizing machine learning methods to extract the necessary data so that the problem can be solved with high accuracy. Here, the mean is calculated to handle missing data and probability is used for calculating yes (positive) and no (negative) values. In comparison to previous controlled techniques, the suggested method enhances accuracy on the Pima dataset by 72.9%. In the study conducted by Ssodia D et al. [7], the Naive Bayes (NB), SVM, and DT algorithms are used among which The NB algorithm is considered the best machine learning method for this test because it has higher accuracy than other classification algorithms, with an accuracy of 76.30%. The study carried out by Mujumdar A et al. [8] suggests the use of numerous classification and common machine learning methods that yield the most precise results. Various classification methods include methods like LR, GBC, LDA, AdaBoost Classifier (ABC), Gaussian NB, Extra Trees Classifier (ETC), Bagging, RF, DT, Perceptron, SVC, and KNN on the PIMA dataset. Out of all, GBC, ETC, and ABC gave the highest accuracy of 77%.

¹<https://www.who.int/news-room/fact-sheets/detail/diabetes>

²<https://www.cdc.gov/diabetes/basics/risk-factors.html>

In the model, author Sara S et al. [9] employ ensemble methods by using various classifiers like DT, KNN, and NB on the PIMA dataset. Here, the highest prediction result of 78% was obtained using DT without bootstrap. Authors of [10] paper used 32 machine-learning algorithms used on the Pima Dataset for prediction namely DT, KNN, RF, NB, AB, LR, and SVM. For some criteria, including precision, accuracy, recall, and F measure, all models exhibit good results. Each model has an accuracy of 70%. For the train/test and test phases, LR and SVM yielded accuracies of 77% and 78%, respectively. Another neural network approach was used to predict diabetes in which the hidden layers in the network model were 1, 2, and 3, and occasionally 200, 400, and 800. In 400 epochs, the accuracy of the second hidden layer was 88.6%, the highest of all models used in PIDD. The research conducted by Zou et al. [11] uses two datasets for better diabetes prediction. The first dataset is from a hospital physical examination in Luzhou, China and the second dataset is Pima Indians. When RF is used, the accuracy for the Luzhou dataset is 80.84%, and for the Pima Indians dataset, the accuracy is 77.21%. Talha Mahboob and others [12] propose ANN for diabetes prediction. The accuracy of the ANN approach was 75.7%, the RF method was 74.7%, and the K-means clustering method was 73.6%. ANN outperforms other methods. Authors of paper [13] found that NB is better than the DT method J48. The results predicted that NB got 76.3021% accuracy followed by J48 with an accuracy of 73.8281%.

N.Gupta et al. [14] uses various classification techniques namely Multilayer Perception(MLP), J48, JRIP, and Bayes Network on Pima Indians dataset. The paper determined that J48 has the highest accuracy of 81.33%. A.Iyer et al. [15] proposed DT and NB algorithm for Diabetes Prediction on the PIMA dataset. Based on the cross-validation technique, the DT achieves an accuracy of 74.8698%. Based on the percentage split(70:30) technique, J48 gets 76.9565% accuracy and NB achieves an accuracy of 76.652%. The paper [16] predicted diabetes in females by taking the PIMA dataset. Various supervised learning algorithms have been used such as SVM, KNN, NB, RF, CT, NN, AB, and LR. When 10-fold cross-validation is applied, LR outperforms other techniques. The AUC for LR is 0.825. Aziz Perdana and others [17], in their research paper have used the KNN method for classification. By employing KNN with a value of k=22, they achieved the highest accuracy of 83.12%. R. Sivanesan et al. [18] analyzed the performance of the J48 Decision Tree with different metrics. Based on the training set, the accuracy of correctly classified was 84.11%. After using 10-fold cross-validation, the accuracy of correctly classified was 73.82%.

3. Methodology

In our study, we use two diabetes-related datasets from UCI and Kaggle. These datasets contained different numbers of features and multiple inconsistencies. We use the following steps to ascertain the best performance of the models.

- (1) Load and preprocess the data. Remove inconsistencies like missing values, outliers, inconsistent data and duplicate rows. Perform feature scaling and normalization. Split the data into training and test sets.
 - (2) Select the models and perform hyperparameter tuning to achieve the best results.
 - (3) Finalize the model based on the performance metric.
- Dataset Collection The module covers gathering data and preprocessing it for further analysis. For the study, two datasets were identified from the UCI repos-

itory.

