RISK PREDICTION OF NON-COMMUNICABLE DISEASE IN EARLY-STAGE

A PROJECT REPORT

Submitted by

ASWIN SIVAKUMAR (960519106023)

R.C.SUTHAN (960519106065)

in partial fulfillment for the award of the degree

of

BACHELOR OF ENGINEERING

IN

ELECTRONICS AND COMMUNICATION ENGINEERING

CAPE INSTITUTE OF TECHNOLOGY ,LEVENJIPURAM

ANNA UNIVERSITY ::CHENNAI 600 025

MAY 2023

BONAFIDE CERTIFICATE

Certified that this project report "**RISK PREDICTION OF NON-COMMUNICABLE DISEASE IN EARLY-STAGE"** is the bonafide work of "**ASWIN SIVAKUMAR (969519106023) , SUTHAN.R.C (960519106065)"** who carried out the project work under my supervision.

SIGNATURE

Mrs. K.JAI DEVI M.E., MISTE.,

HEAD OF THE DEPARTMENT

ASSOCIATE PROFESSOR

Department of Electronics and Communication Engineering

Cape Institute of Technology

Levenjipuram,

Tirunelveli – 627114

SIGNATURE

Ms. V. RANI M.E.,

SUPERVISOR

ASSISTANT PROFESSOR

Department of Electronics and Communication Engineering

Cape Institute of Technology

Levenjipuram,

Tirunelveli - 627114

Submitted for viva –vice examination held at Cape Institute of Technology on ..... Conducted by Anna University Chennai.

INTERNAL EXAMINER **EXTERNAL EXAMINER**

ACKNOWLEDGEMENT

First of all we would like to thank the **GOD almighty** for his blessings and grace which enabled us to complete this project in time.

We wish to thank **Er.I.KRISHNA PILLAI M.Tech.,** our respected and honourable chairman for this encouragement.

We wish to thank **Dr.K.V.IYAPPA KARTHIK B.E., M.B.A.,** our respected prochairman for his excellent support in completion of this project.

We wish to thank **Er.J.B. RENIN JEYA GEM M.E.,** our respected chief Executive officer for his encouragement to accomplish this project.

We express our profound thanks to to our principal **Dr.B.THANUKUMARI M.E., Ph.D.,** for her kind encouragement in completing this project.

We express our sincere thanks to our vice principal **Dr .Dev R NEWLIN M.E., Ph.D.,** for his cooperation and encouragement which help me in completion of this project.

We express our heartfelt and sincere thanks to our HOD **K.JAI DEVI M.E.,** Department of Electronics and Communication Engineering ,Cape Institute of Technology for her advice and encouragement in completing this project.

We express our sincere gratitude and heartfelt thanks to our guide **K.JAIDEVI M.E., MISTE.,** Assistant professor ,Department of Electronics and Communication Engineering for giving valuable suggestions in making this project grand success.

We extend our sincere thanks to the entire faculty in Electronics and communication Engineering Department for their co-operation and encouragement.

Finally ,We would like to express our thanks to our parents and friends who have helped us for the successful completion of our project.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF ABBREVIATION

SVM = Support Vector Machine

NN = Neural Network

DT = Decision Tree

CBGM = Capillary Blood Glucose Measurements

CGM = Continuous Glucose Monitoring

FDA = Food and Drug Administration

DCCT = Diabetes Control and Complications Trial

FBG = Fasting Blood Glucose

ML = Machine Learning

ROC = Receiver Operating Characteristics

FS = Feature Selection

CS = Case Selection

LD = Listwise Detection

SSA = Squirrel Search Algorithm

ABSTRACT

Human body turns the food consumed into energy, but when insulin doesn't act in its way to convert the blood glucose into energy, then the glucose remains in the bloodstream and causes a life-threatening health issue called Diabetes Mellitus or Diabetes. According to the growing morbidity in recent years, in 2040, the world's diabetic patients will reach 642 million, which means that one of the ten adults in the future is suffering from diabetes. There is no doubt that this alarming figure needs great attention. With the rapid development of machine learning, machine learning has been applied to many aspects of medical health. So, for efficiently and effectively diagnosing the Diabetes Mellitus, a method is proposed using the ML Grid Search algorithm. In this method, a database called Pima Indian Diabetic Dataset is used. This system has two phases: the training phase and the test phase. In training phase, preprocessing, feature selection and instance evaluation is done. In test phase, preprocessing, instance evaluation and disease prediction is done. For feature selection, Modified Squirrel Search Algorithm is used and for classification, Support Vector Machine (SVM), a machine learning method as the classifier for diagnosis of diabetes. The machine learning method focus on classifying diabetes disease from high dimensional medical dataset. The experimental results obtained show that support vector machine can be successfully used for diagnosing diabetes disease.

CHAPTER 1

INTRODUCTION

1.1 OVERVIEW

Diabetes is one of the common and rapidly increasing diseases in the world. It is a major health problem in most of the countries. Diabetes is a condition in which your body is unable to produce the required amount of insulin needed to regulate the amount of sugar in the body. This leads to various diseases including heart disease, kidney disease, blindness, nerve damage and blood vessels damage. There are two general reasons for diabetes:

(1) the pancreas does not make enough insulin or the body does not produce enough insulin. Only 5-10 % of people with diabetes have this form of the disease (Type-1).

(2) Cells do not respond to the insulin that is produced (Type-2). Insulin is the principal hormone that regulates uptake of glucose from the blood into most cells (muscle and fat cells).

If the amount of insulin available is insufficient, then glucose will not have its usual effect so that glucose will not be absorbed by the body cells that require it. Diabetes mellitus being one of the major contributors to the mortality rate. Detection and diagnosis of diabetes at an early stage is the need of the day. Diabetes disease diagnosis and interpretation of the diabetes data is an important classification problem. A classifier is required and to be designed that is cost efficient, convenient and accurate. Artificial intelligence and Soft Computing Techniques provide a great deal of human ideologies and are involved in human

related fields of application. These systems find a place in the medical diagnosis. A medical diagnosis is a classification process. A physician has to analyse lot of factors before diagnosing the diabetes which makes physician's job difficult. In recent times, machine learning and data mining techniques have been considered to design automatic diagnosis system for diabetes. Recently, there are many methods and algorithms used to mine biomedical datasets for hidden information including Neural networks (NNs), Decision Trees (DT), Fuzzy Logic Systems, Naive Bayes, SVM, cauterization, logistic regression and so on. These algorithms decrease the time spent for processing symptoms and producing diagnoses, making them more precise at the same time. The Support Vector Machine (SVM) is a novel learning machine introduced first by Vapnik and has been applied in several financial applications recently, mainly in the area of time series prediction and classification.

1.2 BACKGROUND

Diabetes is a chronic disease that occurs when the human pancreas does not produce enough insulin, or when the body cannot effectively use the insulin it produces, which leads to an increase in blood glucose levels. Normally, after a meal, the body breaks the food down into glucose, which is carried by the blood to cells throughout the body. The cells use insulin, a hormone made in the pancreas, to convert the blood glucose into energy. People with diabetes have problems in doing such a conversion leading to fatigue and many other serious complications.

