VISVESVARAYA TECHNOLOGICAL UNIVERSITY
Jnana Sangama, Belagavi – 590018

2022 - 2023

A Project Report On

"PREDICTION OF DIABETES THROUGH MEDICAL DATASET USING MACHINE LEARNING"

Submitted in partial fulfillment of the requirements for the award of degree of

BACHELOR OF ENGINEERING

In

INFORMATION SCIENCE AND ENGINEERING

Submitted by

AYUSH KUMAR (1AT19IS021)
HARSHITHA V (1AT19IS038)
INCHARA A (1AT19IS043)

Under the Guidance of
Mr. OMPRAKASH B
Assistant Professor,
Dept. Of ISE
Atria Institute of Technology

ATRIA
INSTITUTE OF TECHNOLOGY

ATRIA INSTITUTE OF TECHNOLOGY
DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING

Anandnagar, Bengaluru-560 024

ATRIA INSTITUTE OF TECHNOLOGY
Anandnagar, Bengaluru 560024

Department of Information Science and Engineering

CERTIFICATE

This is to certify that the work entitled **“PREDICTION OF DIABETES THROUGH MEDICAL DATASET USING MACHINE LEARNING”** carried out by **AYUSH KUMAR (1AT19IS021) , HARSHITHA V (1AT19IS038) and INCHARA A (1AT19IS043)**, are bonafide student of **ATRIA INSTITUTE OF TECHNOLGY**, Bengaluru, in partial fulfilment for the award of the **Bachelor of Engineering in Information Science & Engineering of Visvesvaraya Technological University**, Belagavi, during the academic year 2022-2023. It is certified that all corrections/suggestions indicated for Internal Assessment have been incorporated in the Report deposited in the departmental library. The project report has been approved as it satisfies the academic requirements in respect of Project work prescribed for the said Degree.

Signature of the Guide	**Signature of the HOD**	**Signature of the Principal**
Mr. Omprakash B	Dr. Shanthi Mahesh	Dr. Y Vijaya Kumar
Assistant Professor	Head of Department	Principal
Dept. of ISE, Atria IT	Dept. of ISE, Atria IT	Atria IT

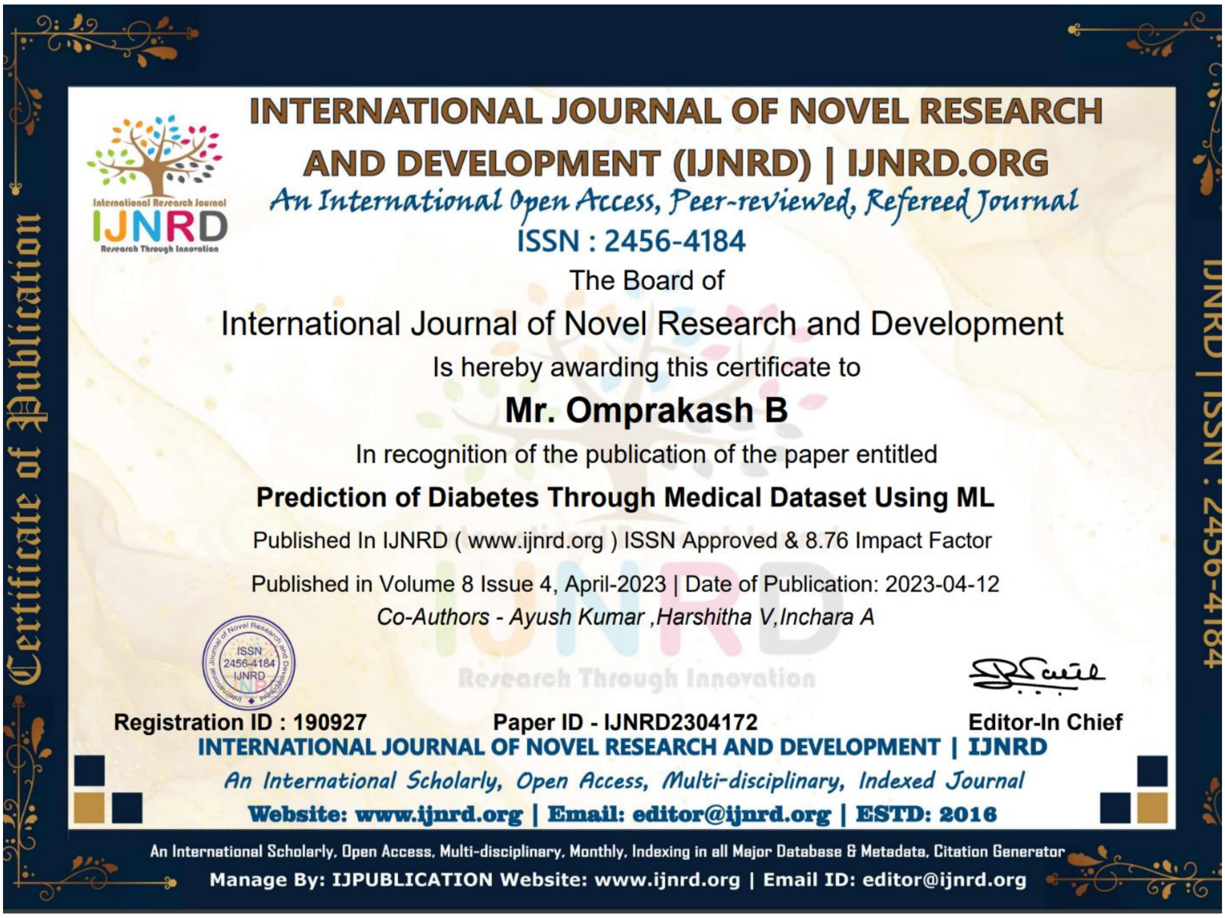
External Viva

Name of the Examiner **Signature with Date**

1. ______________________ ________________________

2. ______________________ ________________________

PUBLICATION CERTIFICATE



Certificate of Publication

International Research Journal
IJNRD
Research Through Innovation

INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG

An International Open Access, Peer-reviewed, Refereed Journal

ISSN : 2456-4184

The Board of

International Journal of Novel Research and Development

Is hereby awarding this certificate to

Mr. Omprakash B

In recognition of the publication of the paper entitled

Prediction of Diabetes Through Medical Dataset Using ML

Published In IJNRD (www.ijnrd.org) ISSN Approved & 8.76 Impact Factor

Published in Volume 8 Issue 4, April-2023 | Date of Publication: 2023-04-12

Co-Authors - Ayush Kumar ,Harshitha V,Inchara A

International Journal of Novel Research and Development ISSN 2456-4184 IJNRD

Registration ID : 190927 **Paper ID - IJNRD2304172** **Editor-In Chief**

INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT | IJNRD

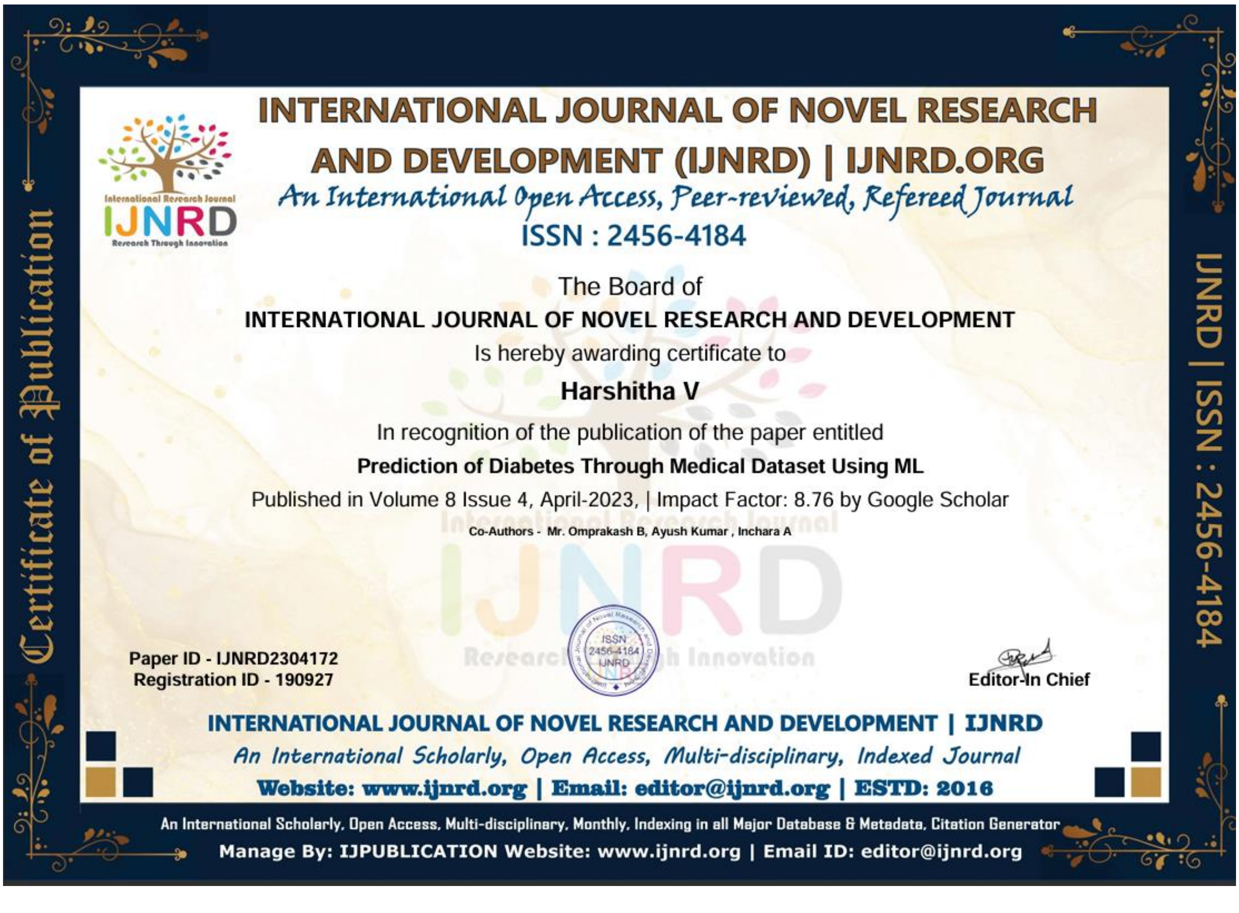
An International Scholarly, Open Access, Multi-disciplinary, Indexed Journal

Website: www.ijnrd.org | Email: editor@ijnrd.org | ESTD: 2016

An International Scholarly, Open Access, Multi-disciplinary, Monthly, Indexing in all Major Database & Metadata, Citation Generator

Manage By: IJPUBLICATION Website: www.ijnrd.org | Email ID: editor@ijnrd.org

IJNRD | ISSN : 2456-4184



Certificate of Publication

International Research Journal
IJNRD
Research Through Innovation

INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG

An International Open Access, Peer-reviewed, Refereed Journal

ISSN : 2456-4184

The Board of

INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT

Is hereby awarding certificate to

Harshitha V

In recognition of the publication of the paper entitled

Prediction of Diabetes Through Medical Dataset Using ML

Published in Volume 8 Issue 4, April-2023, | Impact Factor: 8.76 by Google Scholar

Co-Authors - Mr. Omprakash B, Ayush Kumar , Inchara A

Paper ID - IJNRD2304172
Registration ID - 190927

ISSN 2456-4184 IJNRD

Editor-In Chief

INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT | IJNRD

An International Scholarly, Open Access, Multi-disciplinary, Indexed Journal

Website: www.ijnrd.org | Email: editor@ijnrd.org | ESTD: 2016

An International Scholarly, Open Access, Multi-disciplinary, Monthly, Indexing in all Major Database & Metadata, Citation Generator

Manage By: IJPUBLICATION Website: www.ijnrd.org | Email ID: editor@ijnrd.org

IJNRD | ISSN : 2456-4184

DECLARATION

We, **Ayush Kumar (1AT19IS021), Harshitha V (1AT19IS038) , and Inchara A (1AT19IS043)** hereby declare that the project entitled **“Prediction of diabetes through medical dataset using machine learning”** is carried out under the guidance of **Mr. Omprakash B**, AssistantProfessor, Department of Information Science and Engineering. The project work is submitted to Visvesvaraya Technology University in partial fulfillment of the requirementsfor the Bachelor of Engineering in Information Science and Engineering award for the academic year 2022-2023.

Place: Bengaluru
Date:

Signature of the Students

Ayush Kumar (1AT19IS021)

Harshitha V (1AT19IS038)

Inchara A (1AT19IS043)

ACKNOWLEDGEMENT

The foundation for any successful venture is laid out not just by the individual accomplishment the task, but also by several other people who believe that the individual can excel and put in their every bit in every endeavor he/she embarks on, at every stage in life. And the success is derived when opportunity meets preparation, also support by a well-coordinator approach and attitude.

We would like to express my sincere gratitude to the respected principal **Dr. Y Vijaya Kumar**, for providing a congenial environment to work in. I also like to express my sincere gratitude to **Dr. Shanthi Mahesh**, Prof & Head, Department of Information Science & Engineering, for her continuous support and encouragement.

We take this opportunity to express our deep sense of gratitude to our guide **Mr. Omprakash B**, Asst. Professor, Department of Information Science & Engineering, for his valuable guidance and help throughout the course of the academic project.

We take this opportunity to express our deep sense of gratitude to Coordinators **Dr. K S Ananda Kumar**, and **Mr. Srinivas B V** for their continued support, advise and valuable input during the project phases.

Last, but not the least we would like to thank my family, who has acted as a beacon of light throughout my life. My sincere gratitude goes out to all my comrades and well-wishers who have supported me though all the venture.

ABSTRACT

Data mining is the process of looking at data from multiple perspectives and combining them with desired data. It is about discovering knowledge or knowledge. Among the many software tools for data analysis, data mining is the most widely used. This allows users to evaluate data from multiple perspectives and dimensions, and group and save relationships. Technically, data mining can be thought of as a step to follow in searching for patterns or analyzing relationships between different sources in large datasets. Current developments in data mining and machine learning are improving the conditions of primary health care by improving research in the field of biomedicine. Regular recording is essential. New medical devices and technologies for diagnosis create mixed data and big data. Therefore, to deal with this poor biomedical data, intelligent data mining and machine learning methods are required to generate demand from the collected raw data calculated as medical data mining. In medical records, medical records only look for patterns and associations that can provide important information for an accurate diagnosis. This technology is used in many medicines (medical applications) and helps to improve diagnosis. Accuracy of classification of medical data and estimation of its value are the main tasks/challenges of medical data mining. Better classifications are needed to improve the predictive value of additional clinical data, as misclassifications can lead to poor estimates. When medical information is used only for medical information, the basic and difficult problems are classification and prediction. Artificial neural network (ANN) and logistic regression (LR) are often used to perform these functions. In our presented research, a hybrid data mining model is proposed for classifying and estimating medical data using LR and ANN, a cross-validated model (CVS) and a percentage selection method (FSM). The performance of the proposed hybrid model will be evaluated based on classification accuracy.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

Sl. No.	Acronym	
1	ANN	Artificial Neural Network
2	CVS	Cross Validation Sample
3	FSM	Finite State Machine
4	LR	Logistic Regression
5	CDC	The centre for disease control and prevention
6	FMDP	Fused model for diabetes prediction
7	SVM	Support Vector Machine
8	ML	Machine Learning
9	T2DM	Type 2 Diabetes Mellitus
10	BMI	Body Mass Index
11	NB	Naïve Bayes
12	NN	Neural Network
13	DL	Deep learning
14	AI	Artificial Intelligence
15	DM	Diabetes Mellitus
16	UMASS	University of Massachusetts
17	RF	Random Forest
18	UML	Unified Modelling Language
19	FS	Function Specification
20	ARFF	Attribute Relation File Format
21	UCI	University of California, Irvine

CHAPTER 1

INTRODUCTION

Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. Insulin is a hormone that regulates blood glucose. Hyperglycemia, also called raised blood glucose or raised blood sugar, is a common effect of uncontrolled diabetes and over time leads to serious damage to many of the body's systems, especially the nerves and blood vessels.

In 2014, 8.5% of adults aged 18 years and older had diabetes. In 2019, diabetes was the direct cause of 1.5 million deaths and 48% of all deaths due to diabetes occurred before the age of 70 years. Another 460 000 kidney disease deaths were caused by diabetes, and raised blood glucose causes around 20% of cardiovascular deaths.

Data Mining

Data mining is nothing but the process of viewing data in different angle and compiling it into appropriate information. It can also be referred to as knowledge or data discovery. Out of the many software tools used for data evaluation, the one which is widely used is the data mining. This allows the user to evaluate the data from distinct angles and dimensions, grouping and summarizing the identified relations. Technically the data mining can be considered as the sequence of steps followed for searching patterns or identifying correlations among large numbers of fields within a huge relational database.

Data

Data is nothing but the fact, wording or numbers which can be handled by the computer systems. Many organizations are involved in collecting bulky and ever-increasing amount of data in various formats (types) and in different databases . These include:

- Business/Operational data such as cost, selling, pay sheets, accounting and stock.
- Non-operational data such as macro economical data, forecast data and industry sales.
- Meta data – i.e. data within data like data dictionary and logical database design.

Information

The patterns, relationships or associations between all these *data* are called *information*. For example, the information related to what products are sold is obtained by analyzing retail sale transaction data.

Knowledge

The past information related to patterns and future trends can be transformed into knowledge. For example, by analyzing the detailed information of retail supermarket, a retailer or manufacturer can determine the items that are suitable for promotional efforts.

Data Warehouses

Due to the sudden advances in capturing the data, transmission of data, processing power, and storage efficiency of data, allows the organizations to combine different databases to data warehouses. The mechanism of managing and retrieval of data from centralized repositories is called data warehousing.

Components of Data Mining

The five major components of data mining are:

1. Collecting, converting and storing business data within a data warehouse.
2. Storing and handling the data within a multidimensional db system.
3. Allowing IT professionals and business analysts to access data.
4. Using application software for analyzing the data.
5. Representing data in multiple data formats like graph or table.

Applications of Data Mining

Data mining is used in a variety of applications such as future healthcare, market analysis, learning, manufacture engineering, detecting fraud, detecting unwanted entries, detection of lies, customer division, financial banking, corporate surveillance, analysis of research, criminal investigation, retail industry, bio informatics, and telecommunication industry.

Techniques Used in Data Mining

Data mining includes the following major techniques:

1. Association or relation: This is the well known, the most familiar and the straight forward technique in data mining. In this, we find the relationship among multiple objects of same form to recognize the patterns.
2. Classification: This may be used for building a concept of the category of item by defining multiple attributes for determining a particular class.
3. Clustering: In this, individual portions of data are grouped mutually to create a composition. In this, single or multiple attributes are used as a basis for creating the cluster.
4. Prediction: This is often used in combination of other techniques of data mining. This usually involves analyzing classification, pattern matching, and relation. Predictions can be made by analyzing the previous events.

5. Sequential Patterns: These are useful approach for determining trends and routine instances of identical events.
6. Decision trees: These are connected to other methods like classification and prediction, and may be used as a component in selection or to supplement the selection and use of exact information within overall composition (structure).

Medical Data Mining

Recent improvements in the area of data mining and machine learning have empowered the research in biomedical field to improve the condition of general health care. In many parts of the world the tendency for maintaining long-lasting records consisting of medical data is becoming an accepted practice. In addition to this, the newer medical equipments and the techniques used in diagnosis, produces composite and huge data. Therefore, to handle these ill-structured biomedical data, intelligent algorithms for data mining and machine learning are required in order to take logical reasoning from the saved raw data, which is considered as medical data mining. Within the medical data, the medical data mining searches for patterns and relationships which can provide useful information for appropriate medical diagnosis [6]. Data mining techniques are applied to different medical domains (health care databases or medical datasets) to improve the medical diagnosis.

To check for any invisible patterns inside the medical datasets, medical data mining is strongly recommended. In medical data mining, the actual tasks (challenges) are the classification and prediction of medical datasets. To manage these tasks, Artificial Neural Networks (ANN) and Logistic Regression (LR) are frequently used.