Dataset 1: Dataset-1³ was gathered by doctors from patients from Sylhet Diabetes Hospital patients in Sylhet, Bangladesh. The dataset contains attributes namely age, gender, polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia, obesity, and class. The dataset contains 520 entries.

Dataset 2: For the research, dataset-2⁴ is the Pima Indians Diabetes dataset from kaggle. The dataset was obtained from the National Institute of Diabetes and Digestive and Kidney Diseases. Based on specific diagnostic metrics included in the collection, the dataset aims to predict diagnostically whether or not a patient has diabetes. These examples were chosen from a larger database under several restrictions. Specifically, all of the patients in this facility are Pima Indian women who are at least 21 years old. The dataset contains attributes: pregnancies, glucose, blood pressure, skin thickness, insulin, bmi, diabetes presence function, age, and outcome. This dataset has 768 entries, of which 268 are positive and the remaining 500 are negative for diabetes.

- Data Pre-Processing: The data pre-processing method improves results by handling superfluous data in the dataset in an efficient manner. But dealing with missing information is a major problem, especially regarding important variables like age, blood pressure, skinfold thickness, insulin, bmi, and blood sugar level. Since the values of these attributes must not have zero, therefore we used the imputation technique to remedy and preserve the integrity of the dataset. Our goal is to maintain data consistency and correctness by expanding the dataset to include all values, that enable more precise and significant analysis and forecasts.
- Model Selection The choice of a suitable machine-learning technique is crucial during this stage of creating a diabetes prediction model. Used different machine-learning methods, each with unique advantages and disadvantages, to accurately predict diabetes. SVM, DT, and NB are a few of the techniques used. The goal is to optimize both predictive accuracy and robustness while maintaining the consistency of the model's predictions by experimenting with different approaches.
 - (1) Support vector classifier: SVM is a supervised machine learning algorithm that is used for both classification and regression tasks. For learning instances, the maximal reserve hyperplane serves as its crucial limit. SVM uses the hinge loss function to calculate the empirical risk and adds a solution system regularization term to optimize the structural risk. SVM can do non-linear classification thanks to the kernel method, one of the most used kernel learning strategies.
 - (2) Decision Trees: A non-parametric supervised learning technique for regression and classification is called a DT. The goal is to create a model that can forecast the value of a target variable using fundamental decision rules inferred from the data attributes.
 - (3) Naïve Bayes: NB uses a set of algorithms based upon Bayes' theorem where the probability of previously occurred events is used to calculate the probability of posterior events. It is called naive as it takes the assumptions of features to be independent of each other. The classification model assigns class labels, represented by feature values, and the class labels are extracted

³<https://archive.ics.uci.edu/dataset/529/early+stage+diabetes+risk+prediction+dataset>

⁴<https://www.kaggle.com/datasets/ucml/pima-indians-diabetes-database>

from the dataset. For an item to be classified, it finds the maximum probability of each occurrence under the conditions that the item is considered as the item to be classified.

Table 1. Distinctive Paper Examination

Author	Algorithm	Dataset	Performance Metric
Soni [1]	RF	PIMA	77% classification accuracy
Xue [2]	SVM	Sylhet	Accuracy of 96.54%
Tigga [3]	RF	PIMA	Accuracy of 94%
Thomas [4]	Decision Tree	PIMA	Accuracy of 94%
S.You [5]	SVM	PIMA	Accuracy of 70.4%
Rajni [6]	RB-Bayes	PIMA	Accuracy of 72.9%
Sisodia [7]	Naive Bayes	PIMA	Accuracy of 76.30
Mujumdar [8]	Random Forest	PIMA	77% classification accuracy
Saru [9]	Decision Tree	PIMA	Accuracy of 78%
Khanam [10]	SVM	PIMA	Accuracy 77%-78%

The table 1 summarizes several studies using different machine learning algorithms to predict diabetes. It includes the author's name, the algorithm used, the dataset used, and its performance metrics, expressed as a percentage of accuracy. Algorithms exhibiting varying accuracy degrees on datasets include RF, SVM, DT, and NB. This highlights the significance of choosing algorithms depending on certain attributes. The table provides a quick comparative overview of the prediction performance of various algorithms in different papers.