Late diagnosis and/or improper control of diabetes can lead to many serious complications: damage to the eye (leading to blindness), kidney (leading to renal failure), and nerves (leading to impotence and foot disorders with possible

amputation). As well, it increases the risk of heart disease, stroke, and reduces life expectancy. Diabetes has recently become one of the most common diseases around the world, where in year 2000, 171 million people worldwide suffered from diabetes. This number is expected to increase to 366 million by the year 2030. Fig 1.1 shows the prevalence of diabetes around the world in year 2000 and the expected numbers in year 2030. The prevalence of diabetes in the Middle East is expected to increase 164%. It is interesting to note here that within the UAE, where this research is carried out, diabetes currently affects 19.2 per cent of its population, which makes it the tenth highest prevalence ratio worldwide. In addition, diabetes in the Gulf is a regional challenge. Gulf countries such as Kuwait, Qatar, Saudi Arabia, and Bahrain all feature in the top ten countries in diabetes prevalence worldwide. Therefore, finding an effective management process for this disease is of most importance, especially for those countries listed above.

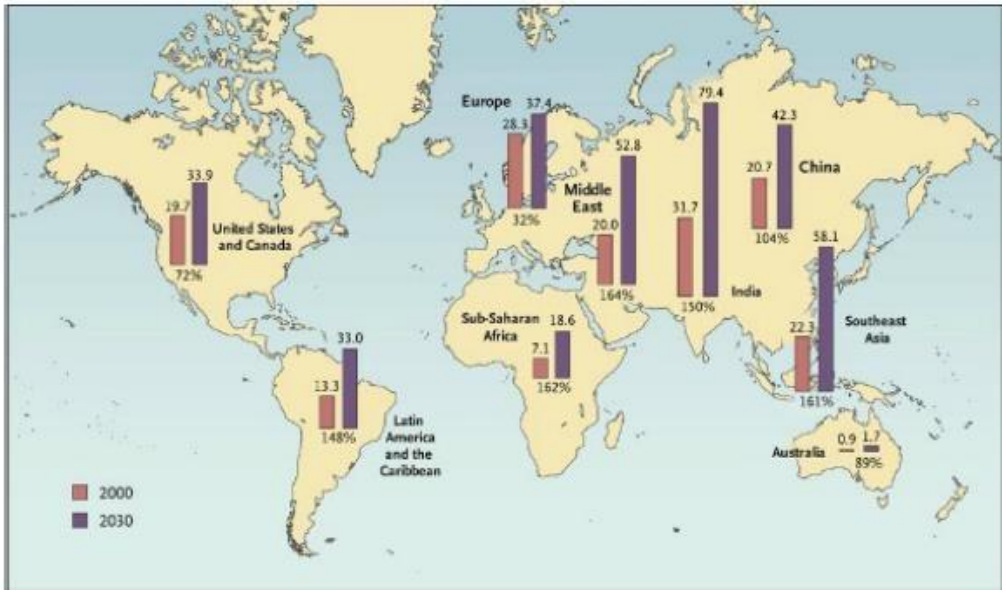

Fig 1.1: Global prevalence of diabetes in 2000 to 2030

There are three types of diabetes, namely:

1. Type 1 Diabetes Mellitus (T1DM) or sometimes called juvenile diabetes, is the type of diabetes that results from stopping the insulin generation by the pancreas beta cells. It is the most severe type of diabetes among all of the other ones. It is prevalent among children and requires several insulin injections each day to bring the patient's glucose levels under control.

2. Type 2 Diabetes Mellitus (T2DM), or the so-called adult-onset diabetes, is the most common type of diabetes, where it compromises 90% of diabetic population worldwide. People can develop T2DM at any age. This form of diabetes usually starts with insulin resistance, which eventually leads to the loss of the pancreas ability to produce enough insulin in response to food intake.

3. Gestational diabetes is the type of diabetes affecting some women during pregnancy only.

There is no cure for diabetes as of yet, nevertheless, an early diagnosis of this disease, followed by a suitable medication, a balanced diet, and regular physical activity go a long way in controlling blood glucose levels and decreasing the risk of developing complications. Controlling the blood glucose level of diabetic patients and keeping it within the normal range (70 mg/dL -120 mg/dL) is therefore the focal goal of physicians. However, the main challenge in blood glucose control is the desire to keep its level as close to the normal range as possible, while keeping the number of hypoglycemia events to a minimum. Hypoglycemia is a condition that occurs when blood glucose drops dangerously low (below 60 mg/dL).

This event may occur due to a number of reasons, such as taking insulin at the wrong time or taking extra dosages, not eating enough during meals or

delaying them, and excessive exercising or changing the time of exercising. Effects of hypoglycemia vary from mild dysphoria to conditions that are more dangerous; these present as seizures, unconsciousness, and possibly permanent brain damage or death. Hypoglycemia is treated simply by the oral intake of carbohydrate food by the diabetic patient. It is interesting to note here that the Diabetes Control and Complications Trial (DCCT)found the occurrence of hypoglycemia is three times higher in intensively insulin-treated group of diabetic patients when compared to the group receiving standard treatment.

Traditionally, self-monitoring of blood glucose requires the drawing of a blood sample several times a day. The need for reducing the number of daily Capillary Blood Glucose Measurements (CBGM) is due to the pain associated with the needles' use. Nowadays, the availability of Continuous Glucose Monitoring (CGM) devices, placed on the patient body, makes it possible to obtain subcutaneous glucose reading information in real time, i.e., every few minutes. These real time measurements reduce the need for the painful CBGM. Many types of CGM sensors have been approved by the Food and Drug Administration (FDA), including, the Freestyle Navigator (Abbott Diabetes Care), the Seven Plus (DexCom), and the Guardian Real-Time (Medtronic Diabetes). CGM are a minimally invasive device whereby it measures the glucose level in the patient’s interstitial fluid, but not in his/her blood streams1. As a result, the measured glucose levels lag 8-10 minutes behind the capillary blood glucose values. The CGM readings, therefore, must be calibrated occasionally against the readings of the traditional glucose meter (finger stick) to reflect the actual blood glucose values.

One of the main advantages of CGM is its ability to detect hypoglycemia and sound an alarm in such conditions. The CGM device detects hypoglycemia using a predefined threshold of glucose level (for example < 60 mg/dl). Sometimes, detecting hypoglycemia is too late for a patient to take corrective actions; therefore, a better approach is to predict such a condition before its occurrence. It is interesting to note here that the development of CGM is considered a breakthrough in diabetes treatment and management fields, due to the wealth of useful information it supplies to the diabetic patients.

1.3 ANALYSIS FROM VARIOUS HEALTH ORGANIZATIONS

World Health Organization (WHO) in the year 2015 their guidelines provide the functionality of the blood sugar level in the human body. Whenever the sugar level or glucose is high in the circulatory system, the beta cells in the pancreas discharges the insulin to the circulation system.

- The people with diabetes rose from 108 million in 1980 to 422 million in 2014. This is the early stage of diabetes which people suffer a lot.
- The global popularity of diabetes among adults over 18 years of age rose from 4.7% in 1980 to 8.5% in 2014. (“Roglic and World Health Organization - 2016 - Global Report on Diabetes.)
- Between 2000 and 2016, there was a slight difference in diabetes condition were reduced a bit but there was a 5% increase in premature mortality from diabetes.
- Mainly Diabetes occurrence has been increasing more rapidly in low and middle- income countries than in high-income countries.