Challenges in Medical Data Mining

There are various challenges to be addressed by medical data mining with respect to data stored in healthcare databases. Many researchers are trying to address these challenges :

1. One of the challenges is format of medical data. The format in which the medical data is stored in different databases is different since there is no standard format.
2. To extract meaningful information, the quality of medical data is significant. The medical data available in real world is noisy, heterogeneous, redundant, and incomplete. The medical data available in large quantities may often be unreliable. These problems could be due to errors in the machines in which the medical data is measured or because of the mistakes by people/users. Even the medical data could change due to system or human errors. All these lead to incomplete and noisy data which really challenge the medical data mining model. All the necessary steps should be taken to maintain the quality of medical data.

3. Sharing of medical data is another important challenge. Neither the health care organizations nor the patients like to share their personal information.
4. The way in which the data warehouses are built to share the medical data is expensive and takes more time. Usually the medical data available in real world is stored in distributed computing environment on various platforms. It can be on the internet, databases or on individual systems. Practically, it is very difficult to collect all the medical data to a single centralized repository because of the technical reasons and the organizational limitations. Hence, there is a demand for developing algorithms and tools which allows for mining distributed medical data.
5. If there is a huge amount of records and attributes in dataset then handling these datasets will be difficult. In addition to this, if the medical data is heterogeneous i.e. the medical data may contain images, audio and video, complex data, temporal data, spatial data, time series, natural language text and so on, then, it is very difficult to handle these datasets. New methodologies and tools need to be developed to mine the required information.
6. Since the medical field is dynamic, the medical information evolves endlessly because of this; the equivalent medical actions and process vary as well.
7. The medical data mining model's performance mainly depends on the techniques used and the efficiency of algorithms. Improper use of the techniques and algorithms may affect the performance of the medical data mining process.
8. Improving the accuracy of the classification and improving the prediction rate of medical datasets is the main task/challenge of medical data mining. Since the wrong classification may lead to poor prediction, there is a need to perform the better classification which further improves the prediction rate of the medical datasets. When medical data mining is applied to the medical datasets, the important and difficult challenges are the classification and the prediction.

1.1 Motivation

Diabetes is the fast-growing disease among the people even among the youngsters. About 422 million people worldwide are suffering from diabetes, the majority living in low-and middle-income countries, and 1.5 million deaths are directly attributed to diabetes each year. Both the number of cases and the prevalence of diabetes have been steadily increasing over the past few decades. In understanding diabetes and how it develops, we need to understand what happens in the body without diabetes. Sugar (glucose) comes from the foods that we eat, specifically carbohydrate foods Carbohydrate foods provide our body with its main energy source everybody, even those people with diabetes, needs carbohydrate . Carbohydrate foods include bread , cereal , pasta , rice , fruit , dairy products and vegetables (especially starchy vegetables). When we eat these foods, the body breaks them down into glucose. The glucose moves around the body in the bloodstream. Some of the glucose is taken to our brain to help us think clearly and function. The remainder of the glucose is taken to the cells of our body for energy and also to our liver, where it is stored as energy that is used later by the body. In order for the body to use glucose for energy, insulin is required. Insulin is a hormone that is produced by the beta cells in the pancreas. Insulin works like a key to a door. Insulin attaches itself to doors on the cell, opening the door to allow glucose to move from the blood stream, through the door, and into the cell. If the pancreas is not able to produce enough insulin (insulin deficiency) or if the body cannot use the insulin it produces (insulin resistance), glucose builds up in the bloodstream (hyperglycaemia) and diabetes develops . In this paper we aim to develop a prediction system using machine learning to detect and classify the presence of diabetes in e-healthcare environment using Ensemble Decision Tree Algorithms for high feature selection. A significant attention has been made to the accurate detection of diabetes which is a big challenge for the research community to develop a diagnosis system to detect diabetes in a successful way in the e healthcare environment. The existing diagnosis systems have some drawbacks, such as high computation time, and low prediction accuracy. To handle these issues, we have proposed diagnosis system using machine learning methods, such as pre-processing of data, feature selection, and classification for the detection of diabetes disease in e-healthcare environment. Model validation and performance evaluation metrics have been used to check the validity of the proposed system. We have proposed a filter method based on the Decision Tree algorithm for highly important feature selection.

Two ensemble learning Decision Tree algorithms, such as Ada Boost and Random Forest are also used for feature selection and compared the classifier performance with wrapper-based feature selection algorithms also. Machine learning classifier Decision Tree has been used for the classification of healthy and diabetic subjects. The experimental results show that the Decision Tree algorithm based on selected features improves the classification performance of the predictive model and achieved optimal accuracy. Additionally, the proposed system performance is high as compared to the previous state-of-the-art methods. High performance of the proposed method is due to the different combinations of selected features set. Furthermore, the experimental results statistical analysis demonstrated that the proposed method would be effectively detected diabetes disease.

Common Techniques Used in Medical Data Mining for Classification and Prediction

Many techniques are used for the classification and prediction of medical datasets. The most common techniques used for handling these challenges are LR and ANN.

- **Logistic Regression**

LR is one in every of the data mining ways used for analyzing issues wherever the end result is determined based on one or additional variables. A dichotomous variable is used to measure the outcome. In LR, the non-independent variable is dichotomous or binary i.e., it consists of data represented as 0 (FALSE, failure, etc.) or as 1 (TRUE, success, etc.). In various biomedical fields such as cancer analysis, survival forecast, kidney transplant etc. [9] [10], LR has been widely used. Even in statistics, it is a well-established and a powerful method. It is suggested that LR has to be compared to data mining techniques while performing medicinal data mining .LR is implemented on the health care databases for detecting the patterns which are useful for either forecasting or determining the diseases along with take the remedial measures for handling such diseases.

- **Artificial Neural Networks**

The human brain architecture is the main inspiration behind the development of the model. ANNs are successfully used in various disciplines such as environmental science, study of human mind, study of numbers, study of medicine. ANNs are also being used in many commercial regions like accounts and audits, funding, managing and decision making, promotion and manufacture etc. ANN models or “neural nets” are also called by different names. Whatever the name is; each one of these models tries to give good performance through compact interconnection of uncomplicated computational elements. For many years these models have been studied with a hope of achieving the performance like humans in the field of speech and image recognition .

1.2 Existing System

In present day scenario multiple diabetes prediction systems are available with varying levels of efficiencies. Present system we are facing some of the issues in the collection of the patient data which in the real time is very difficult, and also the design is ineffective in order to treat the treat the diabetes in a effective manner. There is no comprehensive sensing and analysis for the patients suffering from the diabetes.

Advantages

- This helps in the Cognitive intelligence towards the patient’s status and network resources.

- Provides the information about the early detection of the diabetes .

Disadvantages

- System is not comfortable **.**
- Collection of the data in the real time is very difficult.
- This system also lacks the monitoring of the psychological indications of diabetic patients. Absence of sharing of data mechanism and analysis in personal form of data from different resources which includes the sports, diet and also the lifestyle of the patients

1.3 Proposed System

This work mainly concentrates to obtain the better classification accuracy with less number of the given attributes with which we can lessen the total time needed time for prediction and also it helps in the improvement of the accuracy in the classification.

The classification of real time data streams is a challenging task in data mining. To manage these problems for classifying data streams an ensemble classifier is developed. For mining the noisy structures in data streams, decision trees were used. The ensemble model was dynamic that is it automatically updates and it represents the latest concepts of the data streams. During each iteration, a new data for training is developed from the original dataset used for training.

The different steps in the process of knowledge discovery in general medical data mining

The Steps involved in medical data mining process are:

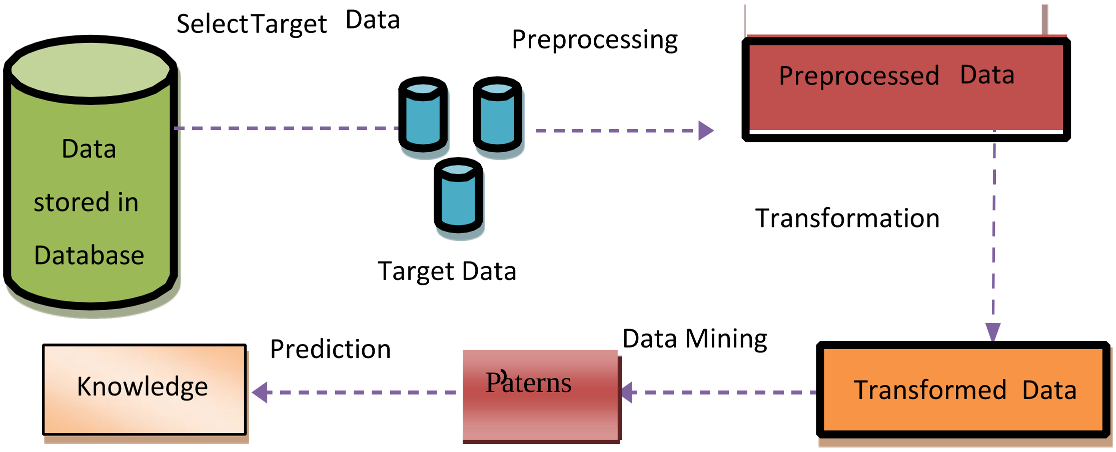


Figure 1.1: Steps in Knowledge Discovery Process in Data Mining

Step 1: Databases

The main repository of data is the database. A huge quantity of information is stored in large and different databases. Because of the digital revolution, there exists a variety of inexpensive ways togather and pile up huge amount of patient records and their medical conditions within a database. Internet for use in healthcare services globally . Newer diagnostic methods can be discovered by extracting useful information from these health care databases.

Step 2: Selecting target data

Selecting the target data, health care dataset consists of huge amount of data of different types, from which I have to select the data according to our requirement i.e. since I require medical data, I select the medical datasets. Since the clinical medical data may contain systemic and human errors, it has to be corrected during reprocessing

Step 3: Pre-processing of medical data

Before doing the classification and prediction, the data is pre-processed for any missing, noisy, invalid and inconsistent attributes. Therefore, prior to applying data mining methods, the data is to be pre-processed and the steps involved in pre-processing are:

- Data Integration: If the information is received from multiple sources, then the information is to be integrated, which consists of removing any inconsistencies in the attribute name or attribute value.
- Discretization: Discretization is used when the data mining algorithms cannot handle continuous attributes. Discretization converts continuous attributes to categorical attributes by considering few discrete values.
- Attribute Selection: Generally the medical datasets include a huge quantity of attributes, out of which a few may not be used at the time of prediction. Hence Feature Selection Methods (FSMs) is applied on these medical data to select the important attributes.

Step 4: Transformation of data

After the pre-processing of the selected medical data is done, I transform the medical data to a unified format by reducing the features or dimensionality of the datasets. The dimension in the datasets will be reduced using the different FSMs.

Step 5: Applying data mining methods

After transforming the data to the required format, I apply different techniques of data mining to identify the valuable data, knowledge and information. This helps in searching the patterns of our interest.

Step 6: Predictions

Finally after identifying the useful patterns, I extract the useful knowledge from these patterns. Prediction is nothing but estimating the uniqueness of a single pattern depending on the explanation of another associated pattern. Predictions are done mainly based on the relationship between a pattern I know and a pattern that is to be predicted. Usually for classification and prediction different techniques like decision trees, Bayesian classifier, LR, ANN, genetic algorithms, adaptive euro fuzzy interfaces etc. are available.

1.4 Objectives

- To increase the accuracy in future prediction of diabetes using less number of variables.
- To increase the awareness related to the diabetes and the various parameters which can be a clear indication of the future onset of diabetes.
- To be able to make a prediction in real time with high precision.
- To keep the entire process fast paced so that people can act early upon their diagnosis.

1.5 Features with Scope

- The system and the approach is cost effective and more comfortable.
- The personalization of the assorted machine learning and cognitive feature algorithms to established customized treatment for the patients.
- Adjusts in the treatment ideology in time based and status of the patients with the continuous storing and collecting of the information of the patients.
- It is able to make predictions in real time based on the data provided by the patient or related to patient.

1.6 Limitations

There are four main types of diabetes: type 1, type 2, gestational diabetes (diabetes in pregnancy), and prediabetes.

Type 1 Diabetes

This type is usually diagnosed in kids, teens, and young adults, but it can happen at any age. Type 1 diabetes occurs when your pancreas doesn’t make insulin. This means you have to take insulin every day. The Centers for Disease Control and Prevention (CDC) estimates that around 5% to 10% of people with diabetes have this type.

Type 2 Diabetes

This type can also show up at any age, but it’s more common if you’re over 40. Type 2 diabetes occurs when your pancreas doesn’t make enough insulin, or your body isn’t using the insulin well. Around 90% to 95% of people with diabetes have this type. While it has historically affected mainly adults, the rate of type 2 diabetes in children and adolescents is rising.

Gestational Diabetes (Diabetes in Pregnancy)

During pregnancy, some women who did not previously have diabetes develop it. This is called gestational diabetes. It usually disappears after the baby is born, but having gestational diabetes increases both your and your baby’s risk of developing type 2 diabetes later on.

Prediabetes

As the name suggests, prediabetes increases your risk of developing type 2 diabetes. In this stage, your blood sugar levels are higher than they should be, but not high enough to be diagnosed with type 2. The CDC says that 96 million adults in the United States have prediabetes. That’s more than a third of adults. Unfortunately, more than 84% don’t know they have it.

While our proposed system is capable of detecting all these types of diabetes in advance, it is not capable of identifying the type of diabetes which the patient is currently suffering from.

1.7 Organization of Report

In this report submitted it has been carefully divided into seven chapters for better understanding and organization. They are Introduction, Literature Survey, System Requirement Specification, System Design, Implementation & Testing, Results of Discussion, Conclusion. The work propose a new hybrid data mining model for selecting the important attributes based on Entropy Evaluation method, Mean Evaluation method, and Threshold Evaluation method on Pima Indian Diabetes dataset, Bupa Diabetes Disorder dataset, and Spectf dataset chosen from UCI data repository. FSMs like Forward Selection (FS) and Backward Elimination (BE) based on wrapper model generates different subsets of important attributes by eliminating unnecessary or unsuitable attributes, thus by improving the performance of learning algorithms and most importantly producing reduced set of attributes. Our main objectives of the research work are:

- To build an efficient hybrid data mining classification and predictive model for medical datasets to scale down the procedural steps involved in diagnosing the disease.
- Generation of Important Attributes: FSMs are applied on the attributes of the medical datasets to get different subsets of important attributes, to develop economical healthcare solutions.

CHAPTER 2

LITERATURE SURVEY

In order to have a better understanding of the works previously done on this domain we went through the papers related to this domain. This gave us an understanding of the previously available models and also techniques and algorithms which might be useful for our work. Insights drawn from the papers and some of the notable works have been mentioned below.

2.1 General Working features of the existing system.

The general working features of the existing system is-

- Collection of various parameters impacting the chances of a person acquiring diabetes in future.
- Developing a dataset of these parameters and processing it through the proper algorithm and testing and training it to derive results from the data set.
- Recording these predictions in a sheet and developing reports.

2.2 Research Papers

2.2.1 PAPER 1
TOPIC : Prediction of Diabetes Empowered With Fused Machine Learning
Year of Publish : Jan 11 , 2022
Author : Usama Ahmed , Ghassan F.Issa , Muhammad Adnan Khan ,Shabib Aftab , Muhammad Farhan Khan and Munir Ahmad

This article proposes a Fused Model for Diabetes Prediction (FMDP). The proposed FMDP model consists of two main phases. The first phase consist of Training Layer while the second phase consists of Testing Layer. The Training Layer is divided into different stages, including data acquisition, pre-processing, classification, performance evaluation, an machine-learning fusion. The dataset used in this research is taken from the UCI Machine Learning Repository . In the Data Acquisition stage, a dataset that has enough features can be used to predict diabetes. Data is cleaned, normalized, and divided in to training and test dataset during the preprocessing stage. Preprocessed data can be used to train Support Vector Machines (SVMs) and Artificial Neural Networks (ANNs) for the prediction. We can select several Machine-learning algorithms for the classification to achieve the required accuracy. However, in the proposed model, we used only two widely used ML algorithms (SVMs and ANNs). The proposed fuzzy decision system has achieved the accuracy of 94.87.

2.2.2 PAPER 2

Topic : Machine Learning Tools for Long-Term Type 2 Diabetes Risk Prediction
Year Of Publish : Jan 20 , 2021
Author: Nikos Fazakis , Otilia Kocsis , Elias Dritsas, Sotiris Alexiou, Nikos Fakotakis and Konstantinos Moustakas

In this research, several strengths and limitations are high- lighted. In terms of the former, to our knowledge, it is the first to assess various ML models and provide participants with personalized long-term risk prediction of T2DM occurrence and appropriate guidance regarding lifestyle interventions. Also, the research findings were derived from across-sectional study on a representative English cohort (e.g., elderly office workers) with follow-up data; thus, we may identify causal and temporal associations between elderly lifestyle and T2DM. Another positive aspect of this work is that, during the balanced dataset creation, we drew instances of the initially ‘‘Non-Diabetics’’ class from the reference waves, whose class label was finally defined in the follow-up waves. This approach may give us a view of features behaviour for participants diagnosed with T2DM in the follow-up examination, contributing to T2DM prognosis. Moreover, our study revealed the importance of different risk factors in T2DM prediction for elder persons. The results of feature selection techniques coincided with the corresponding literature aboutT2DM risk factors. The selected features for the ML model straining and testing are among the symptoms/factors that doctors consider for quantifying long-term risk prediction or identifying its occurrence.

2.2.3 PAPER 3

Topic : Early Prediction of Diabetes Using an Ensemble of Machine Learning Models
Year Of Publish : September 2022
Author: Aishwariya Dutta , Md. Kamrul Hasan , Mohiuddin Ahmad , Md. Abdul Awal ,Md. Akhtar Islam Mehedi Masud 7 and Hossam Meshref

In this article Employing the suggested ML-based ensemble model, in which pre-processing plays a critical role in ensuring robust and accurate prediction, enabled this research to achieve its goal of making an early prediction of diabetes. The quality of the dataset was improved due to the presented pre-processing technique; the key considerations were selecting features and filling in missing values. The implementation of these pre-processing methods is required, which necessitates doing an exhaustive examination of the ablative processes in order to choose the most suitable approaches. In addition, when compared to previous research, this study produces a more accurate estimation despite including only four to five features, namely the body mass index (BMI) of the respondent, their present age, their average systolic pressure, and their average diastolic pressure, as well as their occupation, which is easily explicable. A weighted ensemble of machine learning classifiers may enhance the categorization consequences according to the suggested framework. This is accomplished by assigning a weight to the probability of the outcomes produced by the ensemble candidates' models. In terms of its potential to forecast diabetic disease classes in various medical settings, we anticipate that the model that we have developed would display both generality and flexibility. In addition, the extensive DDC dataset that was introduced from the South Asian country of Bangladesh (2011 and 2017–2018), which was the first dataset in this location, will continue to be helpful in future studies that involve the use of demographic information.

2.2.4 PAPER 4

TOPIC : Prediction and diagnosis of future diabetes risk: a machine learning approach .

Year Of Publish : August 2019

Author: Roshan Birjais, Ashish Kumar Mourya , Ritu Chauhan, Harleen Kaur1

In this paper they have used three techniques of machine learning –Gradient boosting, logistic regression and Naive Bayes to do the better diagnosis of diabetes disease. Using these three algorithms on Pima Indian diabetes data set, we can do the diagnosis whether the person is diabetic (1) or non-diabetic (0). With minor changes in the life style and in the eating habits, pre diabetic patients can be prevented from being diabetic, if not forever but at least for some duration of their lifetime. The results of the implementation show that gradient boosting has prediction accuracy of 86%, which is greater than the other two techniques used. The data set used is Pima

Indians diabetes dataset. In any research, data pre-processing is an important step in order to build the better and reliable model for the process of prediction.

2.2.5 PAPER 5

TOPIC : Machine learning and artificial intelligence based Diabetes Mellitus detection and self management: A systematic review

Year Of Publish : 30 June 2020

Author: Jyotismita Chaki , S. Thillai Ganesh , S.K Cidham , S. Ananda Theertan

This article presents several scientific problems that researchers have not been able to tackle in prior diabetic detection studies. Significant work is however also required to enhance the efficiency of various diabetic detection techniques. The research challenges that need to be tackled are laid out below.