4. Results & Conclusion

Among the various machine learning algorithms evaluated in our study, DT emerged as the most accurate model exhibiting a diabetes prediction accuracy of 98%, refer table 2. We utilized various machine learning algorithms such as SVM, DT, and NB on the two datasets. The dataset from the Pima Indian dataset contains fewer parameters. Consequently, due to this limitation, the model's performance is not satisfactory. Conversely, when dealing with the Sylhet dataset, which encompasses more parameters and accounts for both genders, the model achieved higher accuracies. The highest accuracy defines the effectiveness of DT in handling the intricacies of datasets and making precise predictions. The study compared the accuracy of a machine learning algorithm using two different datasets. This comparative analysis revealed that the developed model significantly enhanced the accuracy and precision of diabetes prediction across various datasets.

Table 2. Comparison between accuracy of PIMA Dataset and Sylhet Dataset

Algorithm	PIMA Dataset	Sylhet Dataset
SVM	77.2%	93.2%
Decision Tree	74%	98%
Naive Bayes	77.2%	90.3%

The model's adaptability to different data sources underscores its relevance to real-

world situations. Additionally, the study lays the groundwork for further extensions. The natural progression involves examining the risk of mortality within the next few years among individuals with diabetes. Expanding the analysis to include predictors of diabetes in individuals without the condition, can furnish valuable insights for health interventions. This broadening of the research scope holds promise for advancing diabetes prediction and prevention strategies.

References

- [1] Soni M, Varma S. Diabetes prediction using machine learning techniques. *International Journal of Engineering Research & Technology (Ijert)* Volume. 2020;9.
- [2] Xue J, Min F, Ma F. Research on diabetes prediction method based on machine learning. In: *Journal of Physics: Conference Series*; Vol. 1684; IOP Publishing; 2020. p. 012062.
- [3] Tigga NP, Garg S. Prediction of type 2 diabetes using machine learning classification methods. *Procedia Computer Science*. 2020;167:706–716.
- [4] Thomas J, Joseph A, Johnson I, et al. Machine learning approach for diabetes prediction. *International Journal of Information*. 2019;8(2).
- [5] YOU S, KANG M. A study on methods to prevent pima indians diabetes using svm. *Korean Journal of Artificial Intelligence*. 2020;8(2):7–10.
- [6] Rajni R, Ananddeep A. Rb-bayesalgorithm for the predictionof diabetic in pima indian dataset. *International Journal of Electrical and Computer Engineering*. 2019;9(6):4866.
- [7] Sisodia D, Sisodia DS. Prediction of diabetes using classification algorithms. *Procedia computer science*. 2018;132:1578–1585.
- [8] Munjundar A, Vaidehi V. Diabetes prediction using machine learning algorithms. *Procedia Computer Science*. 2019;165:292–299.
- [9] Saru S, Subashree S. Analysis and prediction of diabetes using machine learning. *International journal of emerging technology and innovative engineering*. 2019;5(4).
- [10] Khanam JJ, Foo SY. A comparison of machine learning algorithms for diabetes prediction. *Iet Express*. 2021;7(4):432–439.
- [11] Zou Q, Qu K, Luo Y, et al. Predicting diabetes mellitus with machine learning techniques. *Frontiers in genetics*. 2018;9:515.
- [12] Alam TM, Iqbal MA, Ali Y, et al. A model for early prediction of diabetes. *Informatics in Medicine Unlocked*. 2019;16:100204.
- [13] Diwani SA, Sam A. Diabetes forecasting using supervised learning techniques. *Adv Comput Sci an Int J*. 2014;3:16–18.
- [14] Gupta N, Rawal A, Narasimhan V, et al. Accuracy, sensitivity and specificity measurement of various classification techniques on healthcare data. *IOSR Journal of Computer Engineering (IOSR-JCE)*. 2013;11(5):70–73.
- [15] Iyer A, Jeyalatha S, Sambaly R. Diagnosis of diabetes using classification mining techniques. *arXiv preprint arXiv:150203774*. 2015;.
- [16] Bhoi SK, et al. Prediction of diabetes in females of pima indian heritage: a complete supervised learning approach. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*. 2021;12(10):3074–3084.
- [17] Perdana A, Hermawan A, Avianto D. Analyze important features of pima indian database for diabetes prediction using knn. *Jurnal Sidelokom (Sistem Informasi dan Komputer)*. 2023;12(1):70–75.
- [18] Sivanesan R, Dhivyaa KDR. A review on diabetes mellitus diagnoses using classification on pima indian diabetes data set. *International Journal of Advance Research in Computer Science and Management Studies*. 2017;5(1).