- Diabetes has a major cause of blindness, kidney failure, heart attacks, stroke, and lower limb amputation Nephrology DialysisTransplantation (2019).
- In 2016, an appraised 1.6 million deaths were directly caused by diabetes. Another 2.2 million deaths were attributable to high blood glucose in 2012(Roglic and World Health Organization 2016).
- Almost every half of all deaths attributable to high blood glucose occur before the age of 70 years. World Health Organization approximations that diabetes was the seventh leading cause of death in 2016(Roglic and World Health Organization 2016).
- Having a healthy diet, regular physical activity, maintaining normal body weight, and avoiding tobacco use are ways to prevent ourselves or to delay the onset of type 2 diabetes.
- Diabetes can be treated and its significances avoided or delayed with diet, physical activity, medication, and regular screening and treatment for impediments(Roglic and World Health Organization 2016).
- Diabetes results from the interaction between the genetic predisposition and behavioural and environmental risk factors(Roglic and World Health Organization 2016).

The genetic basis of diabetes is yet to be identified, however, there is strong evidence that such modifiable risk factor as obesity and physical activities are the main non-genetic disease determinants of the disease.

All the types of Diabetes can lead to impediments in many parts of the body, and can increase the overall risk of dying ahead of time. Possible difficulties may include kidney failure, leg amputation, vision loss and nerve damage. Adults with

diabetes have type 1 or type 2 diabetes have a major increased risk of heart attacks and strokes. In pregnancy, poorly controlled diabetes increases the risk of fetal death and other impediments.

Diabetes interferes with the body's ability to process celluloses for energy, leading to high levels of blood sugar. These persistently high blood sugar levels increase a person's risk of developing serious health problems.

1.4 GLOBAL HEALTH AGENDA IN DIABETES

Diabetes is recognized as an important cause of premature death and disability. It is one of four priorities Non-Communicable Diseases (NCDs) targeted by world leaders in the 2011 Political Declaration on the Prevention and Control of NCDs.

The declaration recognizes that the incidence and impacts of diabetes and other NCDs can be largely prevented or reduced with an approach that incorporates evidence-based, affordable, cost-effective, population-wide and multi sectoral interventions. To catalyse national action, the World Health Assembly adopted a comprehensive global monitoring framework in 2013, comprised of nine voluntary global targets to reach by 2025. This was accompanied by the WHO Global action plan for the prevention and control of NCDs 2013 – 2020 (WHO NCD Global Action Plan), endorsed by the 66th World Health Assembly, which provides a roadmap and policy options to attain the nine voluntary global targets. Diabetes and its key risk factors are strongly reflected in the targets and indicators of the global monitoring framework and the WHO NCD Global Action Plan. These commitment s were deepened in 2015 by the United Nations General Assembly’s adoption of the 2030 Agenda for Sustainable

Development. In this context, countries have agreed to take action to achieve ambitious targets by 2030 to reduce premature mortality from NCDs by one-third; to achieve universal health coverage; and to provide access to affordable essential medicines. To halt the rise in obesity and type 2 diabetes it is imperative to scale-up population-level prevention. Policy measures are needed to increase access to affordable, healthy foods and beverages; to promote physical activity; and to reduce exposure to tobacco.

Mass media campaigns and social marketing can influence positive change and make healthy behaviours more the norm. These strategies have the potential to reduce the occurrence of type 2 diabetes and may also reduce complications associated with diabetes. To reduce avoidable mortality from diabetes and improve outcomes, access to affordable treatment is critical. Lack of access to insulin in many countries and communities remains a critical impediment to successful treatment efforts. Inadequate access to oral hypoglycaemic medication, and medication to control blood pressure and lipids, is also a barrier. Improved management in primary care with ongoing support by community health workers can lead to better control of diabetes and fewer complications.

1.5 HISTORY OF DIABETES

Physicians have observed the effects of diabetes for thousands of years. For much of this time, little was known about this fatal disease that caused wasting away of the body, extreme thirst, and frequent urination. It wasn't until 1922 that the first patient was successfully treated with insulin.

One of the effects of diabetes is the presence of glucose in the urine. Ancient Hindu writings, many thousands of years old, document how black ant and flies were attracted to the urine of diabetics. The Indian physician Sushruta in 400 B.C.

described the sweet taste of urine from affected individuals, and for many centuries to come, the sweet taste of urine was key to diagnosis. Around 250 B.C., the name “diabetes” was first used. It is a Greek word that means “to syphon”, reflecting how diabetes seemed to rapidly drain fluid from the affected individual. The Greek physician Aretaeus noted that as affected individuals wasted away, they passed increasing amounts of urine as if there was “liquefaction of flesh and bones into urine”. The complete term “diabetes mellitus” was coined in 1674 by Thomas Willis, personal physician to King Charles II. Mellitus is Latin for honey, which is how Willis described the urine of diabetics (“as if imbued with honey and sugar”).

Up until the mid-1800s, the treatments offered for diabetes varied tremendously. Various “fad” diets were prescribed, and the use of opium was suggested, as were bleeding and other therapies. The most successful treatments were starvation diets in which calorie intake was severely restricted. Naturally, this was intolerable for the patient and at best extended life expectancy for a few years.

A breakthrough in the puzzle of diabetes came in 1889. Surgically removed the pancreas from dogs. The dogs immediately developed diabetes. Now that a link was established between the pancreas gland and diabetes, research focused on isolating the pancreatic extract that could treat diabetes. In 1923, Banting and Macloed were awarded the Nobel Prize for the discovery of insulin. Banting split his prize with Best, and Macloed split his prize with Collip. In his Nobel Lecture, Banting concluded the following about their discovery. During the spring of 1922, Best increased the production of insulin to enable the treatment of diabetic patients coming to the Toronto clinic. Over the next 60 years, insulin was further refined

and purified, and long-acting and intermediate types were developed to provide more flexibility. A revolution came with the production of recombinant human DNA insulin in 1978. Instead of collecting insulin from animals, new human insulin could be synthesized.

"Insulin is not a cure for diabetes; it is a treatment. It enables the diabetic to burn sufficient carbohydrates, so that proteins and fats may be added to the diet in sufficient quantities to provide energy for the economic burdens of life." It nearly proves that only by changing the person lifestyle into healthier. When the person's life changes, he starts leading in the peaceful environment and living a happy life. Always maintain a proper diet and make a habit of doing physical activity.

1.6 DIABETES AND ITS TYPES

Nowadays diabetes is a common group of diseases in which the body doesn't produce the insulin level in the body or high level of insulin secretion or exhibits a combination of both. When any of these situations occurs, the body is unable to get sugar from the blood. That leads to high blood sugar levels which may cause the dangerous conditions to health. In order to protect ourselves from this condition, one must always have a healthy diet and physical activities, and always keep the mind and body calm and relaxed. The form of sugar found in the blood which is glucose is one of the main energy sources. A lack of insulin causes sugar to build up in the blood. Lack of insulin may lead to many health problems.