1.No automated optimization technique

Deep learning has typically obtained promising outcomes in the field of DM detection and diagnosis, but the context of DL models is not fully known and is perceived to be a black box. For instance, several scientists have modified the established DL algorithms, like Deep NN or CNN, to enhance classification efficiency.

2.Training with inadequate data

DL software typically needs a significant number of diabetic data for training. When the training range is limited, it cannot yield sufficient results in terms of precision.

3. The integration of DL, AI, and cloud computing

In general, rural areas struggle from a shortage of human resources, particularly in medicine. Therefore, in these situations, AI may play a crucial role in resolving this constraint in the context of telemedicine. DL, AI, and cloud computing will be combined in the future to diagnose DM.

2.2.6 PAPER 6

TOPIC : Ethical and Legal Issues for Medical Data Mining

Year Of Publish : 2017

Published By : Ashwinkumar U.M. and Dr. Anandakumar K.R.

This Paper emphasizes on individuation and specialty of medical data processing health care connected data processing is one among the foremost reward full and difficult areas in application information, mining and knowledge discovery. The challenges square measure because of the information sets that square measure massive, complex, heterogeneous, hierarchical , statistic and of variable quality. The accessible health care datasets square measure fragmented and distributed in nature, thereby creating the method of knowledge integration a challenged task. the main problems associated with tackle square measure moral, legal and social aspects. because of the shortage of domain information on the analyst’s behalf it becomes necessary for a full of life collaboration between domain specialist and information mineworker with moral and legal clearance from specialized hospitals. Medical datasets represent a big a part of medical analysis. moral considerations, particularly problems with confidentiality have resulted within the introduction of rigorous rules in doing this manner of analysis. The deserves and demerits of those new rules square measure debated everywhere the planet. The introduction of rules for individual consent can prove pricey to Indian physicians. tries square measure being created to evolve a accord within which moral considerations square measure given due respect while not discouraging analysis.

2.2.7 PAPER 7

TOPIC : Design and Evaluation of Logistic Regression Model for Pattern Recognition Systems

Year Of Publish : 2019

Author: Pranav Rao and Manikandan J

In this paper, an effort is created to style pattern recognition systems exploitation logistical regression model and few mapping functions area unit planned for a similar. The performance of planned logistical regression model and mapping functions area unit assessed by evaluating the model exploitation style of digital circuits, commonplace datasets from UMASS info and datasets relating wireless device network applications. it's determined that for a majority of cases, the popularity accuracy is increased on exploitation planned mapping functions for each binary and multi category pattern classification issues

2.2.8 PAPER 8

TOPIC : Input Feature Extraction for Multilayered Perceptrons Using Supervised Principal Component Analysis

Year Of Publish : 2020

Author: Santosh S, Meenakshi Y

Recent enhancements within the space of knowledge mining and machine learning have sceptred the analysis in medicine field to enhance the condition of general health care. In several elements of the planet the tendency for maintaining long records consisting of medical information is turning into Associate in Nursing accepted follow. additionally to the present, the newer medical equipment’s and also the techniques employed in diagnosing, produces composite and big information. Therefore, to handle these ill structured medicine information, intelligent algorithms for data processing and machine learning square measure needed so as to require logical reasoning from the saved data, that is taken into account as medical data processing the newer medical equipment’s and also the techniques employed in diagnosing,produces composite and big information. Therefore, to handle these ill-structured medicine information, intelligent algorithms for data processing and machine learning square measure needed so as to require logical reasoning from the saved data, that is taken into account as medical data processing. among the medical information, the medical data processing searches for patterns and relationships which may offer helpful info for applicable diagnosing . data processing techniques square measure applied to totally different medical domains (health care databases or medical datasets) to enhance the diagnosing.

2.2.9 PAPER 9

TOPIC : Machine Learning-Based Application for Predicting Risk of Type 2 Diabetes Mellitus (T2DM) in Saudi Arabia: A Retrospective Cross-Sectional Study

Year Of Publish : 30 October 2020

Author: Asif Hassan and Tabrej Khan

In this paper survey, they focused on the following closed-ended research questioners for identifying participants at high risk of diabetes:

1. Choose the region of your residence.
2. How old are you?
3. What is your Gender?
4. What is your Body Mass Index (BMI)?
5. What is your Waist size?
6. Do you daily engage in at least 30 minutes of physical activity?
7. How often do you eat fruits and vegetables?
8. Have you ever taken hypertension medicine?

2.2.10 PAPER 10

TOPIC : 5G-Smart Diabetes: Towards Personalized Diabetes Diagnosis with Healthcare Big Data Clouds

Year of Publish : 2020

Author: Likitha V

The proposed research work uses feature selection methods like forward selection for Diabetes disorder medical dataset. LR, NN and NN with 10 fold CVS are applied on feature selection methods using Cross Validation Sample and Percentage Split as test options. From the experimental results it is identified that for Reduced Diabetes Disorder dataset with NN using percentage split of 66%, prediction accuracy of 84.52% is achieved. For all datasets used in the research work gives better classification accuracy with reduced subset of features. From the experimental results it is observed that the reduced subsets of attributes gives more efficient results than that obtained by using full set of attributes.

CHAPTER 3

SYSTEM REQUIREMENT SPECIFICATION

The specification is that the main demand for hardware and software system apparatuses that are castoff to run a project or task. For a system, necessities to be specified by the designer. Each hardware and software necessities are given below:

3.1 Hardware Requirements

Hardware is acknowledged because the physical pc resources that square measure demarcated by package or software package application that has some needs specification. Hardware needs List stays escorted through a Hardware Compatibility List, solely within the circumstance of disk- operating. A Hardware Compatibility List consists of verified, matched and intermittently mismatched hardware expedients for a certain disk package or specific solicitation.

- Processor: Intel Core i5
- RAM: 4GB
- Hard Disk Space: 250 GB
- Graphics Card: Microsoft DirectX 9 graphics device along with WDDM driver

3.2 Software Requirements

Software demand stipulation may be a requirements description for software-system and it's conjointly comprehensive rationalization of performance of system that has developed. conjointly embraces the definite use cases so as that designate communications of users with code. To stipulate the necessities, designers got to have clear and higher understanding of the merchandise that they're operating to develop. Following signifies the minimum code specifications:

- Operating System: Windows 8.1 or above
- Language: Python
- IDE: Anaconda Navigator
- Database: Microsoft SQL Server, AWS Cloud Services

3.3 Specific Requirements

3.3.1 User interfaces

The main UI appeared in this task is the principle structure separated into sub parts which are

- Menu to record tasks (results/logs).
- Division where the graphs are plotted.
- Result division for yield and conclusion to be appeared.
- Toolbar to permit client to choose data set.

3.3.2 Hardware interfaces

No equipment interfaces expected to run this product be that as it may, Diabetes identification are gotten by analyzing the pre-recorded data sets captured with a computer.

- Processor required ought to be least i3
- RAM size of 8GB

3.3.3 Software interfaces

Programming introduced in this SRS needn't bother with some other programming interface than the working framework and RStudio.

3.3.4 Communications interfaces

No web association is required to run this product in this way there won't be any correspondence interfaces.

3.4 Functional Requirements

- User will have the option to upload the data set to the product.
- User will have the option to analyze different logs at a given data set.
- User will have the option to see the final pragmatic result of the provided data set.

3.5 Software System Attributes

3.5.1 Performance

- Process will take close to ~5 minutes.
- Accuracy level of the yield picture will be higher than 70%.

3.5.2 Availability

Clients can utilize the framework on PC condition. To have the option to utilize the framework on PC condition, the framework is clicked and begun by client.

3.5.3 Security

Since there is no basic data to be kept, there are no security imperatives.

3.5.4 Portability

Since the framework will be created on PC condition, it can work just on PC.

3.5.5 Usability

- Software will acknowledge configurations of data set in fixed format
- Software will have the option to keep up any size of dataset.

3.5.6 Ease of use

Since the created framework is a clinical situated undertaking, it must be a specialist benevolent UI and this interface ought to be basic and simple to utilize.

CHAPTER 4

SYSTEM DESIGN

Medical data mining is strongly suggested to review any unseen or hidden pattern within the medical data. The actual challenges will be in the efficient projection, classification of the datasets. In order to address these, I am proposing a datamining model for the prediction of diabetes datasets with the help of integrating statistical approaches.

4.1 Proposed Framework for Prediction of Diabetes Datasets

When selecting the model some of the important attributes, reduction of attributes for the classification and prediction of diseases for a given medical data set is proposed. The proposed framework for classification and prediction of medical datasets.

The proposed framework implementation process is presented with following steps:

- The first step is the selection of the datasets.
- The second step is preprocessing for missing attributes.
- The fourth step is in applying algorithms.
- Last step is to find the accuracy of classification for each ANN, LR, SVM and RF with CVS and percentage split as test alternatives.

4.2 Framework for Data mining Model for categorization and Prediction of Medical Datasets

Data mining model involves the task of classification and prediction, then the attribute choice is one of the vital steps. In many existing researches works on medical datasets some researchers make use of full set of attributes and others use less number of attributes for prediction. If the model uses full set of attributes then it takes more time for prediction and the classification accuracy attained is not satisfactory. In order to obtain better classification accuracy with less number of attributes and to lessen the time taken for prediction, I suggest a data mining model for better attribute selection. Proposed model, focuses on identifying the important attributes that are capable enough for predicting the existence or nonexistence of disease within the medical datasets. This results in reducing the amount of time taken for prediction and also increases the classification accuracy. By identifying the important attributes, I lessen the quantity of medical tests that are necessary for prediction. In our proposed research work, I suggest a data mining model for classification and prediction of medical datasets using LR and ANN with CVS and percentage split as test alternatives. The performance of the model is measured by classification accuracy.

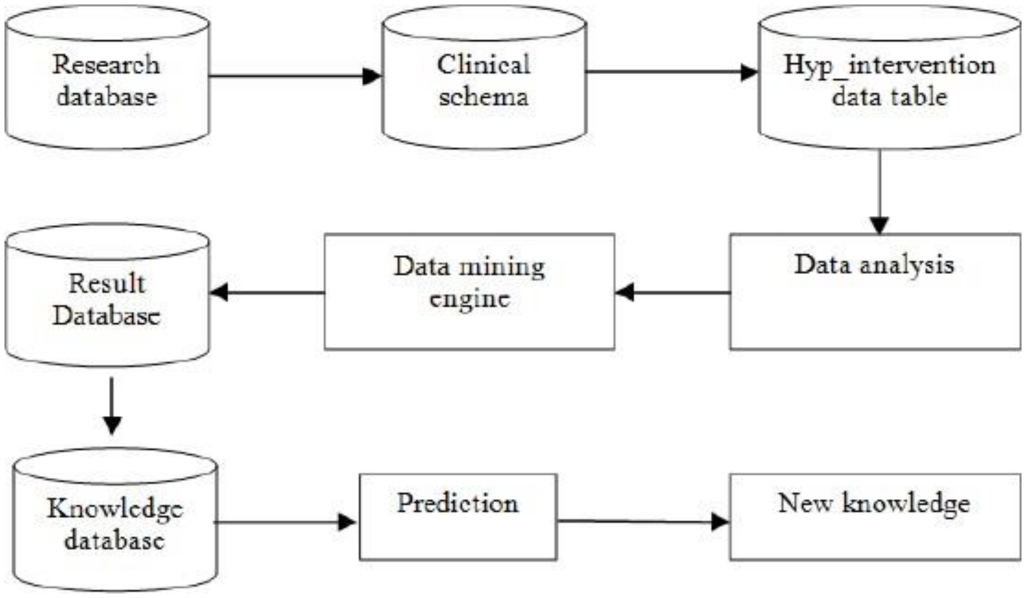


Fig 4.1: System outline of the diabetes system

The above figure 4.1 represents the architecture diagram for the system, this depicts the high-level understanding of the system and the concepts used in it. This is the set of the graphical representation of the system and the principles which have been included in it, and also it involves all the elements and the components in it. The important activity is in the solution architecture, within the pace of business needs providing the key decisions of the design and also the technical leadership for the IT systems project the above architecture includes the research database, clinical schema, data mining engine, the result database, knowledge database and also the prediction with the new knowledge in it.

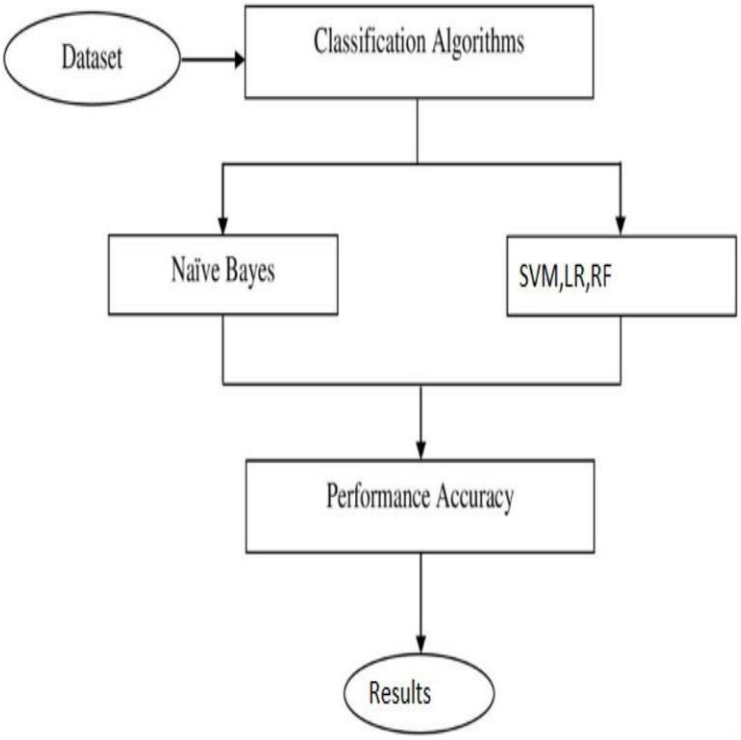


Fig 4.2: Flow chart of the diabetes system model

Flowchart may be a sort of diagram that represents Associate in Nursing rule, progress or method. flow sheet may also be outlined as a diagrammatical illustration of Associate in Nursing rule (step by step approach to unravel a task). This shows the steps as boxes of assorted sorts, and their order by connecting the boxes with arrows. Flowchart's area unit employed in analyzing, designing, documenting or managing a method or program in varied fields. Flowchart's area unit employed in coming up with and documenting straightforward processes or programs. Like different styles of diagrams, assist visualize what's occurring and thereby help perceive a method, and maybe conjointly realize lessobvious options inside the method, like flaws and bottlenecks. There area unit sorts differing kinds of flowcharts: every type has its own set of boxes and notations

Above flowchart shows the typical representation of the system in a easier way for the user to understand the details of the attributes and the components involved in it. This involves the dataset which includes all the dataset of the patients and also the algorithms like the random forest, logistic regression and others. The performance accuracy is been involved in it and in the end it's the outcome or the results.

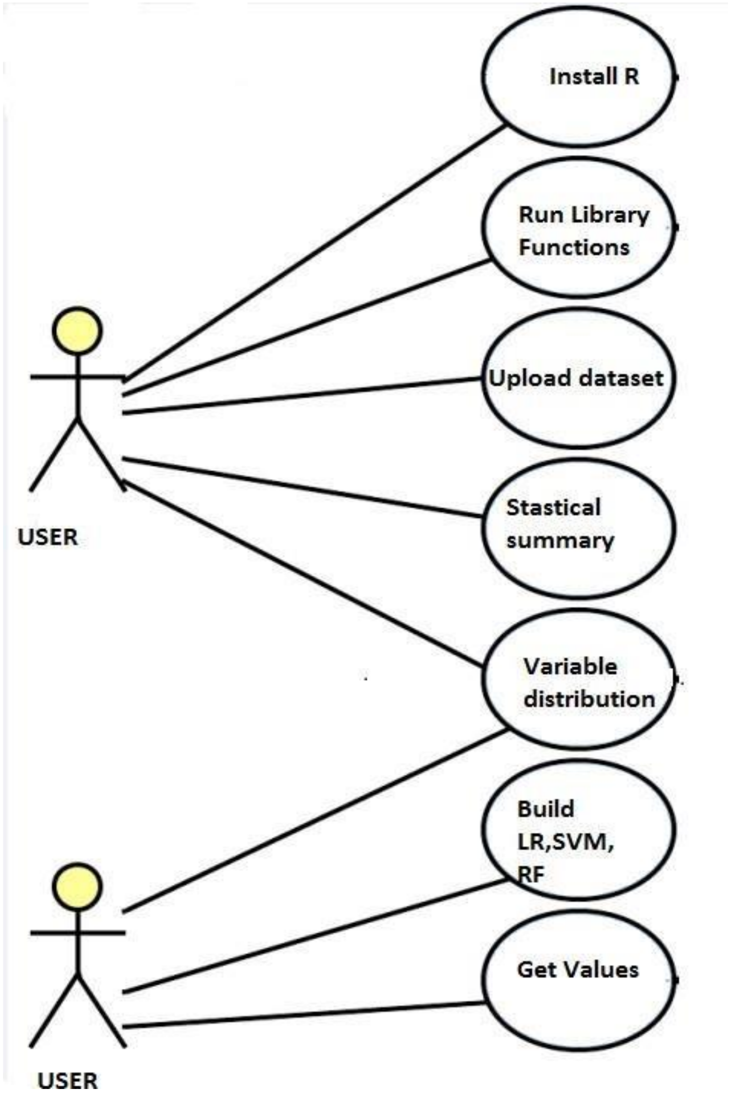


Fig 4.3: Use Case Diagram of diabetes system

Use case is a list of event steps in defining the interactions between a role Unified Modeling Language (UML) and a system to achieve a goal usually used by system engineers. The actor may be a external system or a person. use case diagrams are used at a higher level in system engineering than within software engineering, often showing missions or stakeholder goals. The basic necessities may then be captured in the Systems Modeling Language or as written agreement statements.

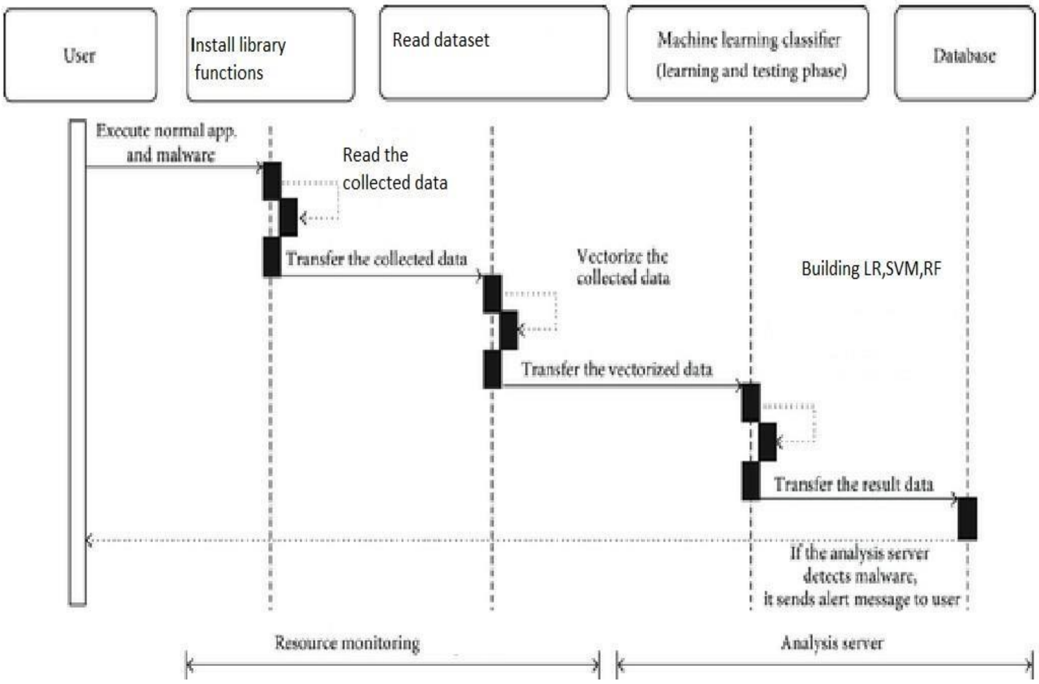

User
Install library functions
Read dataset
Machine learning classifier (learning and testing phase)
Database
Execute normal app. and malware
Read the collected data
Transfer the collected data
Vectorize the collected data
Transfer the vectorized data
Building LR,SVM,RF
Transfer the result data
If the analysis server detects malware, it sends alert message to user
Resource monitoring
Analysis server

Fig 4.4: Sequence Diagram of the diabetes system

Above fig 4.4 represents the interactions arranged in between the given time space. Object. This shows the objects and classes involved into it and also the message order exchanged in between the objects needed to carry the functionality in the scenario. Here the diagrams are typically in association with use case realizations in the Logical View of the system under development. Sequence diagrams are sometimes called event diagrams or event scenarios**.** It also shows, as parallel vertical lines different processes or objects that live simultaneously, and, as horizontal arrows, the messages exchanged between them, in the order in which they occur. Also allows in the description of runtime summary in a graphical manner.