Revolutionizing Healthcare: Unleashing the Power of Machine Learning in Remote Diagnosis and Health Management_Gaurav_parasar

ORIGINALITY REPORT

20%

SIMILARITY INDEX

13%

INTERNET SOURCES

12%

PUBLICATIONS

12%

STUDENT PAPERS

PRIMARY SOURCES

1	studenttheses.uu.nl Internet Source	2%
2	Submitted to King's Own Institute Student Paper	1%
3	Submitted to Hong Kong Baptist University Student Paper	1%
4	turcomat.org Internet Source	1%
5	Submitted to University of Hertfordshire Student Paper	1%
6	Submitted to University of Portsmouth Student Paper	1%
7	Minshui Huang, Yongzhi Lei, Xifan Li, Jianfeng Gu. "Damage Identification of Bridge Structures Considering Temperature Variations-Based SVM and MFO", Journal of Aerospace Engineering, 2021 Publication	1%

8	Submitted to UC, Boulder Student Paper	1 %
9	climateerinvest.blogspot.com Internet Source	1 %
10	Submitted to Asia Pacific University College of Technology and Innovation (UCTI) Student Paper	1 %
11	fritz.ai Internet Source	1 %
12	Bashar Hamad Aubaidan, Rabiah Abdul Kadir, Mohamad Taha Ijab. "Chapter 45 Enhancing Diabetes Prediction and Classification Using the Bidirectional Neighbor Graph Algorithm", Springer Science and Business Media LLC, 2024 Publication	1 %
13	Quan Zou, Kaiyang Qu, Yamei Luo, Dehui Yin, Ying Ju, Hua Tang. "Predicting Diabetes Mellitus With Machine Learning Techniques", Frontiers in Genetics, 2018 Publication	1 %
14	Submitted to South University Student Paper	1 %
15	www.researchgate.net Internet Source	1 %
16	Submitted to Aspen University	

	Student Paper	1 %
17	Submitted to University at Buffalo Student Paper	1 %
18	2021.igem.org Internet Source	1 %
19	www.sciencegate.app Internet Source	<1 %
20	Gita Mahmoudabadi, Kelsey Homyk, Adam Catching, Ana Mahmoudabadi, Helen Foley, Arbel D. Tadmor, Rob Phillips. "Machine learning models can identify individuals based on a resident oral bacteriophage family", Cold Spring Harbor Laboratory, 2024 Publication	<1 %
21	ijrps.com Internet Source	<1 %
22	www.frontiersin.org Internet Source	<1 %
23	Ludmil Dakovski. "Learning and classification with prime implicants applied to medical data diagnosis", Proceedings of the 2007 international conference on Computer systems and technologies - CompSysTech 07 CompSysTech 07, 2007 Publication	<1 %

24	bsj.uobaghdad.edu.iq Internet Source	<1 %
25	es.slideshare.net Internet Source	<1 %
26	indjst.org Internet Source	<1 %
27	ojs.wiserpub.com Internet Source	<1 %
28	www.mdpi.com Internet Source	<1 %
29	"Intelligent Internet of Things for Healthcare and Industry", Springer Science and Business Media LLC, 2022 Publication	<1 %
30	Abeer El-Sayyid El-Bashbishy, Hazem M. El-Bakry. "Pediatric diabetes prediction using deep learning", Scientific Reports, 2024 Publication	<1 %
31	Anjali Jain, Alka Singhal. "Bio-inspired Approach for Early Diabetes Prediction and Diet Recommendation", SN Computer Science, 2024 Publication	<1 %
32	Bhuvaneswari Amma N.G.. "En-RfRsK: An ensemble machine learning technique for	<1 %

prognostication of diabetes mellitus",
Egyptian Informatics Journal, 2024

Publication

33	hrcak.srce.hr Internet Source	<1 %
34	jnas.nbu.gov.ua Internet Source	<1 %
35	"Third International Conference on Image Processing and Capsule Networks", Springer Science and Business Media LLC, 2022 Publication	<1 %
36	"Communication and Intelligent Systems", Springer Science and Business Media LLC, 2022 Publication	<1 %
37	"Healthcare Transformation with Informatics and Artificial Intelligence", IOS Press, 2023 Publication	<1 %

Exclude quotes Off

Exclude matches < 5 words

Exclude bibliography On