The most common types of diabetes are:

- Type 1 Diabetes
- Type 2 Diabetes
- Gestational Diabetes

1.6.1 Type 1 Diabetes

In Type 1 diabetes, the human body does not make insulin. The person's immune system attacks and destroys the cells in the pancreas that make insulin. Type 1 diabetes was normally detected in children and young adults, even though it appears at any age. To stay alive People with type 1 diabetes, need to take insulin each day. Type 1 Diabetes patients were in need to take insulin because the insulin which the body needs to produce may stop doing the work, that insulin secretion work is done by external abundance.

Symptoms of type 1Diabetes

Sometimes the person may progress the ketoacidosis which has a major complication in diabetes, symptoms of this shows like,

- Rapid Breathing
- Dry Skin and Mouth
- Flushed Face
- Fruity Breath Odour
- Nausea
- Vomiting or Stomach Pain

Type 1 diabetes symptoms tend to begin hastily and considerably. It is most often seen in adolescents, young adults, and children. Type 1 diabetes people can notice a quick and sudden weight loss. Conversely, type 1 diabetes can develop at any age.

1.6.2 Type 2 Diabetes

According to the thrifty genotype hypothesis, the high predominance of type 2 diabetes and obesity is a sign of genetic variants that have undergone positive selection during historical periods of the erratic food supply. The recent

development in the number of validated type 2 diabetes and obesity resistance, coupled with the access to empirical data, enables us to look for evidence in support of the thrifty genotype.

According to the hypothesis, found no evidence for significant differences for the derived/inherited allele test. None of the study loci showed strong evidence for selection based on the iHS score. Found out on the high FST for rs 7901695 at TCF7L2, this is the largest type 2 diabetes effect size found to date.

Symptoms of Type 2 diabetes

In fact, the person might have type 2 diabetes for years and unable to find for so long. Type 2 diabetes results in showing the symptoms of:

- Increased Thirst
- Frequent Urination
- Increased Hunger
- Unintended Weight Loss
- Fatigue
- Blurred Vision
- Slow-Healing Sores
- Frequent Infections

1.7 PROBLEM STATEMENT

Controlling the blood glucose levels for diabetic patients has attracted significant research interest. Researchers used sensors and machine learning algorithms to help patients in maintaining their blood glucose level within the normal range. However, this tight glycemic control results in threefold increases in severe hypoglycemia occurrences. On the other hand, predicting hypoglycemia using machine learning techniques has received less research interest.

This paper investigates the hypoglycemia predication problem using machine learning techniques. In particular, two approaches are developed and compared, namely, time sensitive artificial neural network (TS-ANN) and tree based temporal classification (TBTC) techniques.

In the TS-ANN, the ANN predicts the future subcutaneous glucose measurements, using previous values of blood glucose and other parameters, and then uses those predicted values to decide on the occurrence or non-occurrence of hypoglycemia events within a specific horizon. In the TBTC approach, features that best discriminate hypoglycemia from non-hypoglycemia events are extracted from the patient's glucose signal, and then fed to a decision tree for predicting the occurrence of the hypoglycemia events.

CHAPTER 2

LITERATURE SURVEY

2.1 D. Tian, J. Zhou, Y. Wang, Y. Lu, H. Xia, and Z. Yi, "A dynamic and self-adaptive network selection method for multimode communications in heterogeneous vehicular telematics," IEEE Transactions on Intelligent Transportation Systems, vol. 16, no. 6, pp. 3033–3049, 2015.

With the increasing demands for vehicle-to-vehicle and vehicle-to-infrastructure communications in intelligent transportation systems, new generation of vehicular telematics inevitably depends on the cooperation of heterogeneous wireless networks. In heterogeneous vehicular telematics, the network selection is an important step to the realization of multimode communications that use multiple access technologies and multiple radios in a collaborative manner. This paper presents an innovative network selection solution for the fundamental technological requirement of multimode communications in heterogeneous vehicular telematics. To guarantee the QoS satisfaction of multiple mobile users and the efficient utilization and fair allocation of heterogeneous network resources in a global sense, a dynamic and self-adaptive method for network selection is proposed. It is biologically inspired by the cellular gene network, which enables terminals to dynamically select an appropriate access network according to the variety of QoS requirements and to the dynamic conditions of various available networks. The experimental results prove the effectiveness of the bioinspired scheme and confirm that the proposed network selection method provides better global performance when compared with the utility function method with greedy optimization.

2.2 M. Chen, Y. Ma, Y. Li, D. Wu, Y. Zhang, C. Youn, "Wearable 2.0: Enable Human-Cloud Integration in Next Generation Healthcare System," IEEE Communications, Vol. 55, No. 1, pp. 54–61, Jan. 2017.

With the rapid development of the Internet of Things, cloud computing, and big data, more comprehensive and powerful applications become available. Meanwhile, people pay more attention to higher QoE and QoS in a "terminal-cloud" integrated system. Specifically, both advanced terminal technologies (e.g., smart clothing) and advanced cloud technologies (e.g., big data analytics and cognitive computing in clouds) are expected to provide people with more reliable and intelligent services. Therefore, in this article we propose a Wearable 2.0 healthcare system to improve QoE and QoS of the next generation healthcare system. In the proposed system, washable smart clothing, which consists of sensors, electrodes, and wires, is the critical component to collect users' physiological data and receive the analysis results of users' health and emotional status provided by cloud-based machine intelligence.

2.3 M. Chen, Y. Ma, J. Song, C. Lai, B. Hu, "Smart Clothing: Connecting Human with Clouds and Big Data for Sustainable Health Monitoring," ACM/Springer Mobile Networks and Applications, Vol. 21, No. 5, pp. 825C845, 2016.

Traditional wearable devices have various shortcomings, such as uncomfortableness for long-term wearing, and insufficient accuracy, etc. Thus, health monitoring through traditional wearable devices is hard to be sustainable. In order to obtain healthcare big data by sustainable health monitoring, we design "Smart Clothing", facilitating unobtrusive collection of various physiological indicators of human body. To provide pervasive intelligence for smart clothing

system, mobile healthcare cloud platform is constructed by the use of mobile internet, cloud computing and big data analytics. This paper introduces design details, key technologies and practical implementation methods of smart clothing system. Typical applications powered by smart clothing and big data clouds are presented, such as medical emergency response, emotion care, disease diagnosis, and real-time tactile interaction. Especially, electrocardiograph signals collected by smart clothing are used for mood monitoring and emotion detection. Finally, we highlight some of the design challenges and open issues that still need to be addressed to make smart clothing ubiquitous for a wide range of applications.

2.4 M. Qiu and E. H.-M. Sha, "Cost minimization while satisfying hard/soft timing constraints for heterogeneous embedded systems," ACM Transactions on Design Automation of Electronic Systems (TODAES), vol. 14, no. 2, p. 25, 2009.

In high-level synthesis for real-time embedded systems using heterogeneous functional units (FUs), it is critical to select the best FU type for each task. However, some tasks may not have fixed execution times. This article models each varied execution time as a probabilistic random variable and solves heterogeneous assignment with probability (HAP) problem. The solution of the HAP problem assigns a proper FU type to each task such that the total cost is minimized while the timing constraint is satisfied with a guaranteed confidence probability. The solutions to the HAP problem are useful for both hard real-time and soft real-time systems. Optimal algorithms are proposed to find the optimal solutions for the HAP problem when the input is a tree or a simple path. Two other algorithms, one is optimal and the other is near-optimal heuristic, are proposed to solve the general problem. The experiments show that our algorithms can

effectively reduce the total cost while satisfying timing constraints with guaranteed confidence probabilities. For example, our algorithms achieve an average reduction of 33.0% on total cost with 0.90 confidence probability satisfying timing constraints compared with the previous work using worst-case scenario.