Above illustration shows the user and the system interaction, how the data is been provided to the user and the installations of the packages and the sequential representation of the flow in the system.

If a data mining model involves the task of classification and prediction, then the attribute choice is one of the vital steps. In many existing research works on medical datasets some researchers make use of full set of attributes and others use less number of attributes for prediction. If the model uses full set of attributes then it takes more time for prediction and the classification accuracy attained is not satisfactory. In order to obtain better classification accuracy with less number of attributes and to lessen the time taken for prediction, we suggest a hybrid data mining model for better attribute selection.

In our proposed hybrid model, we focus on identifying the important attributes that are capable enough for predicting the existence or nonexistence of disease within the medical datasets. This results in reducing the amount of time taken for prediction and also increases the classification accuracy. By identifying the important attributes, we lessen the quantity of medical tests that are necessary for prediction. In our proposed research work, we suggest a hybrid data mining model for classification and prediction of medical datasets using LR and ANN with CVS and percentage split as test alternatives. Classification accuracy is used to measure the performance of the hybrid model.

The detailed framework for the proposed hybrid data mining model for classification and prediction of medical datasets is shown in Figure 3.2.

The different steps involved are:

1. Selection and preprocessing of medical datasets.
2. For selecting important records within the attribute we propose Entropy Evaluation method is suggested on medical datasets to select important attributes.

   By using Entropy Evaluation method, we calculate the information gain or entropy value of each attribute. Information gain of each attribute helps in selection of important attributes.
3. Based on the information gain of each attribute, count of number of 1’s obtained based on Mean value for each attributes and count of 1’s obtained based on Entopy value for each attribute, we apply FSMs like FS and BE. FS and BE methods generate different subsets of attributes for each hybrid technique and for each medical dataset.
4. For the different subsets of attributes obtained from FSM based on Entropy value, we evaluate the performance of LR ,SVM, RF, NN and NN with k-fold using percentage split and CVS as test alternatives.
5. The evaluation or predictions are based on classification accuracy. The subset that gives better classification accuracy is considered as the best one and attributes in that subset are the most important.

CHAPTER 5

IMPLEMENTATION AND TESTING

Implementation is stage/phase of undertaking where hypothetical outline, dreams and plans are changed over to a working framework. Along these lines it is considered as a standout amongst the most imperative stage in building another framework effectively and giving certainty to the client that the new framework fabricated will be viable and works faultlessly. The major steps involved in implementation stage are analyzing the problem, planning, and careful investigation of implementation constraints, evaluating and optimizing the design. The use of classification, FSMs and prediction on medical datasets will result in scaling down the procedural steps involved to diagnose the disease. This provides an economical solution for healthcare schemes and diagnosis using medical software systems.

Medical data mining is strongly suggested to review any unseen data. The actual challenges are the efficient classification and prediction of medical datasets. To address these challenges, a hybrid data mining model for the classification and prediction of medical datasets by integrating the statistical techniques like Entropy Evaluation method, Mean Evaluation method and Threshold Evaluation Method with FSMs based on greedy algorithm of wrapper model is been applied.

5.1 Proposed Framework for Classification and Prediction of Medical

Datasets Hybrid framework selection of important attributes, reduction of attributes for the classification and prediction of diseases for a given medical data set is proposed. The proposed framework for grouping and prediction of medical datasets is shown in Figure.

The proposed framework implementation process is presented in the steps below :

- First is the selection of the medical datasets.
- The second step is preprocessing for handling missing attributes.
- The third step is to select important attributes based on Entropy Evaluation method, Mean Evaluation method, and Threshold Evaluation method.
- The fourth step is to apply FSMs like FS and BE on the attributes obtained in step3.

These result in different subsets of attributes obtained for both FS and BE method.

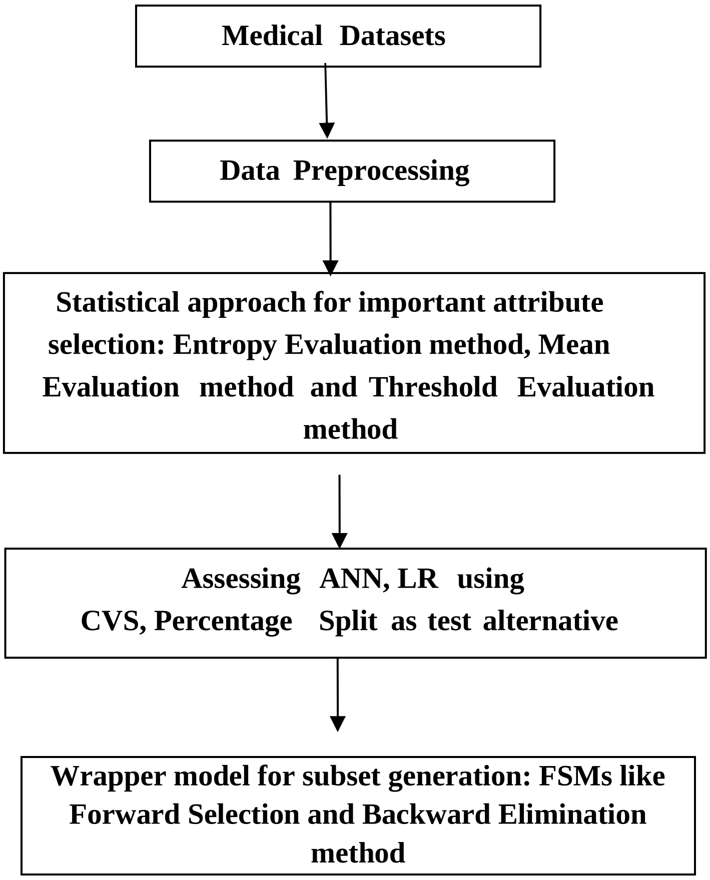

Medical Datasets
Data Preprocessing
Statistical approach for important attribute selection: Entropy Evaluation method, Mean Evaluation method and Threshold Evaluation method
Assessing ANN, LR using CVS, Percentage Split as test alternative
Wrapper model for subset generation: FSMs like Forward Selection and Backward Elimination method

Figure 5.1: Proposed Framework for Classification and Prediction of Medical Datasets

Lastly it is to find the classification accuracy for each subset of attributes using ANN and LR with CVS and percentage split as test alternatives.

5.2 Framework for Hybrid Data Processing Mining Model for Classification and Prediction of Medical Datasets

The task of classification, prediction, then the attribute choice is one of the vital steps. In many existing researches works on medical datasets some researchers make use of full set of attributes and others use less number of attributes for prediction. If the model uses full set of attributes then it takes more time for prediction and the classification accuracy attained is not satisfactory. In order to obtain better classification accuracy with less number of attributes and to lessen the time taken for prediction, suggest a hybrid data mining model for better attribute selection. Here on identifying the important attributes that are capable enough for predicting the existence or nonexistence of disease within the medical datasets. This results in reducing the amount of time taken for prediction and also increases the classification accuracy. By identifying the important attributes, lessen thequantity of medical tests that are necessary for prediction. In our proposed research work, suggest a hybrid data mining model for classification and prediction of medical datasets through FSMs using LR and ANN with CVS and percentage split as test alternatives. Also accuracy of the classification is used in measuring the performance of hybrid model. The framework is elaborated data mining model for classification and prediction of medical datasets is shown in figure.

The different steps involved are:

- Selection and preprocessing of medical datasets.
- For selecting important records within the attribute propose Entropy Evaluation method, Mean Evaluation method, and Threshold Evaluation method are suggested on medical datasets to select important attributes. By using Entropy Evaluation method, I calculate the information gain or entropy value of each attribute. Information gain of each attribute helps in selection of important attributes. By using Mean Evaluation method, I calculate the Mean value of each attribute. Based on Mean value I transform the record values in medical dataset to 1's and 0's. The records with value 1 within the attribute are important and considered for selection. the number of 1's for each attribute. By using Threshold Evaluation method, I calculate the Threshold value of each attribute. Based on the Threshold value I transform the medical dataset to 1's and 0's. The records with value 1 within the attribute are important and considered for selection number of one for every attribute.
- Formulated over the basis of the statistics information gain of every attribute, count of number of 1's obtained based on Mean value for each attributes and count of 1's obtained based on Threshold value for each attribute, I apply FSMs like FS and BE. FS and BE methods generate different subsets of attributes for each hybrid technique and for each medical dataset.

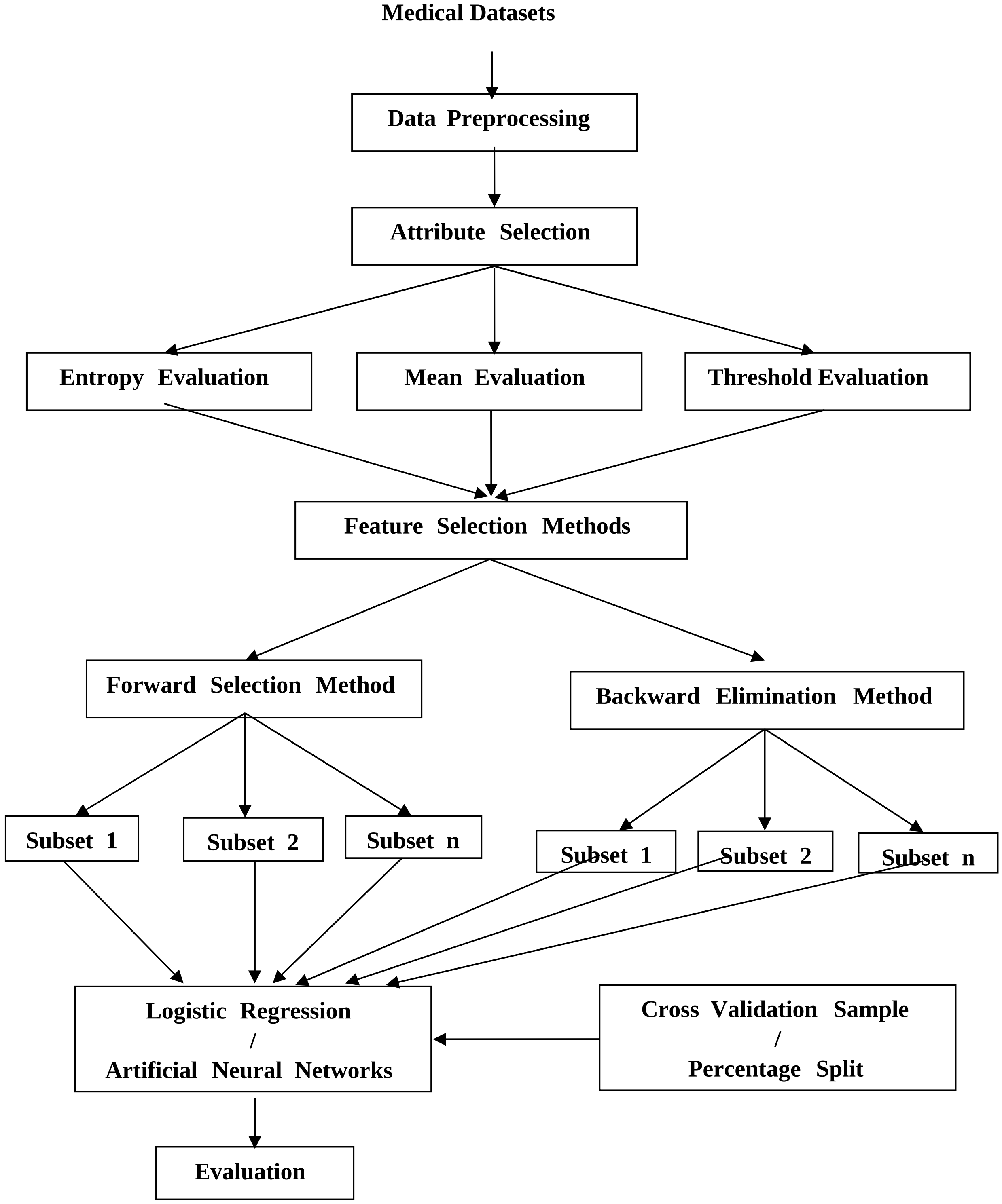


Figure 5.2: Detailed Framework for Hybrid Data Mining Model for Classification and Prediction of Medical Datasets

Use of FSM helps in reducing the number of attributes required for classification and prediction of disease.

- For the different subsets of attributes obtained from FSM based on Entropy value, Mean value and Threshold value, I evaluate the performance of LR and ANN using percentage split and CVS as test alternatives.

- The evaluation or predictions are based on classification accuracy. The subset that gives better classification accuracy is considered as the best one and attributes in that subset are the most important.

5.3 Feature Selection Methods

Steps in any analysis of the information mining dependent on classification and prediction models is feature selection. It's much more important when I construct a medical data mining model, the medical dataset may generally consists of further information than the actual information needed to construct the model. Incase preserving the attribute columns which are not actually needed, then it leads to wastage of memory and more CPU time is needed for the training process and the quality of the explored pattern may be deteriorated by these extra attributes namely because of the following reason:

- It is difficult to discover meaningful pattern from data because some attributes may be redundant, noisy and
- For identifying excellent pattern, the majority of data mining algorithms need huge training dataset but the data used for training is extremely minute in few of the applications of data.

Also provides the assists in solving the various problems by having few little information of high value rather than much of the data of lower value. In the suggested work to select the best subset of attributes for all the three techniques I propose FS and BE algorithms.

5.3.1 Forward Selection and Backward Elimination Methods

A medical dataset consists of many different attributes upon which the disease is predicted. But these entire attributes need not be considered during prediction. Some attributes need not be useful during prediction and some may be very much important. I need to identify the important attributes among the all. Identifying these will help in reducing the procedure involved during diagnosis. Thus, it is economical to find the important attributes I make use of FSMs like FS and BE. Both of these methods give different subsets of attributes. The attribute subset that gives better classification accuracy is the important one. FS method decreases the quantity of models that I may need to study. In FS I start with a blank model and during each step of the algorithm, Selecting an attribute based on the ascending order of the value which gives the better performance of classification when added to present group or set of attributes. The procedure for choosing the best attribute to add into an iteration of forward attribute selection is referred to as forward attribute selection. The algorithm for FS is shown below:

Algorithm: Forward Selection

Input: Medical dataset

Output: Subsets of attributes

Method: Keep the class attribute and the 1st attribute.

Create two variables key1 and key2. Key1 contains the list of attribute names based on ascending order of entropy value and/or number of 1's. Key2 stores the attribute names in original order. Compare key1 list and key2 list.

- If both are same then remove the attribute from dataset and also remove the attribute from key2 list and evaluate.
- Do step 2 until last attribute in the dataset.

BE method decreases the quantity of models that I may need to study. In BE, I start with a complete model and during each step of the algorithm I select an attribute based on the descending order of the value, which gives the better performance of classification when added to present set of attributes. The process of choosing the least attribute to remove in an iteration of backward attribute elimination is referred to as backward attribute evaluation. The algorithm for BE method is given. The FS and BE algorithms give different subsets of important attributes. Each of these methods gives different combination of attributes.

Some of the advantage sin classifying the data are:

- Complexity is reduced due to reduction in dimension .
- Improvisation of classification accuracy due to noise reduction .

Algorithm: Backward Elimination

Input: Medical dataset

Output: Subsets of attributes

Method:

- Keep the class attribute and all attributes.
- Create two variables key1 and key2. Key1 contains the list of attribute names based on descending order of entropy value and/or number of 1's. Key2 stores the attribute names in original order. Compare key1 list and key2 list.
- If both are same then remove the attribute from dataset and also remove the attribute from key2 list and evaluate.
- Do step 2 until last attribute in the dataset.

5.4 Datasets Considered in the Proposed work

Three real world datasets in Attribute Relation File Format (ARFF). Description is given as below in Table.

Table 1: Specifications of Medical Datasets

Sl. No	Medical Datasets	Number of Instance	Number of attributes	Number of Classes
1	Diabetes	3000	9	2

5.5 Proposed Entropy Evaluation Method for Classification and Prediction of Medical Datasets with Feature Selection Methods

Improving the accuracy classification and the prediction rate of a medical dataset, I need to select the important attributes. To select the important attributes, I need a criterion. The criterion considered here is the Entropy. When Entropy Evaluation method is implied on the medical datasets it gives the Entropy value or information gain of every attribute. Depending on Entropy value of each attribute I apply FSMs like FS and BE. Which results in different subsets of attribute.

5.5.1 Entropy Evaluation Method

The Entropy Evaluation method for classification and prediction of medical datasets with FSMs is shown in the below Figure. The Entropy Evaluation method can be applied on any medical datasets irrespective of the type of attribute value (real values or binary). The steps involved in Entropy Evaluation method are:

- Selection of medical datasets.
- Preprocessing is carried out for any missing attribute values in datasets.
- For the preprocessed data, I apply Entropy Evaluation method. The Entropy Evaluation method is implemented as follows:
- The entropy value is calculated for each attribute in the dataset except for the class attribute using the equation 7.

$$\text{Info (D)} = \sum_{i=1}^{m-1} p_i \log_2(p_i) \ldots\ldots \text{eqn}(7)$$

Where ,

D-attribute.

i-attribute index.

p_{i-} probability that an attribute in D belongs class. m-total count of attributes. *Info (D)* is

also described as entropy of D.

- The amount of additional information that is required to arrive at an exact classification is calculated using equation 8.

$$\mathrm{Info}_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \text{ X } \mathrm{Info}(D_j)\text{.......eqn(8)}$$

Where ,

|Dj|/|D| is weight of j partition. j -partition index.

v -total number of partition.

InfoA(D) total no of expected data to classify the attribute D based on partitioning by A.