2.5 J. Wang, M. Qiu, and B. Guo, "Enabling real-time information service on telehealth system over cloud-based big data platform," Journal of Systems Architecture, vol. 72, pp. 69–79, 2017.

A telehealth system covers both clinical and nonclinical uses, which not only provides store-and-forward data services to be offline studied by relevant specialists, but also monitors the real-time physiological data through ubiquitous sensors to support remote telemedicine. However, the current telehealth systems do not consider the velocity and veracity of the big-data system in the medical context. Emergency events generate a large amount of the real-time data, which should be stored in the data center, and forwarded to remote hospitals. Furthermore, patients' information is scattered on the distributed data center, which cannot provide a high-efficient remote real-time service. In this paper, we propose a probability-based bandwidth model in a telehealth cloud system, which helps cloud broker to provide a high-performance allocation of computing nodes and links. This brokering mechanism considers the location protocol of Personal Health Record (PHR) in cloud and schedules the real-time signals with a low information transfer between different hosts. The broker uses several bandwidth evaluating methods to predict the near future usage of bandwidth in a telehealth context. The simulation results show that our model is effective at determining the

best performing service, and the inserted service validates the utility of our approach.

2.6 D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar, "Big data in health care: using analytics to identify and manage high-risk and high-cost patients," Health Affairs, vol. 33, no. 7, pp. 1123–1131, 2014.

The US health care system is rapidly adopting electronic health records, which will dramatically increase the quantity of clinical data that are available electronically. Simultaneously, rapid progress has been made in clinical analytics—techniques for analyzing large quantities of data and gleaning new insights from that analysis—which is part of what is known as big data. As a result, there are unprecedented opportunities to use big data to reduce the costs of health care in the United States. We present six use cases—that is, key examples—where some of the clearest opportunities exist to reduce costs through the use of big data: high-cost patients, readmissions, triage, decompensation (when a patient's condition worsens), adverse events, and treatment optimization for diseases affecting multiple organ systems. We discuss the types of insights that are likely to emerge from clinical analytics, the types of data needed to obtain such insights, and the infrastructure—analytics, algorithms, registries, assessment scores, monitoring devices, and so forth—that organizations will need to perform the necessary analyses and to implement changes improve care while reducing costs.

2.7 L. Qiu, K. Gai, and M. Qiu, "Optimal big data sharing approach for tele-health in cloud computing," in Smart Cloud (SmartCloud), IEEE International Conference on. IEEE, 2016, pp. 184–189.

The rapid development of tele-health systems has received driving engagements from various emerging techniques, such as big data and cloud

computing. Sharing data among multiple tele-health systems is an adaptive approach for improving service quality via the network-based technologies. However, current implementations of data sharing in cloud computing is still facing the restrictions caused by the networking capacities and virtual machine switches. In this paper, we focus on the problem of data sharing obstacles in cloud computing and propose an approach that uses dynamic programming to produce optimal solutions to data sharing mechanisms. The proposed approach is called Optimal Telehealth Data Sharing Model (OTDSM), which considers transmission probabilities, maximizing network capacities, and timing constraints. Our experimental results have proved the flexibility and adoptability of the proposed method.

2.8 Y. Zhang, M. Qiu, C.-W. Tsai, M. M. Hassan, and A. Alamri, "Healthcps: Healthcare cyber-physical system assisted by cloud and big data," IEEE Systems Journal, 2015.

The advances in information technology have witnessed great progress on healthcare technologies in various domains nowadays. However, these new technologies have also made healthcare data not only much bigger but also much more difficult to handle and process. Moreover, because the data are created from a variety of devices within a short time span, the characteristics of these data are that they are stored in different formats and created quickly, which can, to a large extent, be regarded as a big data problem. To provide a more convenient service and environment of healthcare, this paper proposes a cyber-physical system for patient-centric healthcare applications and services, called Health-CPS, built on cloud and big data analytics technologies. This system consists of a data collection layer with a unified standard, a data management layer for distributed storage and

parallel computing, and a data-oriented service layer. The results of this study show that the technologies of cloud and big data can be used to enhance the performance of the healthcare system so that humans can then enjoy various smart healthcare applications and services.

2.9 K. Lin, J. Luo, L. Hu, M. S. Hossain, and A. Ghoneim, "Localization based on social big data analysis in the vehicular networks," IEEE Transactions on Industrial Informatics, 2016.

Location-based services, especially for vehicular localization, are an indispensable component of most technologies and applications related to the vehicular networks. However, because of the randomness of the vehicle movement and the complexity of a driving environment, attempts to develop an effective localization solution face certain difficulty. In this paper, an overlapping and hierarchical social clustering model is first designed to classify the vehicles into different social clusters by exploring the social relationship between them. By using the results of the OHSC model, we propose a social-based localization algorithm (SBL) that use location prediction to assist in global localization in the vehicular networks. The experiment results validate the performance of the OHSC model and show that the presented SBL algorithm demonstrates superior localization performance compared with existing methods.

2.10 J. Wan, S. Tang, D. Li, S. Wang, C. Liu, H. Abbas and A. Vasilakos, "A Manufacturing Big Data Solution for Active Preventive Maintenance", IEEE Transactions on Industrial Informatics, DOI: 10.1109/TII. 2017.2670505, 2017.

Industry 4.0 has become more popular due to recent developments in cyber-physical systems, big data, cloud computing, and industrial wireless networks.

Intelligent manufacturing has produced a revolutionary change, and evolving applications, such as product lifecycle management, are becoming a reality. In this paper, we propose and implement a manufacturing big data solution for active preventive maintenance in manufacturing environments. First, we provide the system architecture that is used for active preventive maintenance. Then, we analyze the method used for collection of manufacturing big data according to the data characteristics. Subsequently, we perform data processing in the cloud, including the cloud layer architecture, the real-time active maintenance mechanism, and the offline prediction and analysis method. Finally, we analyze a prototype platform and implement experiments to compare the traditionally used method with the proposed active preventive maintenance method. The manufacturing big data method used for active preventive maintenance has the potential to accelerate implementation of Industry 4.0.

CHAPTER 3

SYSTEM ANALYSIS

3.1 EXISTING SYSTEM

Diabetes mellitus is a healthcare burden in India. Seventy-four percent of India's population lives in rural areas with limited access to healthcare resources. Telemedicine can play a big role in screening people with diabetes at grassroots level. In the telescreening model, single field 45-degree photographs are used for detecting diabetic retinopathy. The American Academy of Ophthalmology does not recommend single-field fundus photography as an adequate substitute for a comprehensive ophthalmic examination because it may lead to a higher rate of underdiagnosis. We conducted a telescreening project using single-field fundus photography to determine its accuracy compared to the traditional camp-based screenings. In this we compared the prevalence of diabetic retinopathy between an ophthalmologist-based and an ophthalmologist-led model on two different samples of people self-reporting with diabetes in rural South India. Between 2004 and 2005 in rural South India, 3522 people with diabetes mellitus underwent ophthalmologist-based diabetic retinopathy screening and 4456 people with diabetes underwent ophthalmologist-led (telescreening) diabetic retinopathy screening. The two population groups were randomly separated. In the ophthalmologist-based program, a trained retina specialist travels along with the camp team and screens patients at the camp site for diabetic retinopathy.

3.1.1 Disadvantages

- Time consuming
- Wait for doctor

- Travel

To achieve near-universal coverage, the screening method should be community-based, and the point of delivery should be within easy reach of the population. Such screening can be either ophthalmologist-based or ophthalmologist-led8 . In the ophthalmologist-based program, a trained retina specialist travels along with the camp team and screens patients at the camp site for diabetic retinopathy. In the ophthalmologist-led program (telescreening), fundus photographs are transmitted to the base hospital for further evaluation and grading. There has been speculation about whether telemedicine overestimates or underestimates diabetic retinopathy. The present study compares the prevalence of diabetic retinopathy in an ophthalmologist based model with an ophthalmologist-led model in people self-reporting with diabetes in rural South India.

Between January 2004 and December 2005, 39 free diabetic retinopathy screening camps were conducted in the rural areas of three districts of Tamil Nadu funded by Lions Club International. These camps were randomized into ophthalmologist-based and ophthalmologist-led groups. Of these, 21 camps were ophthalmologist-based and 18 were ophthalmologist-led (telescreening). A customized mobile van with in-built ophthalmic examination facility and satellite connectivity (courtesy of the Indian Space Research Organization) was used for telescreening; clinical examination was performed by an optometrist, and a social worker assisted him in villages.

Diabetes mellitus is a global healthcare burden. In 2011 there were 366 million people with diabetes globally, and this is expected to increase to 552 million by

20301 . Eighty percent of people with diabetes live in low- and middleincome countries and an estimated 183 million people (50%) with diabetes are undiagnosed2 . The Indian Council of Medical Research–India Diabetes (ICMR-INDIAB) national study reported that in India 62.4 million people have type 2 diabetes and 77 million people have pre-diabetes3 . These numbers are projected to increase to 101 million by 20301 . Diabetic retinopathy is a major cause of blindness among those of working age4 . There are approximately 93 million people with DR, 17 million with proliferative DR, 21 million with diabetic macular edema, and 28 million with visionthreatening diabetic retinopathy worldwide5 . The Chennai Urban Rural Epidemiology Study reported that nearly 25% of the Chennai population was unaware of a condition called diabetes.

To achieve near-universal coverage, the screening method should be community-based, and the point of delivery should be within easy reach of the population. Such screening can be either ophthalmologist-based or ophthalmologist-led.

In the ophthalmologist-based screening camps, the data sheet and routine evaluation were the same as that for the telescreening camps. Fundus evaluation to screen for diabetic retinopathy was done by binocular indirect ophthalmoscopy with magnifier (Keeler Ltd, Windsor, UK) by a retina specialist. Diabetic retinopathy was graded clinically using Klein's classification (Modified Early Treatment Diabetic Retinopathy Study scales)9 . Sight-threatening diabetic retinopathy was defined as those eyes that had severe non-proliferative diabetic retinopathy, proliferative diabetic retinopathy, severe diabetic macular edema, or a combination of these. These patients were referred to the base hospital for further management. The data sheet variables from both types of screening camps were

entered into Microsoft Access sheets. The data was analyzed by the Statistical Package for the Social Sciences

In the telescreening model, single-field 45-degree photographs were used for screening. The American Academy of Ophthalmology has recommended that singlefield fundus photography was not an adequate substitute in an urban population for a comprehensive ophthalmic examination because it may lead to a higher rate of underdiagnosis. However, level I evidence suggested that single-field images could serve as a screening tool for diabetic retinopathy to identify those patients who needed referral for further ophthalmic evaluation and treatment.

The results of our study support telescreening (ophthalmologist-led model) as a screening tool for people with diabetes mellitus living in rural areas for referral and further management to higher urban healthcare centers. Several other studies data are required to ascertain the role of telescreening and its accuracy at grassroots levels. Telescreening for diabetic retinopathy seems to be a good model,without undue fear of missing the diagnosis. Above all, it obviates the need for a physical presence of a retinologist in the field area, especially in countries like India.

We assume that the relative higher prevalence of diabetic retinopathy in the telescreening model is due to overdiagnosis. Possible contributing factors are the older age of this group of patients and increased duration of diabetes compared to those of the ophthalmologist-based model. The patients in the two groups are different, so we cannot pinpoint the cause of such a discrepancy. Mode of transmission in our study was satellite-based. Satellite transmission of data eliminates the need for having local internet connectivity and the need for

infrastructure at the village site. The images transmitted via satellite have high resolution and allow real-time conferencing between the patient and the ophthalmologist at the hub. The technical working group for standardization on telemedicine in India recommends satellite link as the best option for data transmission.

3.2PROPOSED SYSTEM

RISK PREDICTION OF NON-COMMUNCABLE DISEASE IN EARLY STAGE

Classification is one of the most important decision-making techniques in many real-world problems. In this work, the main objective is to classify the data as diabetic or non-diabetic and improve the classification accuracy. For many classification problems, the higher number of samples chosen but it doesn't leadto higher classification accuracy. In many cases, the performance of algorithm is high in the context of speed but the accuracy of data classification is low. The main objective of our model is to achieve high accuracy. Classification accuracy can be increase if we use much of the data set for training and few data sets for testing. This survey has analysed various classification techniques for classification of diabetic and non-diabetic data. Thus, it is observed that techniques like Support Vector Machine, Logistic Regression, and Artificial Neural Network are most suitable for implementing the Diabetes prediction system.

If the amount of insulin available is insufficient, then glucose will not have its usual effect so that glucose will not be absorbed by the body cells that require it. Diabetes mellitus being one of the major contributors to the mortality rate. Detection and diagnosis of diabetes at an early stage is the need of the day.

Diabetes disease diagnosis and interpretation of the diabetes data is an important classification problem. A classifier is required and to be designed that is cost efficient, convenient and accurate. Artificial intelligence and Soft Computing Techniques provide a great deal of human ideologies and are involved in human related fields of application. These systems find a place in the medical diagnosis.

Diabetic nephropathy (DN) is one of the most feared diabetic chronic microvascular complications and the major cause of end-stage renal disease (ESRD). The classical presentation of DN is characterized by hyperfiltration and albuminuria in the early phases which is then followed by a progressive renal function decline. The presentation of diabetic kidney disease (DKD) can vary especially in patients with T2DM where concomitant presence of other glomerular/tubular pathologies and severe peripheral vascular disease can become important confounders. All-cause mortality in individuals with DKD is approximately 30 times higher than that in diabetic patients without nephropathy and a great majority of patients with DKD will die from cardiovascular disease before they reach ESRD. The management of metabolic and hemodynamic perturbations for the prevention and for the delay of progression of DKD is very important. DKD is a global challenge and a significant social and economic burden; research should aim at developing new ideas to tackle this devastating condition.