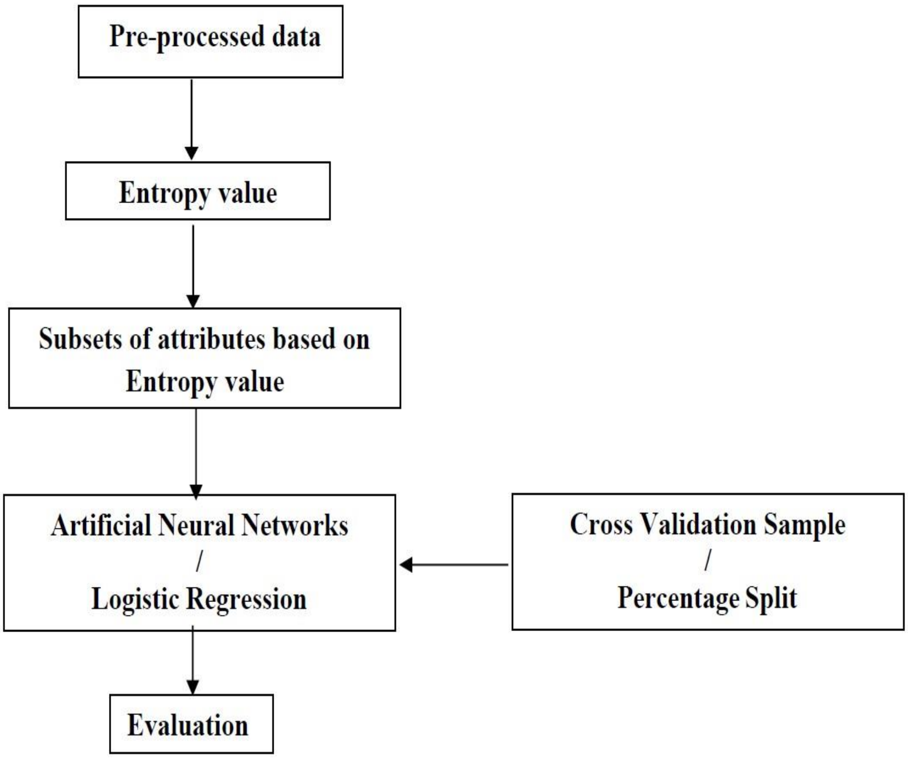

Pre-processed data
Entropy value
Subsets of attributes based on Entropy value
Artificial Neural Networks / Logistic Regression
Cross Validation Sample / Percentage Split
Evaluation

Figure 5.3: Entropy Evaluation Method

The difference in the original information required and the new requirement is the information gain and is calculated using equation 9.

$$\text{Gain (A) =Info (D) – }\mathrm{Info}_A\text{ (D). ..... eqn(9)}$$

The value of entropy for each attribute in the medical information set I apply the FSMs like FS and BE, which generates different subsets of attributes.

5.5.2 For Diabetes Dataset Based on Entropy Evaluation Method

After applying FSMs like FS and BE, on getting different subsets of attribute in Table

Table 2: Different subsets of attributes obtained after applying FS method based on Entropy value or information gain of each attribute of diabetes dataset

Subset No.	Subset of Attributes	No. of Attributes
1	pres, class	2
2	pres, pedi, class	3
3	preg, pres, pedi, class	4
4	preg, pres, skin, pedi, class	5
5	Preg, pres, skin, ins, class	6
6	Preg, pres, skin, insu, pedi, age, class	7
7	Preg, pres, insu, class, mass, pedi, ag	8

Table 3: Different subsets of attributes obtained after applying BE method based on Entropy value or information gain of each attribute of diabetes dataset

Subset No.	Subset of Attributes	No. of Attributes
1	Preg, plas, skin, insu, pedi, age, class	8
2	Preg, plas, skin, insu, mass, age, class	7
3	plas, skin, insu, mass, age, class	6
4	Plas, insu, mass, age, class	5
5	plas, mass, age, class	4
6	plas, mass, class	3
7	plas, class	2

5.5.3 Classification Accuracy Attained for the Diabetes Dataset Based on Entropy Evaluation Method

Full Sets of the Attribute of dataset of diabetes. The accuracy attained for all the attribute set of diabetes dataset is shown in Table below.

Table 4: Classification Accuracy for full attribute set of diabetes dataset

Technique Used for Finding Classification Accuracy	CVS	Percentage Split			
		50%	**66%**	**70%**	**75%**
LR	77.21	77.86	80.07	80.43	**81.77**
ANN	75.39	76.82	74.32	75.21	77.08

From Table IV I can see that for full attributes set of diabetes dataset a classification accurate no of 81.77% is obtained by LR using percentage split of 75%.

Table 5: Classification Accuracy of diabetes dataset for reduced set of attributes using BE method based on information gain

Technique Used for Finding Classification Accuracy	CVS	Percentage Split			
		50%	**66%**	**70%**	**75%**
LR	76.43	78.39	81.22	80.87	**81.77**
ANN	76.30	78.90	79.69	81.73	81.25

From the above Table, I can clearly see that the classification accuracy of 81.77% is attained with only three attributes of diabetes dataset by LR using BE method with percentage split of 75%.

Testing

Testing is a procedure which guarantees excellence and efficiency of the proposed system in nourishing its purposes. Testing is done at various stages in the System designing and implementation process with an objective of developing a transparent, flexible and secured system. Unit testing is used to verify the logic of individual components of the module. It is the first phase during coding. Each individual component is tested against the requirements provided during the design.

Unit Testing

Unit tests are small functions that test your code and help you make sure everything is alright. First, you'll need to download and install testthat. Some dependencies will also be installed.

testthat and RUnit are two main packages in R that tools unit testing.

testwhat is an R package that assist to write these SCTs for interactive R drills. testwhat provides a horde of function to test entity definition, function calls, definitions, for loops, while loops, and countless.

Unit tests

- Each new feature should be accompanied with unit tests, by using the testthat R package.
- For each R-script file named script.R, a correspond test file should be created in inst/testsdirectory, using the writing convention test_<script>.R

RUnit : This package models the common Unit Test framework for R and provides functionality to track results of test case execution and generate a summary report.

The function may be used to approach the degree of morphological assimilation between two or more sets of variables. Input can take one of two forms. First, one can input a single dataset .Alternatively, when evaluate the assimilation between two structure or partition, two datasets may be provided. The generic functions print(), summary(), and plot all work with integration.test. This function calls plot.pls(), which has two extra dispute (with defaults): label = NULL, warpgrids = TRUE.

5.6 Test Cases

It explains the overall description of inputs, conditions, executions of the system, procedure in testing the system and also the results obtained in the end which defines the single tests which is to be executed in order to achieve objective in testing of the software objective, such as to exercise a particular program path or to verify compliance with a specific requirement. the test cases help in underlie testing which is good to obtain the result.

Table 6: Test cases for the Users

Sl.no	Test Input	Expected result	Actual Result	Remarks
1	Provide data set which consists of Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age. Split Ratio value is 70%.	Diabetes outcome s hall be 1 if either of the input parameter is more than normal l range else 0.	Diabetes outcome s hall be 1 if either of the input parameter is more than nor mal range else 0.	Pass
2	Provide data set which consists of Pregnancies, Glucose, Blood Pressure, Ski n Thickness, Insulin, BMI, Diabetes Pedigree Function, Age. Split Ratio value is 80%.	Diabetes outcome s hall be 1 if either of the input parameter is more than nor mal range else 0.	Diabetes outcome s hall be 1 if either of the input parameter is more than nor mal range else 0.	Pass
3	Provide data set which consists of Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BM I, Diabetes Pedigree Function, Age. Split Ratio value is 60%.	Diabetes outcome shall be 1 if either of the input parameter is more than nor mal range else 0.	Diabetes outcome shall be 1 if either of the input parameter is more than normal range else 0.	Pass
4	Provide data set which consists of Pregnancies, Glucose, Blood Pressure, Ski n Thickness, Insulin, BMI, Diabetes Pedigree Function, Age. Split Ratio value is 50%.	Diabetes outcome shall be 1 if either of the input parameter is more than nor mal range else 0.	Diabetes outcome shall be 1 if either of the input parameter is more than normal range else 0.	Pass
5	Provide data set which consists of Pregnancies, Glucose, Blood Pressure, Ski n Thickness, Insulin, BMI, Diabetes Pedigree Function, Age. Split Ratio value is 75%.	Diabetes outcome shall be 1 if either of the input parameter is more than nor mal range else 0.	Diabetes outcome shall be 1 if either of the input parameter is more than normal range else 0.	Pass

5.7 Integration testing

Data might be lost through a gadget, one subsystem may detrimentally affect the other subunit, and the proposed primary element can't be produced when blended. Incorporated exploration is a normalized investigation and can be completed utilizing test results. Finding the general framework execution is the requirement for the coordinated test.

There are two sorts of testing for the incorporation. These are as per the following:

- Top-down integration testing.
- Bottom-up integration testing

Result: Our Undertaking Module Has Passed Both Top down and Base up Approach verification.

5.8 Black box testing

- Black box testing is never really misguided or absent limit
- Organize botch
- Failure in foreign database get to
- Achievement botches
- Determination and end botches

In 'valuable testing', is achieved to affirm an function agrees to its specifics of precisely plays out the sum of its vital limits. So this testing is moreover called 'revelation testing'. It examines the outside direct of the structure. However the manufactured thing might be taken a stab at knowing the prearranged work that have been proposed to achieve; tests are often composed to point out that every breaking point is absolutely operational.

Sl no	**Black box testing**	**Results**
1	Interface Testing	Found bugs - rectified
2	Database Testing	No errors
3	Performance Testing	Medium
4	Program Errors	Found errors- rectified
5	Initialization	No errors
6	Memory Testing	No memory leakages

5.9 Summary

After applying Entropy Evaluation method on the diabetes dataset, diabetes disorder dataset, and Spectf dataset get different subsets of attributes as shown above. For these subsets find the classification accuracy using LR and ANN with CVS and percentage split as test alternatives.

- Attribute set of the dataset for the classification accuracy obtained is 97.33% by RF with percentage split of 75%

CHAPTER 6

RESULTS AND DISCUSSION

The Diabetes Prediction System was trained on a large dataset comprising medical records of individuals, including both diabetic and non-diabetic patients. The dataset was split into training and testing sets to evaluate the performance of the model. The following metrics were used to assess the system's predictive ability

6.1 Levels of Testing

Unit tests are small functions that test your code and help you make sure everything is alright. First, you'll need to download and install testthat. Some dependencies will also be installed. The two main packages in R which implements unit testing: RUnit and testthat. Test what is an R package that can help you write these SCTs for interactive R exercises. Test what provides a bunch of functions to test object definitions, function calls, function definitions, for loops, while loops, and many more. You can also use testwhat to write custom unit tests. Testwhat automatically generates meaningful feedback that's specific to the student's mistake; you can also choose to override this feedback with custom messages.

6.2 Datasets Considered in the Proposed work

Correlations area units displayed in blue and negative correlations in red color. Color scale and also the size the circle area unit relative to the connection coefficients. The recipient Operating Characteristic (ROC) curve is recycled to review the efficiency of an endless classifying for anticipate a binary outcome. In medicine, use for assess tests in radiology and diagnostics. ROC curves are also used in signal detection theory. Most machine learning classifiers production in real-valued count that correlates with the energy of the prophecy that a given case is positive. Turning this real-valued count into yes or no prophecy requires setting a threshold. Scores which are above the threshold are considered as positive and scores below the threshold are prophecy to be negative. The real-world datasets in Attribute Relation File Format (ARFF) and explained in Appendix .

The details is given below.

Table 7: Specifications of Medical Datasets

Sl.No	**Medical Datasets**	**No of Instances**	**No of Attributes**	**No of Classes**
1	PIMA Indian diabetes	345	7	2

Title: PIMA Indian diabetes

The 5 attributes area unit blood tests that are measured to be at risk of PIMAIndian diabetes that may arise from surplus alcohol use. Every line of the dataset constitutes a proof of one male entity. It appears that drinks is some sort of selector on this database.

Attribute details:

- No of times pregnant
- Plasma glucose absorption 2 hours in an oral glucose tolerance test
- Diastolic blood pressure (mm Hg)
- Triceps skin fold thickness (mm)
- 2-Hour serum body fluid insulin(mu U/ml)
- Body mass index (weight in kg/(height in m)^2)
- Diabetes pedigree function
- Age (years)

The classification accuracy attained for the full attribute set of PIMAIndian diabetes dataset is shown in below

Table 8: Classification Accuracy for full attribute set of PIMA Indian diabetes dataset

Technique Used in Finding Classification Accuracy	**Percentage Split**				
	50%	**66%**	**70%**	**75%**	**80%**
LR	83.8	83.2	82	81.2	79.9
NN	83.29	83.11	82.88	82.8	83.4
NN with 10 Fold	83.70	**84.52**	83.35	83.55	83.8
SVM	72.9	69.9	64.7	66.4	65.3
RF	83.8	83	83.7	84.2	**84.5**

The below figure 6.4 represents the Correlation of Diabetes data with all the attributes and the values for it. The diabetes data involves the following set of attributes like blood pressure, age, pregnancies, diabetes pedigree function, skin thickness, BMI, outcome, glucose, insulin. And the values varies from 0.1 to 0.5.

The below representation as shown in figure 6.1 represents the correlation of the dataset also shows the indication in the levels of the values. Each indicated values represented might vary between the range 0.5 to 1.0. Each parameter is represented in the y axis and result are plotted in x-axis.

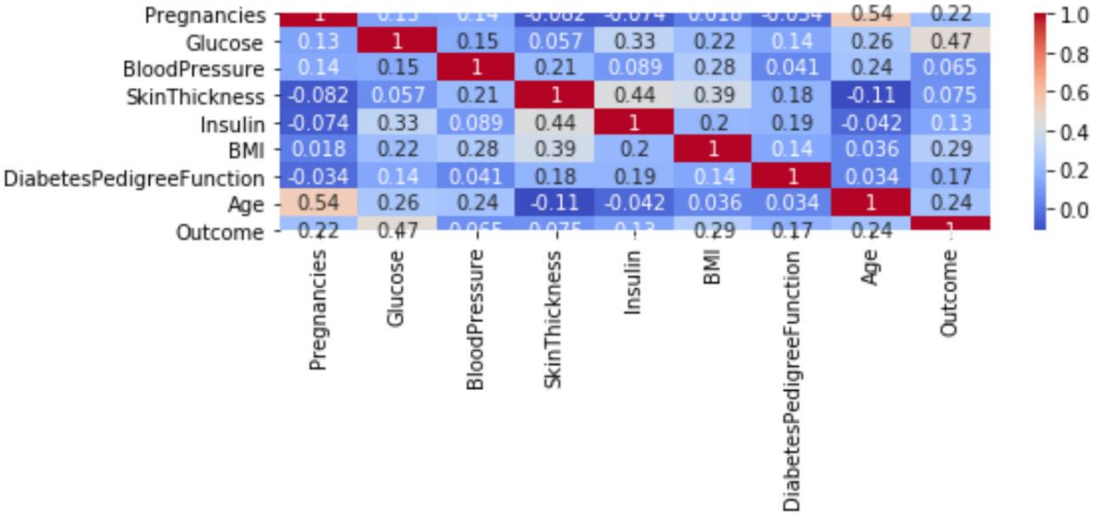
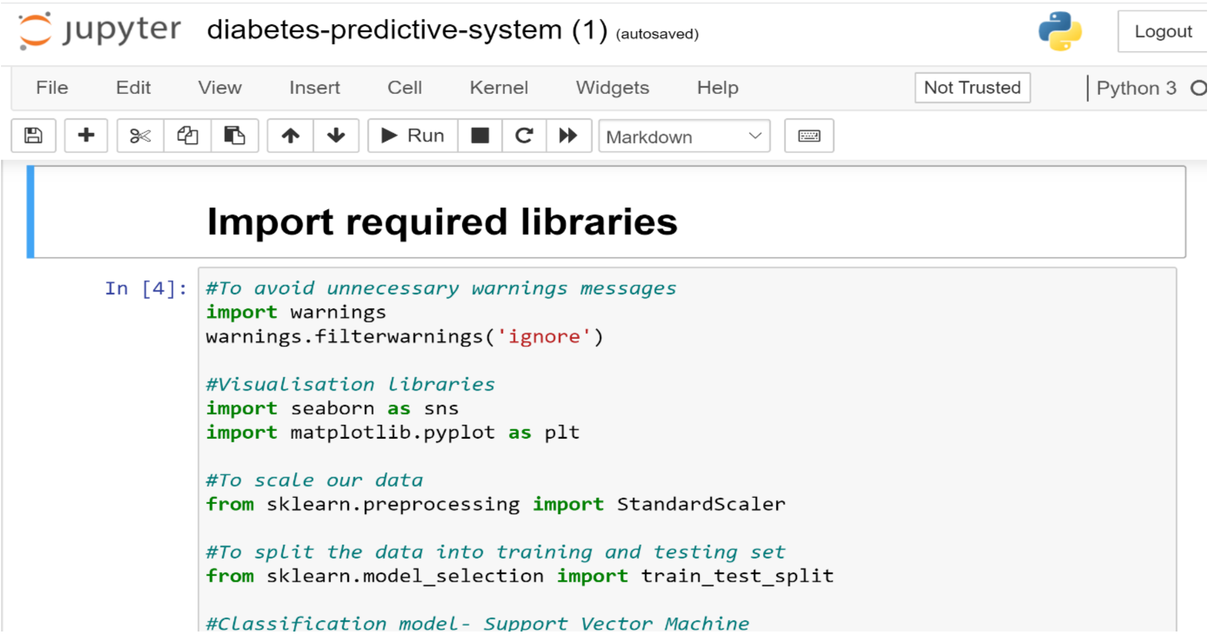


Fig 6.1: Correlation of Diabetes data



Import required libraries

```
In [4]: #To avoid unnecessary warnings messages
        import warnings
        warnings.filterwarnings('ignore')

        #Visualisation libraries
        import seaborn as sns
        import matplotlib.pyplot as plt

        #To scale our data
        from sklearn.preprocessing import StandardScaler

        #To split the data into training and testing set
        from sklearn.model_selection import train_test_split

        #Classification model- Support Vector Machine
```

Fig 6.2: Jupyter Notebook interface

The top figure 6.2 shows the illustration of the Jupyter notebook interface. This shows the define image and therefore the contents in it. This includes the library functions and therefore the run programs. Python could be a free, ASCII text file software system associated artificial language developed as an setting for applied mathematics computing and graphics Since then Python has become one amongst the dominant software system environments for information analysis and Python is especially standard for its graphical capabilities, however it's additionally prized for its GIS capabilities that build it comparatively simple to get raster-based models. additional recently, Python has additionally gained many packages that are designed specifically for analyzing information.

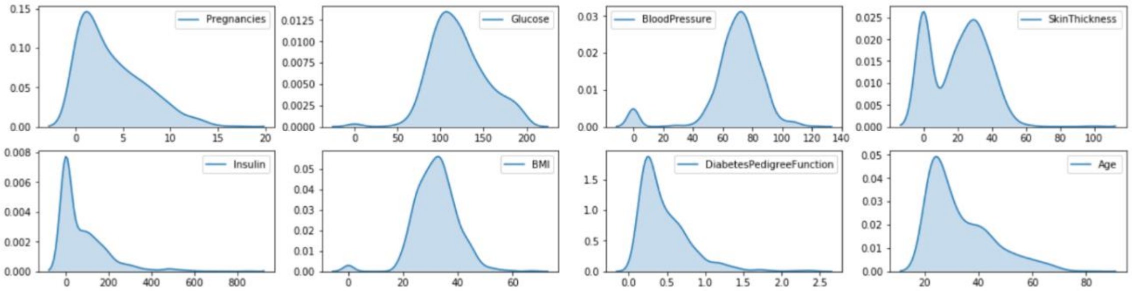


Fig 6.3: Count of each attributes with value.