Diabetes is characterized by chronic hyperglycemia and glucose intolerance due to impaired insulin action and/or secretion. Microand macro-diabetic chronic vascular complications affect the majority of patients with diabetes in both developed and developing countries. The microvascular complications include diabetic nephropathy, retinopathy, and neuropathy, and are responsible for significant morbidity in this patient group. The macrovascular complications

include accelerated coronary heart disease, ischemic stroke, and peripheral vascular disease, and represent the most common cause of mortality in diabetic patients.

Based on the American Diabetes Association classification system, the two major forms of diabetes are type 1 diabetes mellitus (T1DM), an autoimmune disorder characterized by the destruction of the insulin-producing β-cells in the islets of Langerhans which accounts for 5–10% of cases, and type 2 diabetes mellitus (T2DM), a more common form of DM which accounts for up to 90% of cases and is due to diminished insulin action. With the increasing incidence of obesity worldwide, a pronounced increase in T2DM has been noted, compared to T1DM, which now typically affects a younger and increasingly obese patient group. Recently subgroups of patients have been proposed based on six variables (glutamate decarboxylase antibodies, age at diagnosis, body mass index, glycated hemoglobin, and homoeostatic model assessment 2 estimates of β-cell function and insulin resistance). The resulting groups retain a diverse disease progression and risk of diabetic chronic complications.

Diabetic nephropathy (DN) is a morbid and deeply feared complication of diabetes. A striking 45% of T1DM and T2DM diabetics are affected by this microvascular complication. At present, diabetes is the single leading cause of end-stage renal disease (ESRD) in the Western world and the principal cause for patients requiring renal replacement therapy worldwide. However, due to the strong association between DN and cardiovascular disease, a large majority of patients with DN will die even before progression to ESRD, as a result of cardiovascular related events

According to the growing morbidity in recent years, in 2040, the world's diabetic patients will reach 642 million, which means that one of the ten adults in the future is suffering from diabetes. There is no doubt that this alarming figure needs great attention. With the rapid development of machine learning, machine learning has been applied to many aspects of medical health.

Persistent exposure to elevated blood glucose levels leads to damage and disruption of the renal cellular architecture and microvasculature in patients with diabetes. A number of complex pathways mediate these effects and are grouped into four main categories: metabolic, hemodynamic, intracellular, and growth factors/cytokines. As a result, unique ultrastructural changes occur in the kidney nephron, at the level of the glomerulus. The glomerular filtration barrier (GFB) is a complex structure, made of four key components: the mesangium, glomerular basement membrane (GBM), fenestrated glomerular endothelial cells, and podocytes. In DN, hallmark pathological changes occur in the GFB, including GBM thickening, mesangial sclerosis, endothelial dysfunction with glycocalyx damage, podocyte foot process effacement and detachment, and decreased podocyte number. Similar insults affect the renal tubular compartment resulting in progressive deposition of extracellular matrix and secondary tubular interstitial fibrosis.

So, for efficiently and effectively diagnosing the Diabetes Mellitus, a method is proposed using the ML Grid Search algorithm. In this method, a database called Pima Indian Diabetic Dataset is used. This system has two phases: the training phase and the test phase. In training phase, preprocessing, feature selection and instance evaluation is done. In test phase, preprocessing, instance evaluation and disease prediction is done. For feature selection, Modified Squirrel Search Algorithm is used and for classification, Support Vector Machine (SVM),

a machine learning method as the classifier for diagnosis of diabetes. The machine learning method focus on classifying diabetes disease from high dimensional medical dataset. The experimental results obtained show that support vector machine can be successfully used for diagnosing diabetes disease.

3.3 MACHINE LEARNING

Machine learning is the scientific field dealing with the ways in which machines learn from experience. For many scientists, the term “machine learning” is identical to the term “artificial intelligence”, given that the possibility of learning is the main characteristic of an entity called intelligent in the broadest sense of the word. The purpose of machine learning is the construction ofcomputer systems that can adapt and learn from their experience. A more detailedand formal definition of machine learning is given by Mitchel: A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E. With the rise of Machine Learning approaches, we have the ability to find a solution to this issue, we have developed a system usingdata mining which has the ability to predict whether the patient has diabetes or not. Furthermore, predicting the disease early leads to treating the patients beforeit becomes critical. Data mining has the ability to extract hidden knowledge froma huge amount of diabetes-related data. Because of that, it has a significant role in diabetes research, now more than ever. The aim of this research is to develop a system which can predict the diabetic risk level of a patient with a higher accuracy. This research has focused on developing a system based on three classification

methods namely, Support Vector Machine, Logistic regression and Artificial Neural Network algorithms.

3.3.1 Supervised Learning

In supervised learning, the system must "learn" inductively a function called target function, which is an expression of a model describing the data. The objective function is used to predict the value of a variable, called dependent variable or output variable, from a set of variables, called independent variables or input variables or characteristics or features. The set of possible input values ofthe function, i.e., its domain, are called instances. Each case is described by a set of characteristics (attributes or features). A subset of all cases, for which the output variable value is known, is called training data or examples. In order to infer the best target function, the learning system, given a training set, takes into consideration alternative functions, called hypothesis and denoted by h. In supervised learning, there are two kinds of learning tasks: classification and regression. Classification models try to predict distinct classes, such as e.g., blood groups, while regression models predict numerical values. Some of the most common techniques are Decision Trees (DT), Rule Learning, and Instance Based Learning (IBL), such as k-Nearest Neighbours (k-NN), Genetic Algorithms (GA), Artificial Neural Networks (ANN), and Support Vector Machines (SVM).

3.3.2 Unsupervised Learning

In unsupervised learning, the system tries to discover the hidden structure of data or associations between variables. In that case, training data consists of instances without any corresponding labels. Association Rule Mining appeared much later than machine learning and is subject to greater influence from the

research area of databases. Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.

3.3.3 Reinforcement Learning

The term Reinforcement Learning is a general term given to a family of techniques, in which the system attempts to learn through direct interaction with the environment so as to maximize some notion of cumulative reward. It is important to mention that the system has no prior knowledge about the behaviour of the environment and the only way to find out is through trial and failure (trial and error). Reinforcement learning is mainly applied to autonomous systems, due to its independence in relation to its environment.

3.4 DIABETES PREDICTION USING SVM

Support Vector Machine (SVM) is used for classification in predicting diabetes which is a supervised machine learning algorithm. The SVM algorithm is required to find the optimum hyperplane between the two classes to accurately classify all the data points. In between the two classes, the optimum hyperplane maximizes the margin which is used to predict diabetes. The most important data points for training the dataset are Support Vectors. The position of the dividing hyperplane would change when the data points are removed from the training dataset and the data classification is difficult then prediction become difficult. The hyperplane provides maximum number of points in SVM classifier.

SVM technique is used to predict diabetes in a person through some prediction features such as age, number of pregnancies, insulin levels, glucose levels, etc. Several predictor factors are contained in the dataset for diabetes prediction. The person with diabetes is determined by the outcome 1. The labelled data is fed into the training set and the SVM algorithm provides the outcome of new examples. This algorithm makes use of kernel trick to transform the data. An optimal boundary was found between possible outcomes which are based on transformation of data. The SVM uses kernel trick for transforming the given data. The data is classified according to the label provided in the dataset. The missing data are taken concerns. The missing data and duplicate data can create changes prediction result. The testing for the disease become difficult when the feature selection changes due to mistakes in pre-processing (Gill and Mittal 2016).