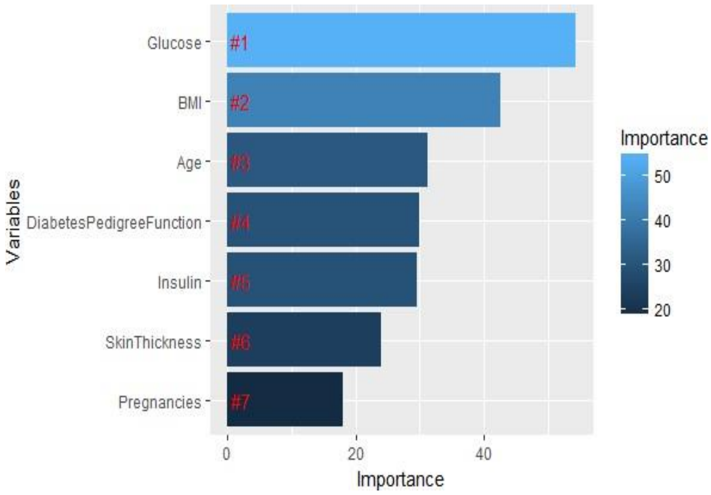
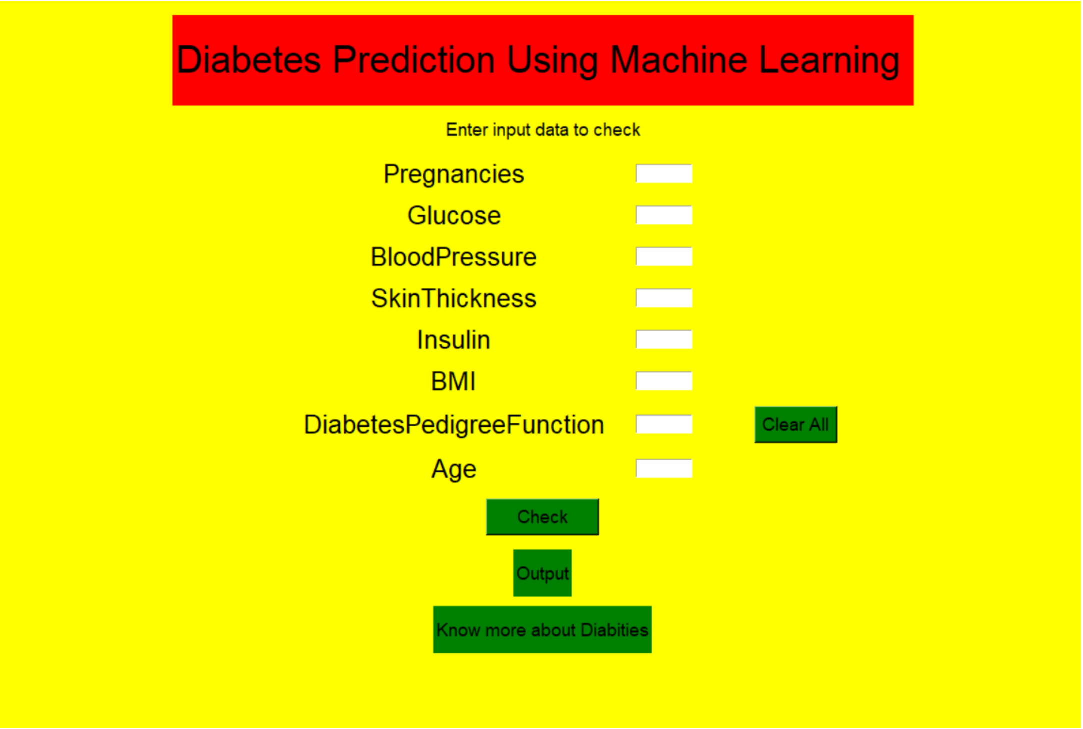


Fig 6.4: Importance of each attributes



Diabetes Prediction Using Machine Learning

Enter input data to check

Pregnancies

Glucose

BloodPressure

SkinThickness

Insulin

BMI

DiabetesPedigreeFunction

Age

Fig 6.5 User interface in tkinter

CHAPTER 7

CONCLUSION

In this research work, a replacement hybrid data mining model for classification and prediction of medical datasets using LR and ANN with FSM like FS and BE has been proposed. The reduction in the number of attributes in data mining helps in building efficient healthcare solutions in medical field. For selecting the important attributes three methods are implemented: Entropy Evaluation method, Mean Evaluation method, and Threshold Evaluation method on medical datasets such as: Pima Indian polygenic disease disorder dataset and Spectf information set chosen from UCI data repository. FSMs are applied to generate the various subset of attributes. For each subset I have evaluated the completion of the ANN, LR and using CVS and also the percentage split as test alternatives.

From the experimental I can identify that:

- For Pima Indian diabetes dataset, the best classification accuracy 83.33% is attained by Threshold Evaluation method by using ANN with percentage split of 75% and only 4 attributes. Result obtained is best than to the accuracy attained for full set of attributes, accuracy attained for Entropy Evaluation method, and Mean Evaluation methods of the proposed hybrid model and also better than some of the existing models.

A new space of analysis in machine learning is that the deep learning. This permits the deep learning process models that embody the multiple of the process layers. The illustration of data with multiple level of abstraction is learned by deep learning technique. Deep learning tries to find intricate structure in massive datasets and by with ever-changing the parameters internally that are to compute representation in every layer enhances learning. Hence backpropagation-based algorithms may be implemented on medical datasets.

REFERENCES

[1] S. Sumathi and S. N. Sivanandam. “Introduction to Data Mining and its Applications”, Studies in Computational Intelligence, Volume 29, Springer, 2006.

[2] R. Tamilselvi and S. Kalaiselvi. “An Overview of Data Mining Techniques and Applications”, International Journal of Science and Research (IJSR), Volume 2, Issue 2, pages 506-509, 2013.

[3] S. D. Gheware, A. S. Kejkar and S. M. Tondare. “Data Mining: Task, Tools, Techniques and Applications”, International Journal of Advanced Research in Computer and Communication Engineering, Volume 3, Issue 10, pages 8095-8098, 2014.

[4] R. Robu and C. Hora. “Medical Data Mining with Extended WEKA”, IEEE 16th International Conference on Intelligent Engineering Systems (INES 2012), pages 347-350, 2012.

[5] Mrs. Bharati M. Ramageri. “Data Mining Techniques and Applications”, Indian Journal of Computer Science and Engineering, Volume 1, Issue 4, pages 301-305, 2010.

[6] Siri Krishan Wasan, Vasudha Bhatnagar and Harleen Kaur. “The Impact of Data Mining Techniques on Medical Diagnostics”, Data Science Journal, Volume 5, pages 119-126, 2006.

[7] Ashwinkumar U.M. and Dr. Anandakumar K.R. “Ethical and Legal Issues for Medical Data Mining”, International Journal of Computer Applications, Volume 1, Issue 28, pages 7-11, 2010.

[8] Raghavendra B. K. and Dr. Jay B. Simha. “Evaluation of Logistic Regression Model with Feature Selection Methods on Medical Datasets”, ACS-International Journal on Computational Intelligence, Volume 1, Issue 2, pages 35-42, 2010.

[9] J. Chhatwal, O. Alagoz, M. J. Lindstorm, C. E. Kahn, K.A. Shaffer and E.S. Burnside. “A Logistic Regression Model Based on the National Mammography Database Format to Aid Breast Cancer Diagnosis”, American Journal of Roentgenology, Volume 192, Issue 4, pages 1117-1127, 2009.

[10] H. Khedmat, G. R. Karami, V. Pourfarziani, S. Assari, M. Rezailashkajani and M. M. Naghizadeh. “A Logistic Regression Model for Predicting Health-Related Quality of Life in Kidney Transplant Recipients”, Transplantation Proceedings, Elsevier, Volume 39, pages 917-922, 2007.

[11] Riccardo Bellazzi and Blaz Zupan. “Predictive Data Mining in Clinical Medicine: Current Issues and Guidelines”, International Journal of Medical Informatics, Elsevier, Volume 77, Issue 2, pages 81-97, 2008.

[12] Raghavendra S. and Dr. Indiramma M. “Performance Evaluation of Logistic Regression and Artificial Neural Network Model with Feature Selection Methods using Cross Validation Sample and Percentage Split on Medical Datasets”, Proceedings of the 2nd International Conference on Emerging Research in Computing, Information Communication and Applications, Volume 2, Elsevier Publication, pages 750-755, 2014.

[13] Raghavendra B.K. and Jay B. Simha. “Performance Evaluation of Logistic Regression and Neural Network Model with Feature Selection Methods and Sensitivity Analysis on Medical Data Mining”, International Journal of Advanced Engineering Technology, Volume 2, Issue 1, pages 289-298, 2011.

[14] Richard P. Lippmann. “An Introduction to Computing with Neural Nets”, IEEE ASSP Magazine, Volume 4, Issue 2, pages 4-22, 1987.

[15] Usama Fayyad, Gregory Piatetsky-Shapiro and Padhraic Smyth. “From Data Mining to Knowledge Discovery in Databases”, Artificial Intelligence Magazine, Volume 17, Issue 3, pages 37-54, 1996.

[16] Jonathan C. Prather, David F. Lobach, Linda K. Goodwin, Joseph W. Hales, Marvin L. Hage and W. Edward Hammond. “Medical Data Mining: Knowledge Discovery in a Clinical Data Warehouse”, Proceedings of the AMIA Annual Fall Symposium, pages 101-105, 1997.

[17] Mark A. Hall and Lloyd A. Smith. “Practical Feature Subset Selection for Machine Learning”, Proceedings of the 21st Australian Computer Science Conference, Springer, pages 181-191, 1998.

[18] Bing Xue, Mengjie Zhang, Will N. Browne and Xin Yao. “A Survey on Evolutionary Computation Approaches to Feature Selection”, IEEE Transactions on Evolutionary Computation, Volume 20, Issue 4, pages 606-626, 2016.

[19] Mark A. Hall and Lloyd A. Smith. “Feature Selection for Machine Learning: Comparing a Correlation-Based Filter Approach to the Wrapper”, Proceedings of the 12th International Florida Artificial Intelligence Symposium Conference, AAAI Press, pages 235-239, 1999.

[20] M. Dash and H. Liu. “Feature Selection for Classification”, Intelligent Data Analysis: An International Journal, Elsevier, Volume 1, Issue 3, pages 131-156, 1997.

[21] Il–Seok Oh, Jin-Seon Lee and Byung-Ro Moon. “Hybrid Genetic Algorithms for Feature Selection”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 26, Issue 11, pages 1424-1437, 2004.

[22] Raghavendra B K. and Jay B. Simha. "Evaluation of Feature Selection Methods for Predictive Modeling Using Neural Networks in Credit Scoring", International Journal of Advanced Networking and Applications, Volume 2, Issue 3, pages 714-718, 2010.

[23] Iffat A. Gheyas and Leslie S. Smith. "Feature Subset Selection in Large Dimensionality Domains", Pattern Recognition, Volume 43, Issue 1, pages 5-13, 2010.

[24] Michael L. Raymer, Travis E. Doom, Leslie A. Kuhn and William F. Punch. "Knowledge Discovery in Medical and Biological Datasets Using a Hybrid Bayes Classifier/Evolutionary Algorithm", IEEE Transactions on Systems, Man and Cybernatics, Volume 33, Issue 5, pages 802-813, 2003.

[25] Peter Sykacek and Stephen Roberts. "Adaptive Classification by Variational Kalman Filtering", Advances in Neural Information Processing Systems (NIPS), pages 737-744, 2002.

[26] Stavros J. Perantonis and Vassilis Virvilis. "Input Feature Extraction for Multilayered Perceptrons Using Supervised Principal Component Analysis", Neural Processing Letters, Volume 10, Issue 3, pages 243-252, 1999.

[27] Liping Wei and Russ B. Altman. "An Automated System for Generating Comparative Disease Profiles and Making Diagnoses", Section on Medical Informatics, Stanford University School of Medicine, MSOB X215, 2004.

[28] Ottar Hellevik. "Linear versus Logistic Regression When the Dependent Variable is a Dichotomy", Quality and Quantity: International Journal of Methodology, Springer, Volume 43, Issue 1, pages 59-74, 2009.

[29] Muhammad Kamran Bodla, Sarmad Majeed Malik, Muhammad Tahir Rasheed , Muhammad Numan, Muhammed Zeeshan Ali and Jimmy Baimba Brima. "Logistic Regression and Feature Extraction Based Fault Diagnosis of Main Bearing of Wind Turbines", IEEE 11th International Conference on Industrial Electronics and Applications (ICIEA), pages1628-1633, 2016.

[30] Tam Nguyen, Raviv Raich and Phung Lai, "Jeffreys Prior Regularization for Logistic Regression", IEEE Statistical Signal Processing Workshop (SSP), pages 1-5, 2016.

[31] Anchana Khemphila and Veera Boonjing. "Comparing Performances of logistic Regression, Decision Trees and Neural Networks for classifying Heart Disease Patients", International Conference on Computer Information Systems and Industrial Management Applications (CISIM), pages 193-198, 2010.

[32] C. P. Prathibhamol, K. V. Jyothy and B. Noora. "Multi Label Classification Based on Logistic Regression (MLC-LR)", International Conference on Advances in Computing, Communications and Informatics (ICACCI), pages 2708-2712, 2016.

[33] Pranav Rao and Manikandan J. “Design and Evaluation of Logistic Regression Model for Pattern Recognition Systems”, IEEE Annual India Conference (INDICON), pages 1-6, 2016.

[34] Chao-Ying Joanne Peng , Kuk Lida Lee and Gary M. Ingersoll. “An Introduction to Logistic Regression Analysis and Reporting”, The Journal of Educational Research, Volume 96, Issue 1, pages 3-14, 2002.

[35] S. Le Cessie and J. C. Van Houwelingen. “Ridge Estimators in Logistic Regression”, Applied Statistics, Volume 41, Issue 1, pages 191-201, January 1992.

[36] Pia Veldt Larsen. “Regression and Analysis of Variance”, Master of Applied Statistics, pages 1-13, 2008.

[37] Laurene Fausett. “Fundamentals of Neural Networks: Architectures, Algorithms and Applications”, 3rd Edition, pages 19-44, 2005.

[38] Johnson RC and Brown C. “Cognizers: Neural Networks and Machines that Think”, 1st Edition, Wiley, 1988.

[39] Ranjit Abraham, Jay B. Simha and S. Sitharama Iyengar. “Effective Discretization and Hybrid Feature Selection Using Naïve Bayesian Classifier for Medical Data Mining”, International Journal of Computational Intelligence Research, Volume 5, Issue 2, pages 116-129, 2009.

[40] Qi Cheng, Pramod K. Varshney and Manoj K. Arora. “Logistic Regression for Feature Selection and Soft Classification of Remote Sensing Data”, IEEE Geoscience and Remote Sensing Letters, Volume 3, Issue 4, page 491-494, 2006.

[41] Huan Liu and Rudy Setiono. “Some Issues on Scalable Feature Selection”, Expert Systems with Applications”, Volume 15, Issue 3-4, pages 333-339, 1998.

[42] Seymour Geisser. “Predictive Inference”, Chapman and Hall, ISBN *0-412-03471-9, 1993.*

[43] Ron Kohavi. “A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection”, Proceedings of the 14th International Conference Joint Conference on Artificial Intelligence, Volume 14, Issue 12, pages 1137-1143, 1995.

[44] Pierre A. Devijver, Josef Kittler. “Pattern Recognition: A Statistical Approach”, Prentice-Hall, 1982.

[45] Robert Grossman, Giovanni Seni, John F. Elder, Nitin Agarwal and Huan Liu. “Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions”, Morgan & Claypool, 2010.

[46] Werner Dubitzky, Martin Granzow and Daniel P. Berrar. “Fundamentals of Data Mining in Genomics and Proteomics”, Springer Science and Business Media, page 178, 2007.

[47] Richard R. Picard and R. Dennis Cook. "Cross-Validation of Regression Models". Journal of the American Statistical Association, Volume 79, Issue 387, page 575-583, 1984.

[48] S. Piramuthu. "Evaluating Feature Selection Methods for Learning in Data Mining Applications". European Journal of Operational Research, Elsevier Publication, Volume 156, pages 483-494, 2004.

[49] Meliha Handzic, Felix Tjandrawibawa and Julia Yeo. "How Neural Networks can Help Loan Officers to Make Better Informed Application Decisions", Informing Science, pages 97-109, 2003.

[50] Paul O' Dea, Josephine Griffith and Colm O' Riordan. "Combining Feature Selection and Neural Networks for Solving Classification Problems", Technical Report, Information Technology Department, National University of Ireland, pages 1-10, 2001.

[51] Sameer Singh, Jeremy Kubica, Scott Larsen and Daria Sorokina. "Parallel Large Scale Feature Selection for Logistic Regression", SIAM International Conference on Data Mining (SDM), pages 1171-1182, 2009.

[52] Rich Caruana, Nikos Karampatziakis and Ainur Yessenalina. "An Empirical Evaluation of Supervised Learning in High Dimensions", Proceedings of the 25th International Conference on Machine Learning, pages 96-103, 2008.

[53] Paul Komarek and Andrew W. Moore. "Making Logistic Regression A Core Data Mining Tool with TR-IRLS", Proceedings of the 5th IEEE International Conference on Data Mining, pages 685-688, 2005.

[54] Paul R. Komarek and Andrew W. Moore. "Fast Robust Logistic Regression for Large Sparse Datasets with Binary Outputs", Proceedings of the 9th International Workshop on Artificial Intelligence and Statistics, pages 174-182, 2003.

[55] Wray Buntine. "Learning Classification Trees", Statistics and Computing, Springer, Volume 2, Issue 2, pages 63-73, 1992.

[56] James Dougherty, Ron Kohavi and Mehran Sahami. "Supervised and Unsupervised Discretization of Continuous Features", Proceedings of the 12th International Conference on Machine Learning, pages 194-202, 1995.

[57] Bojan Cestnik. "Estimating Probabilities: A Crucial Task in Machine Learning", Proceedings of the 9th European Conference on Artificial Intelligence, pages 147-149, 1990.

[58] Christopher J. C. Burges. "A Tutorial on Support Vector Machines for Pattern Recognition", Data Mining and Knowledge Discovery, Volume 2, pages 121-167, 1998.

[59] Luis Mariano Esteban Escano, Gerardo Sanz Saiz, Francisco Javier Lopez Lorente, Angel Borque Fernando and Jose Maria Vergara Ugarrizza. “Logistic Regression Versus Neural Networks for Medical Data”, Monografías del Seminario Matemático García de Galdeano, Volume 33, pages 245-252, 2006.

[60] Norma Terrin, Christopher H. Schmid, John L. Griffith, Ralph B. D. Agostino and Harry P. Selker. “External Validity of Predictive Models: A Comparison of Logistic Regression, Classification Trees and Neural Networks” Journal of Clinical Epidemiology, Elsevier, Volume 56, Issue 8, pages 721-729, 2003.

[61] William G. Baxt. “Use of an Artificial Neural Network for Data Analysis in Clinical Decision-making: The Diagnosis of Acute Coronary Occlusion”, Neural Computation, Volume 2, Issue 4, pages 480-489, 1990.

[62] Bahar Tasdelen, Sema Helvaci, Hakan Kaleagasi and Aynur Ozge. “Artificial Neural Network Analysis for Prediction of Headache Prognosis in Elderly Patients”, Turkish Journal of Medical Sciences, Volume 39, Issue 1, pages 5-12, 2009.

[63] Ananya Das, Tamir Ben-Menachem, Gregory S. Cooper, Amithab Chak, Micheal V. Sivak, Judith A. Gonet and Richard CK Wong. “Prediction of Outcome in Acute Lower-gastrointestinal Haemorrhage Based on an Artificial Neural Network: Internal and External Validation of a Predictive Model”, The Lancet, Volume 362, Issue 9392, pages 1261-1266, 2003.

[64] Le Xu, Mo-Yuen Chow and X. Z. Gao. “Comparisons of Logistic Regression and Artificial Neural Network on Power Distribution Systems Fault Cause Identification”, Proceedings of IEEE Mid-Summer Workshop on Soft Computing in Industrial Applications SMCia/05, pages 128-131, 2005.

[65] Fariba Shadabi and Dharmendra Sharma. “Comparison of Artificial Neural Networks with Logistic Regression in Prediction of Kidney Transplant Outcomes”, Proceedings of the 2009 International Conference of Future Computer and Communication (ICFCC), pages 543-547, 2009.

[66] Behzad Eftekhar, Kazem Mohammad, Hassan Eftekhar Ardebili, Mohammad Ghodsi and Ebrahim Ketabchi. “Comparison of Artificial Neural Network and Logistic Regression Models for Prediction of Mortality in Head Trauma Based on Initial Clinical Data”, BMC Medical Informatics and Decision Making, Volume 5, Issue 3, pages 1-8, 2005.

[67] V. S. Bourdes, S. Bonnevay, P.J.G. Lisoba, M.S.H. Aung, S. Chabaud, T. Bachelot, D. Perol and S. Negrier. “Breast Cancer Predictions by Neural Networks Analysis: A Comparison with Logistic Regression”, Proceedings of the 29th International

Conference of the IEEE Engineering in Medicine and Biology Society, pages 5424-5427, 2007.

[68] Jack V. Tu and Michael R.J. Guerriere. “Use of a Neural Network as a Predictive Instrument for Length of Stay in the Intensive Care Unit Following Cardiac Surgery”, Computers and Biomedical Research, Volume 26, Issue 3, pages 220-229, 1993.

[69] Turgay Ayer, Oguzhan Alagoz, Jagpreet Chhatwal, Jude W. Shavlik, Charles .E. Kahn, and Elizabeth S. Burnside. ”Breast Cancer Risk Estimation with Artificial Neural Network Revisited: Descrimination and Calibration”, Cancer, Volume 116, Issue 14, pages 3310-3321, 2010.

[70] Paulo J. Lisboa, and Azzam F. G. Taktak. “The Use of Artificial Neural Networks in Decision Support in Cancer: A Systematic Review”, Neural Networks, Volume 19, pages 408-415, 2006.

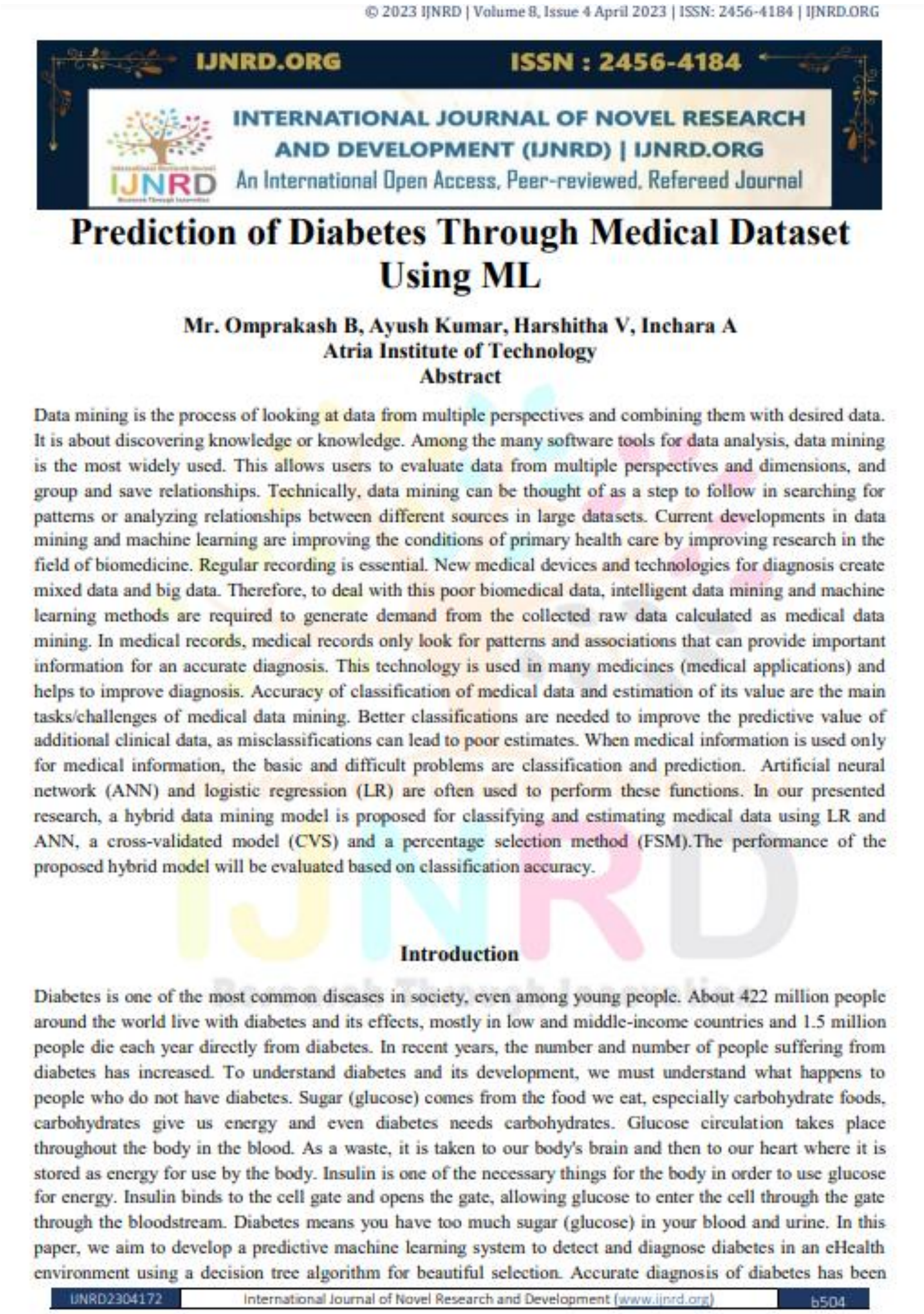
[71] Peter W. F. Wilson, Ralph B. D’Agostino, Daniel Levy, Albert M. Belanger, Halit Silbershatz and William B. Kannel. “Prediction of Coronary Heart Disease Using Risk Factor Categories”, Circulation, Volume 97, pages 1837-1847, 1998.

[72] Jack R. Brzezinski and George J. Knafl. “Logistic Regression Modeling for Context-Based Classification”, Proceedings of 10th International Workshop on Database and Expert Systems Applications (DEXA 99), pages 755-759, 1999.

[73] H. Seker, M. O. Odetayo, D. Petrovic, R. N. G. Naguib, C. Bartoli, L. Alasio, M. S. Lakshmi, G. V. Sherbet and O. R. Hinton. “An Artificial Neural Network Based Feature Evaluation Index for the Assessment of Clinical Factors in Breast Cancer Survival Analysis”, Proceedings of the IEEE Canadian Conference on Electrical and Computer Engineering, Volume 2, pages 1211-1215, 2002.

[74] Munevver Kokuer, Raouf N.G. Naguib, Peter Jancovic, H. Banfield Younghusband and Roger Green. “A Comparison of Multi-Layer Neural Network and Logistic Regression in Hereditary Non-Polyposis Colorectal Cancer Risk Assessment”, Proceedings of the IEEE 27th Annual Conference on Engineering in Medicine and Biology, pages 2417-2420, 2005.

[75] P. J. G. Lisboa and H. Wong. “Are Neural Networks Best Used to Help Logistic Regression? An Example from Breast Cancer Survival Analysis”, IEEE Transactions on Neural Networks, Volume 4, pages 2472-2477, 2001.



© 2023 IJNRD | Volume 8, Issue 4 April 2023 | ISSN: 2456-4184 | IJNRD.ORG

IJNRD.ORG ISSN : 2456-4184

INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG

An International Open Access, Peer-reviewed, Refereed Journal

Prediction of Diabetes Through Medical Dataset Using ML

Mr. Omprakash B, Ayush Kumar, Harshitha V, Inchara A

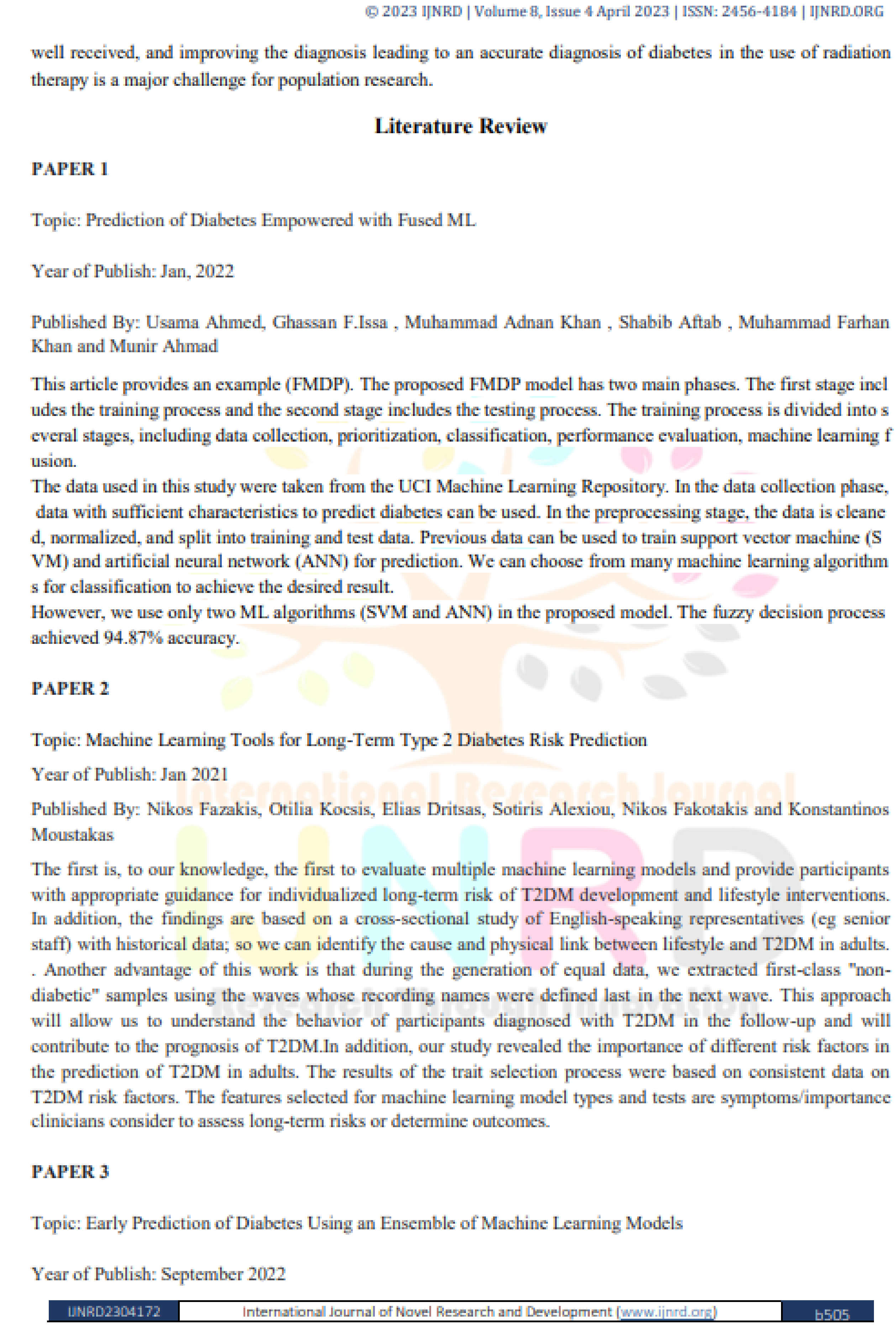
Atria Institute of Technology

Abstract

Data mining is the process of looking at data from multiple perspectives and combining them with desired data. It is about discovering knowledge or knowledge. Among the many software tools for data analysis, data mining is the most widely used. This allows users to evaluate data from multiple perspectives and dimensions, and group and save relationships. Technically, data mining can be thought of as a step to follow in searching for patterns or analyzing relationships between different sources in large datasets. Current developments in data mining and machine learning are improving the conditions of primary health care by improving research in the field of biomedicine. Regular recording is essential. New medical devices and technologies for diagnosis create mixed data and big data. Therefore, to deal with this poor biomedical data, intelligent data mining and machine learning methods are required to generate demand from the collected raw data calculated as medical data mining. In medical records, medical records only look for patterns and associations that can provide important information for an accurate diagnosis. This technology is used in many medicines (medical applications) and helps to improve diagnosis. Accuracy of classification of medical data and estimation of its value are the main tasks/challenges of medical data mining. Better classifications are needed to improve the predictive value of additional clinical data, as misclassifications can lead to poor estimates. When medical information is used only for medical information, the basic and difficult problems are classification and prediction. Artificial neural network (ANN) and logistic regression (LR) are often used to perform these functions. In our presented research, a hybrid data mining model is proposed for classifying and estimating medical data using LR and ANN, a cross-validated model (CVS) and a percentage selection method (FSM).The performance of the proposed hybrid model will be evaluated based on classification accuracy.

Introduction

Diabetes is one of the most common diseases in society, even among young people. About 422 million people around the world live with diabetes and its effects, mostly in low and middle-income countries and 1.5 million people die each year directly from diabetes. In recent years, the number and number of people suffering from diabetes has increased. To understand diabetes and its development, we must understand what happens to people who do not have diabetes. Sugar (glucose) comes from the food we eat, especially carbohydrate foods, carbohydrates give us energy and even diabetes needs carbohydrates. Glucose circulation takes place throughout the body in the blood. As a waste, it is taken to our body's brain and then to our heart where it is stored as energy for use by the body. Insulin is one of the necessary things for the body in order to use glucose for energy. Insulin binds to the cell gate and opens the gate, allowing glucose to enter the cell through the gate through the bloodstream. Diabetes means you have too much sugar (glucose) in your blood and urine. In this paper, we aim to develop a predictive machine learning system to detect and diagnose diabetes in an eHealth environment using a decision tree algorithm for beautiful selection. Accurate diagnosis of diabetes has been



well received, and improving the diagnosis leading to an accurate diagnosis of diabetes in the use of radiation therapy is a major challenge for population research.

Literature Review

PAPER 1

Topic: Prediction of Diabetes Empowered with Fused ML

Year of Publish: Jan, 2022

Published By: Usama Ahmed, Ghassan F.Issa , Muhammad Adnan Khan , Shabib Aftab , Muhammad Farhan Khan and Munir Ahmad

This article provides an example (FMDP). The proposed FMDP model has two main phases. The first stage includes the training process and the second stage includes the testing process. The training process is divided into several stages, including data collection, prioritization, classification, performance evaluation, machine learning fusion.

The data used in this study were taken from the UCI Machine Learning Repository. In the data collection phase, data with sufficient characteristics to predict diabetes can be used. In the preprocessing stage, the data is cleaned, normalized, and split into training and test data. Previous data can be used to train support vector machine (SVM) and artificial neural network (ANN) for prediction. We can choose from many machine learning algorithms for classification to achieve the desired result.

However, we use only two ML algorithms (SVM and ANN) in the proposed model. The fuzzy decision process achieved 94.87% accuracy.

PAPER 2

Topic: Machine Learning Tools for Long-Term Type 2 Diabetes Risk Prediction

Year of Publish: Jan 2021

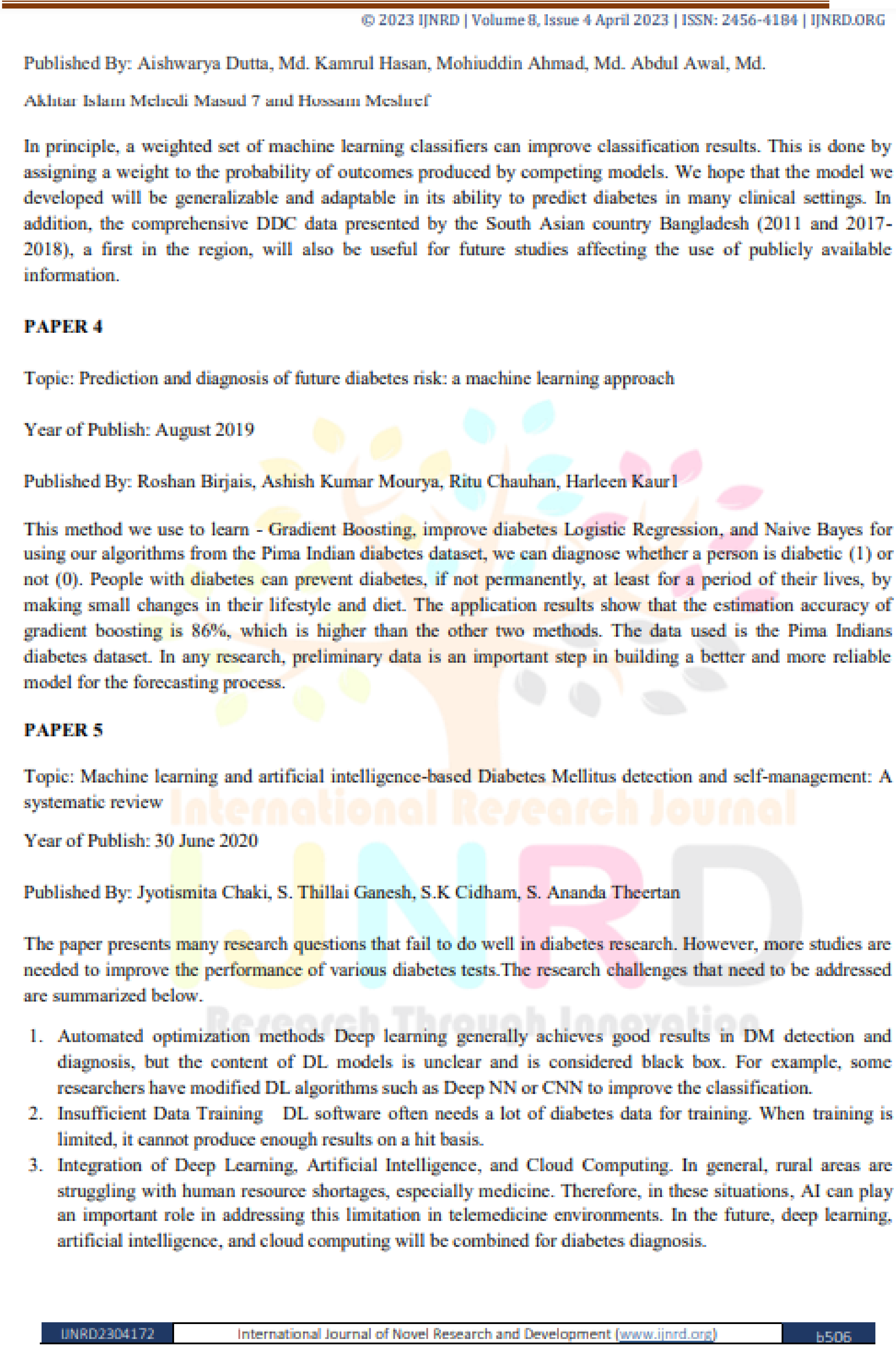
Published By: Nikos Fazakis, Otilia Kocsis, Elias Dritsas, Sotiris Alexiou, Nikos Fakotakis and Konstantinos Moustakas

The first is, to our knowledge, the first to evaluate multiple machine learning models and provide participants with appropriate guidance for individualized long-term risk of T2DM development and lifestyle interventions. In addition, the findings are based on a cross-sectional study of English-speaking representatives (eg senior staff) with historical data; so we can identify the cause and physical link between lifestyle and T2DM in adults. . Another advantage of this work is that during the generation of equal data, we extracted first-class "non-diabetic" samples using the waves whose recording names were defined last in the next wave. This approach will allow us to understand the behavior of participants diagnosed with T2DM in the follow-up and will contribute to the prognosis of T2DM.In addition, our study revealed the importance of different risk factors in the prediction of T2DM in adults. The results of the trait selection process were based on consistent data on T2DM risk factors. The features selected for machine learning model types and tests are symptoms/importance clinicians consider to assess long-term risks or determine outcomes.

PAPER 3

Topic: Early Prediction of Diabetes Using an Ensemble of Machine Learning Models

Year of Publish: September 2022



Published By: Aishwarya Dutta, Md. Kamrul Hasan, Mohiuddin Ahmad, Md. Abdul Awal, Md. Akhtar Islam Mehedi Masud 7 and Hossam Meshref

In principle, a weighted set of machine learning classifiers can improve classification results. This is done by assigning a weight to the probability of outcomes produced by competing models. We hope that the model we developed will be generalizable and adaptable in its ability to predict diabetes in many clinical settings. In addition, the comprehensive DDC data presented by the South Asian country Bangladesh (2011 and 2017-2018), a first in the region, will also be useful for future studies affecting the use of publicly available information.

PAPER 4

Topic: Prediction and diagnosis of future diabetes risk: a machine learning approach

Year of Publish: August 2019

Published By: Roshan Birjais, Ashish Kumar Mourya, Ritu Chauhan, Harleen Kaur1

This method we use to learn - Gradient Boosting, improve diabetes Logistic Regression, and Naive Bayes for using our algorithms from the Pima Indian diabetes dataset, we can diagnose whether a person is diabetic (1) or not (0). People with diabetes can prevent diabetes, if not permanently, at least for a period of their lives, by making small changes in their lifestyle and diet. The application results show that the estimation accuracy of gradient boosting is 86%, which is higher than the other two methods. The data used is the Pima Indians diabetes dataset. In any research, preliminary data is an important step in building a better and more reliable model for the forecasting process.

PAPER 5

Topic: Machine learning and artificial intelligence-based Diabetes Mellitus detection and self-management: A systematic review

Year of Publish: 30 June 2020

Published By: Jyotismita Chaki, S. Thillai Ganesh, S.K Cidham, S. Ananda Theertan

The paper presents many research questions that fail to do well in diabetes research. However, more studies are needed to improve the performance of various diabetes tests.The research challenges that need to be addressed are summarized below.

1. Automated optimization methods Deep learning generally achieves good results in DM detection and diagnosis, but the content of DL models is unclear and is considered black box. For example, some researchers have modified DL algorithms such as Deep NN or CNN to improve the classification.
2. Insufficient Data Training DL software often needs a lot of diabetes data for training. When training is limited, it cannot produce enough results on a hit basis.
3. Integration of Deep Learning, Artificial Intelligence, and Cloud Computing. In general, rural areas are struggling with human resource shortages, especially medicine. Therefore, in these situations, AI can play an important role in addressing this limitation in telemedicine environments. In the future, deep learning, artificial intelligence, and cloud computing will be combined for diabetes diagnosis.

© 2023 IJNRD | Volume 8, Issue 4 April 2023 | ISSN: 2456-4184 | IJNRD.ORG

Methodology

Here is how the LR algorithm uses the PIMA dataset to predict diabetes:

Data preparation: First, the PIMA dataset must be prepared for use by the LR algorithm. This includes cleaning and prioritizing data to eliminate missing or negative outcomes and evaluating inputs for greater consistency.

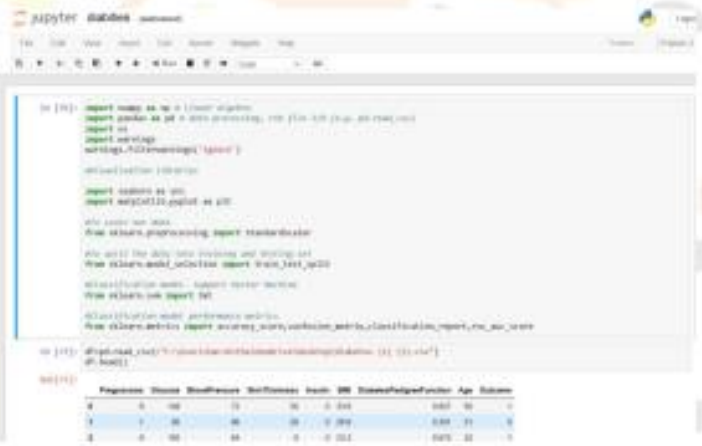
Model training: After the data set is prepared, the LR algorithm can train the data. The LR algorithm works by finding the best-fit coefficients of the equation that divides the data into two classes, in this case diabetic and non-diabetic.

Model Evaluation: Once the LR algorithm has been trained, it can be evaluated on separate datasets to determine its accuracy in predicting diabetes. This test involves comparing the predicted text generated by the LR algorithm with the actual text in the test data.

Perform Evaluation: Once the LR model has been trained and evaluated, it can be used to make predictions on new data. The LR algorithm uses the patient's concept values to predict whether a new patient has diabetes, uses the coefficients learned during training, and calculates the probability that the patient has diabetes. If the result is higher than a threshold, the algorithm classifies the patient as having diabetes mellitus; otherwise, the patient is classified as non-diabetic.

Logistic Regression (LR) is a machine learning algorithm often used for binary classification functions such as predicting diabetes onset in a PIMA dataset. For the PIMA dataset, LR can be used to predict whether a patient has diabetes (labeled 1) or not (labeled 0) based on the values of the input data.

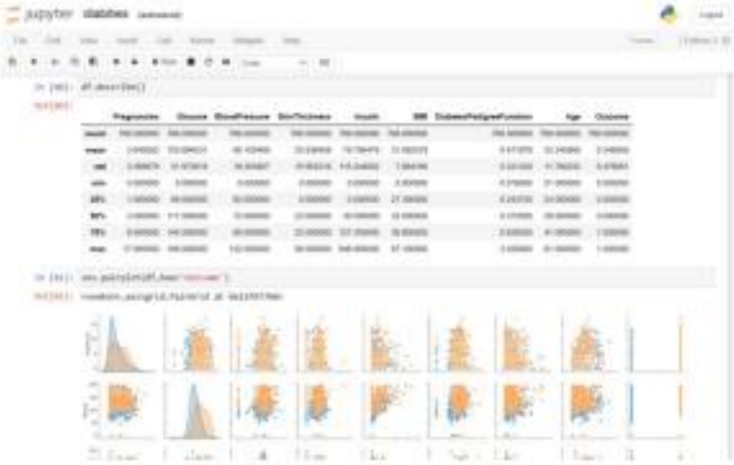
Results and Discussion



```
In [16]: import numpy as np # linear algebra
         import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
         import os
         import warnings
         warnings.filterwarnings('ignore')

         #visualization libraries

         import seaborn as sns
         import matplotlib.pyplot as plt

         #to scale our data
         from sklearn.preprocessing import StandardScaler

         #to split the data into training and testing set
         from sklearn.model_selection import train_test_split

         #classification model - support vector machine
         from sklearn.svm import SVC

         #classification model performance metrics
         from sklearn.metrics import accuracy_score,confusion_matrix,classification_report,roc_auc_score

In [17]: df=pd.read_csv("[illegible]")
         df.head()

Out[17]:
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1

Fig1: Screenshot of importing libraries

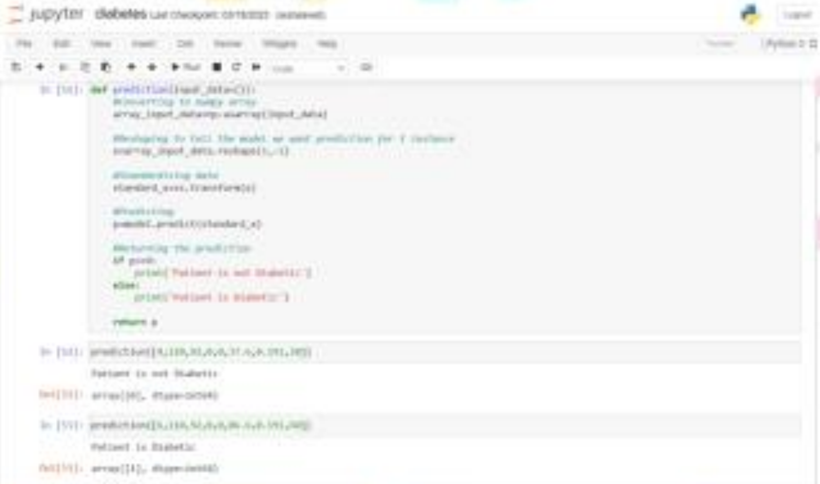


```
In [46]: df.describe()
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
mean	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
std	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
min	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
25%	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
50%	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
75%	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
max	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

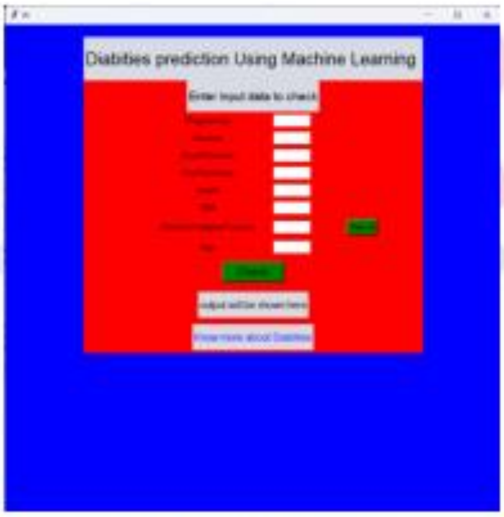
```
In [41]: sns.pairplot(df,hue='Outcome')
Out[41]: <seaborn.axisgrid.PairGrid at [illegible]>
```

Fig 2: Screenshot of attributes in the dataset



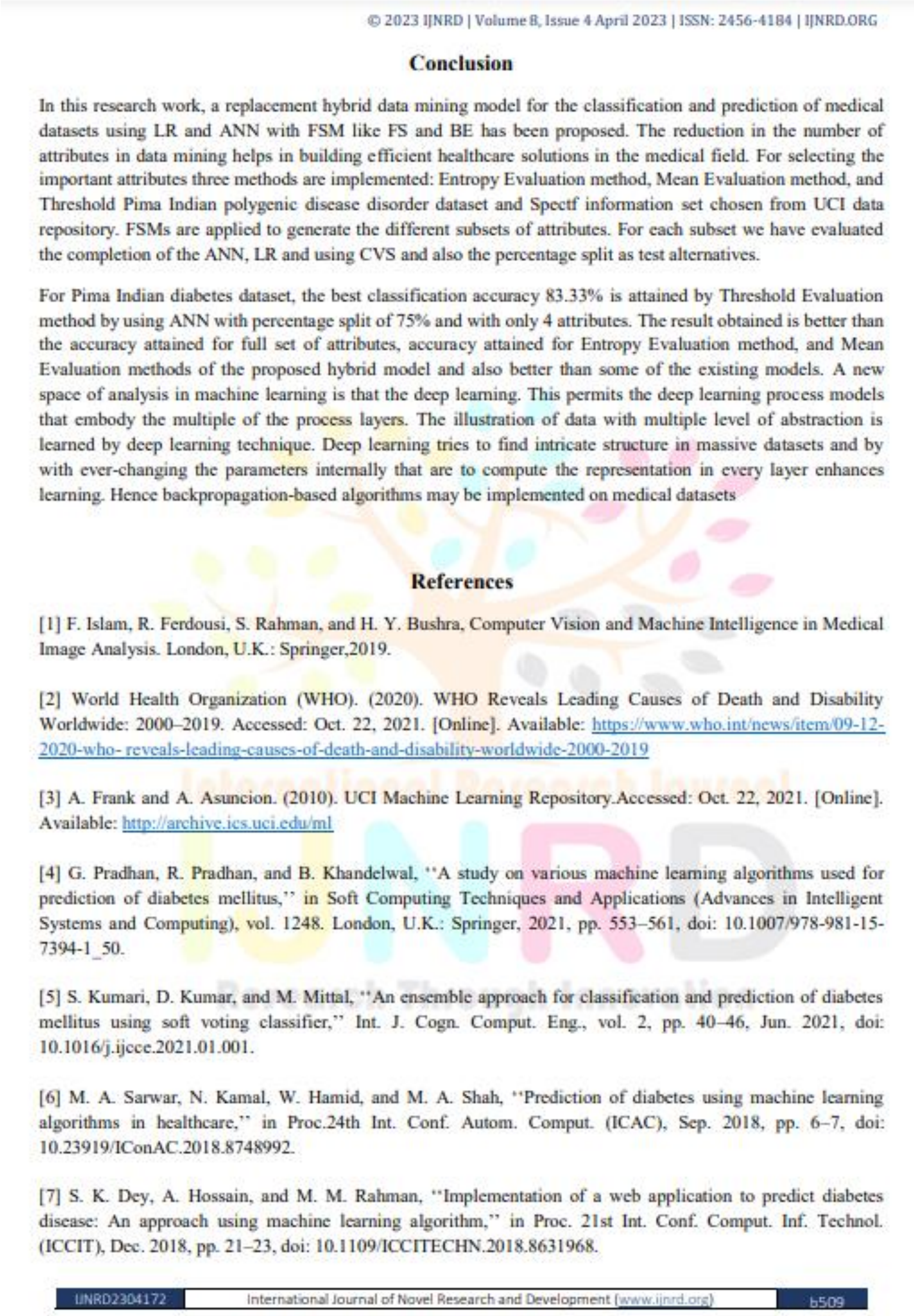
```
In [51]: def prediction(input_data=()):
             #Converting to numpy array
             array_input_data=np.asarray(input_data)

             #Reshaping to tell the model we want prediction for 1 instance
             x=array_input_data.reshape(1,-1)

             #Standardizing data
             standard_x=sc.transform(x)

             #Predicting
             p=model.predict(standard_x)

             #Returning the prediction
             if p==0:
                 print('Patient is not Diabetic')
             else:
                 print('Patient is Diabetic')

             return p

In [53]: prediction([illegible])
         Patient is not Diabetic
Out[53]: array([0], dtype=int64)

In [55]: prediction([illegible])
         Patient is Diabetic
Out[55]: array([1], dtype=int64)
```

Fig 3: Result in Jupyter notebook



Diabities prediction Using Machine Learning
Enter input data to check
Check
output will be shown here
Know more about Diabities

Fig 4: Result in tkinter UI



Conclusion

In this research work, a replacement hybrid data mining model for the classification and prediction of medical datasets using LR and ANN with FSM like FS and BE has been proposed. The reduction in the number of attributes in data mining helps in building efficient healthcare solutions in the medical field. For selecting the important attributes three methods are implemented: Entropy Evaluation method, Mean Evaluation method, and Threshold Pima Indian polygenic disease disorder dataset and Spectf information set chosen from UCI data repository. FSMs are applied to generate the different subsets of attributes. For each subset we have evaluated the completion of the ANN, LR and using CVS and also the percentage split as test alternatives.

For Pima Indian diabetes dataset, the best classification accuracy 83.33% is attained by Threshold Evaluation method by using ANN with percentage split of 75% and with only 4 attributes. The result obtained is better than the accuracy attained for full set of attributes, accuracy attained for Entropy Evaluation method, and Mean Evaluation methods of the proposed hybrid model and also better than some of the existing models. A new space of analysis in machine learning is that the deep learning. This permits the deep learning process models that embody the multiple of the process layers. The illustration of data with multiple level of abstraction is learned by deep learning technique. Deep learning tries to find intricate structure in massive datasets and by with ever-changing the parameters internally that are to compute the representation in every layer enhances learning. Hence backpropagation-based algorithms may be implemented on medical datasets

References

[1] F. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, Computer Vision and Machine Intelligence in Medical Image Analysis. London, U.K.: Springer,2019.

[2] World Health Organization (WHO). (2020). WHO Reveals Leading Causes of Death and Disability Worldwide: 2000–2019. Accessed: Oct. 22, 2021. [Online]. Available: https://www.who.int/news/item/09-12-2020-who- reveals-leading-causes-of-death-and-disability-worldwide-2000-2019

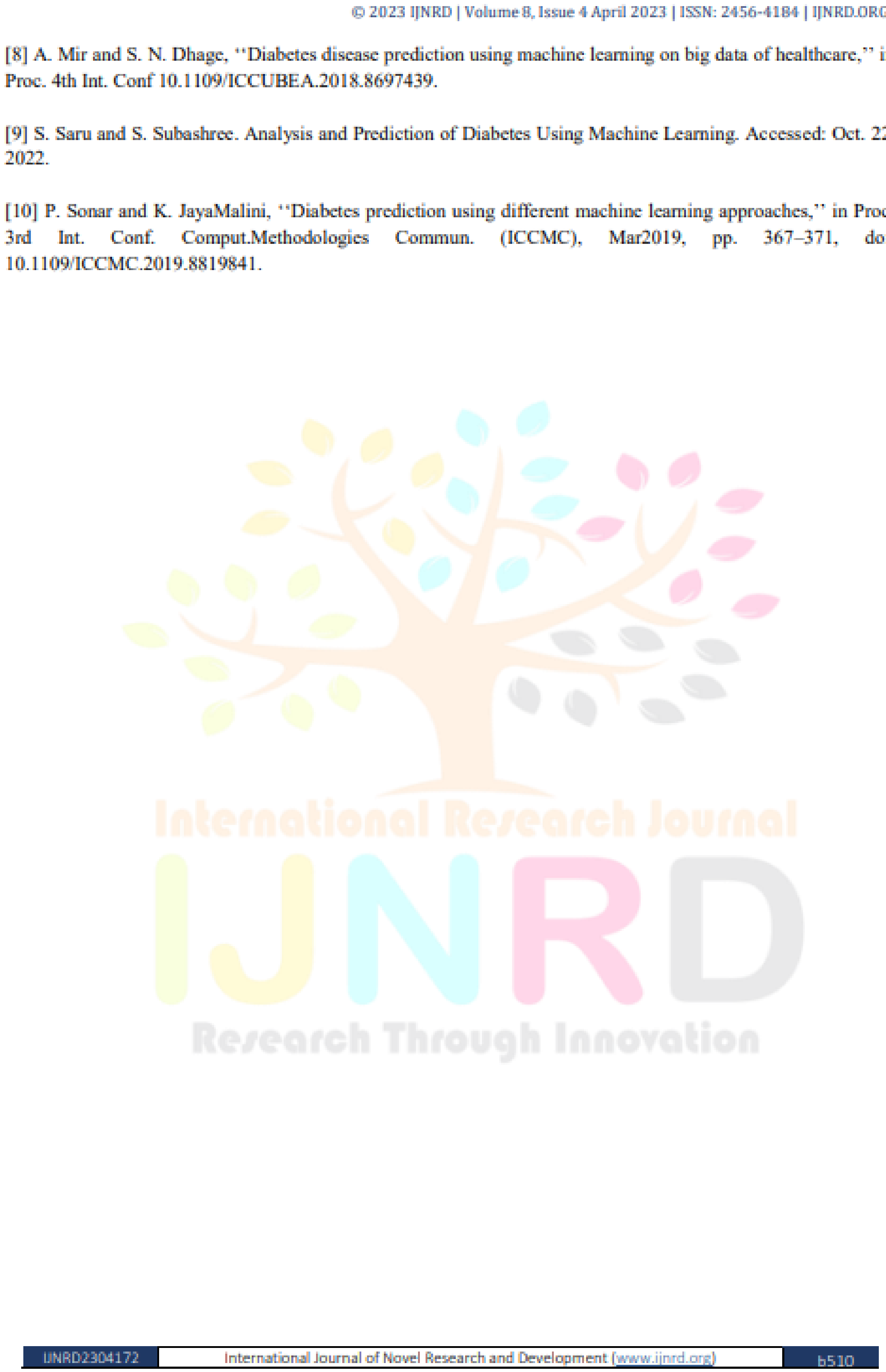
[3] A. Frank and A. Asuncion. (2010). UCI Machine Learning Repository.Accessed: Oct. 22, 2021. [Online]. Available: http://archive.ics.uci.edu/ml

[4] G. Pradhan, R. Pradhan, and B. Khandelwal, ''A study on various machine learning algorithms used for prediction of diabetes mellitus,'' in Soft Computing Techniques and Applications (Advances in Intelligent Systems and Computing), vol. 1248. London, U.K.: Springer, 2021, pp. 553–561, doi: 10.1007/978-981-15-7394-1_50.

[5] S. Kumari, D. Kumar, and M. Mittal, ''An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier,'' Int. J. Cogn. Comput. Eng., vol. 2, pp. 40–46, Jun. 2021, doi: 10.1016/j.ijcce.2021.01.001.

[6] M. A. Sarwar, N. Kamal, W. Hamid, and M. A. Shah, ''Prediction of diabetes using machine learning algorithms in healthcare,'' in Proc.24th Int. Conf. Autom. Comput. (ICAC), Sep. 2018, pp. 6–7, doi: 10.23919/IConAC.2018.8748992.

[7] S. K. Dey, A. Hossain, and M. M. Rahman, ''Implementation of a web application to predict diabetes disease: An approach using machine learning algorithm,'' in Proc. 21st Int. Conf. Comput. Inf. Technol. (ICCIT), Dec. 2018, pp. 21–23, doi: 10.1109/ICCITECHN.2018.8631968.



[8] A. Mir and S. N. Dhage, ''Diabetes disease prediction using machine learning on big data of healthcare,'' in Proc. 4th Int. Conf 10.1109/ICCUBEA.2018.8697439.

[9] S. Saru and S. Subashree. Analysis and Prediction of Diabetes Using Machine Learning. Accessed: Oct. 22, 2022.

[10] P. Sonar and K. JayaMalini, ''Diabetes prediction using different machine learning approaches,'' in Proc. 3rd Int. Conf. Comput.Methodologies Commun. (ICCMC), Mar2019, pp. 367–371, doi: 10.1109/ICCMC.2019.8819841.