The Receiver Operating Characteristic (ROC) curve is evaluated to diagnosis performance of the SVM models. The true positive rate is plotted in ROC curve as function for different cut-off points of the false positive rate. The sensitivity is represented with each ROC plot and decision threshold corresponding to specificity pair.

SVM provide better performance with clear margin separation and accurate in high dimensional. It uses a subset of training points in support vector so it is memory resourceful. The accuracy of SVM classifier is 77.4%. Although, SVM has some drawbacks that it is not suitable for large data set and not perform well when dataset has more noise.

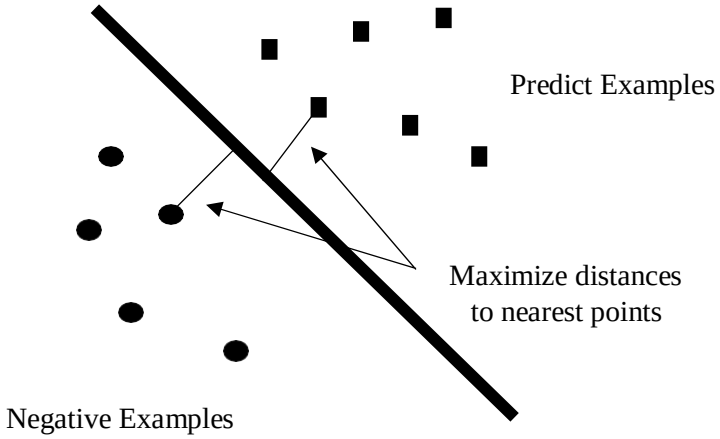

Fig 3.1: SVM Algorithm

Recently, SVM has attracted a high degree of interest in the machine learning research community. Several recent studies have reported that the SVM (support vector machines) generally are capable of delivering higher performance in terms of classification accuracy than the other data classification algorithms. SVM is a technique suitable for binary classification tasks, so we choose SVM to predict the diabetes. The reason is SVM is well known for its discriminative power for classification, especially in the cases where a large number of features are involved, and in our case where the dimension of the feature is 7.

The SVM model uses a feature vector space to represent the different input examples in which each example is mapped as a point in that space, in such a way as to maximize the distance between examples of the different class categories. The main disadvantage of SVM and artificial neural network (ANN) is their underlying models are of black box type. As a result, the classification rules cannot be seen or interpreted easily. Interpreting the classifiers' rules is important, especially in medical diagnosis when we need to convince the doctors with the obtained results.

An SVM classifier is used to extract rules. Two methods for rules extraction have been proposed, namely:

1) **SQRex-SVM**: Sequential Covering Approach for Rule Extraction, which extracts rules using a modified sequential covering algorithm, where feature selection is used to prune irrelevant features from the rules. Rules performance is measured using true positives and false positives along with the area under curve (AUC). These measures are used to determine and select the most accurate and comprehensive rules.

2) **Eclectic Rule Extraction:** the SVM classifier is trained using a labeled dataset until an acceptable accuracy is reached, then support vectors are constructed using the predicted class of SVM classifier as the target class. Finally, a C5 decision tree is used to extract rules from the newly constructed dataset.

The rules developed above have been used to diagnose diabetes patients using the by Oman diabetes dataset collected using a specially designed questionnaire. It contains 3014 patients of at least 20 years old. The Omani dataset includes attributes such as age, gender, Body Mass Index (BMI), and so on. The prevalence of diabetes in this dataset is 9%, thus the dataset is skewed toward the non-diabetic class. To overcome such a drawback, Barakat et al [28] used subsampling and k-means clustering to choose representative instances of the non-diabetic class.

The SVM model is constructed, trained, and then tested to classify the independent test dataset. Accuracy, true positives, and false positive rates were then calculated. Later, rules were extracted using the two proposed methods presented above.

3.5 ML BASED BLOOD GLUCOSE PREDICTION

A Machine Learning (ML) technique was used to predict next morning (fasting) blood glucose (FBG) levels. The authors ran the study on four insulin treated diabetic patients for a period of three months; they provided these patients with three portable commercial devices namely, a blood glucose monitor, a metabolic rate monitor, and a laptop as food intake monitor. The portable blood glucose monitor was used to measure the fasting blood glucose (FBG) level once a day. The patients were attached to the metabolic rate monitors to measure the rate at which the body burns calories. Calories calculating software was designed and implemented to calculate the calories of the food intake. It provided the patient with different meals displayed on the laptop screen, to choose the appropriate one for breakfast, lunch, and dinner. After collecting the data, a machine-learning technique was used to estimate blood glucose levels. However, the estimated FBG level turned out to be inaccurate after the eighth week of data collection; the error rates for the four subjects were 8.9%, 26.1%, 13.5%, and 22.2%. This was due to the small number of blood glucose measurements taken per day (only one glucose reading per day is used).

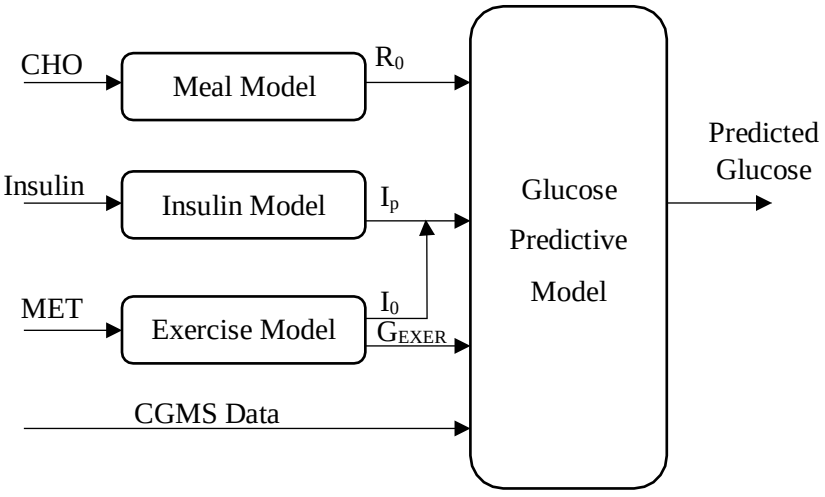

Fig 3.2: Glucose prediction technique

In comparison, used both compartmental (mathematical) models along with a Support Vector Machines for Regression (SVR) to predict the future glucose levels. They used, as shown in Fig 3.2, three compartmental models as already found in the literature: Firstly, a meal model that simulates the consumption and absorption of carbohydrates by the human body; secondly, the insulin model that simulates the absorption and the pharmacokinetics/pharmacodynamics of subcutaneously administered insulin; and finally, the exercise model that simulates the impact of exercising on glucose–insulin interaction.

In order to collect data, seven diabetic patients were monitored for 10 days, using several monitoring devices/techniques. Patients wore the Guardian Real-Time CGM system, which was used to measure the subcutaneous glucose concentration every five minutes. For monitoring physical activities, a Sense Wear body-monitoring device was used. Finally, a special paper-based diary was used to track the food intake, dosage, and time of insulin injections.

CHAPTER 4

SYSTEM DESIGN

4.1 SYSTEM ARCHITECTURE

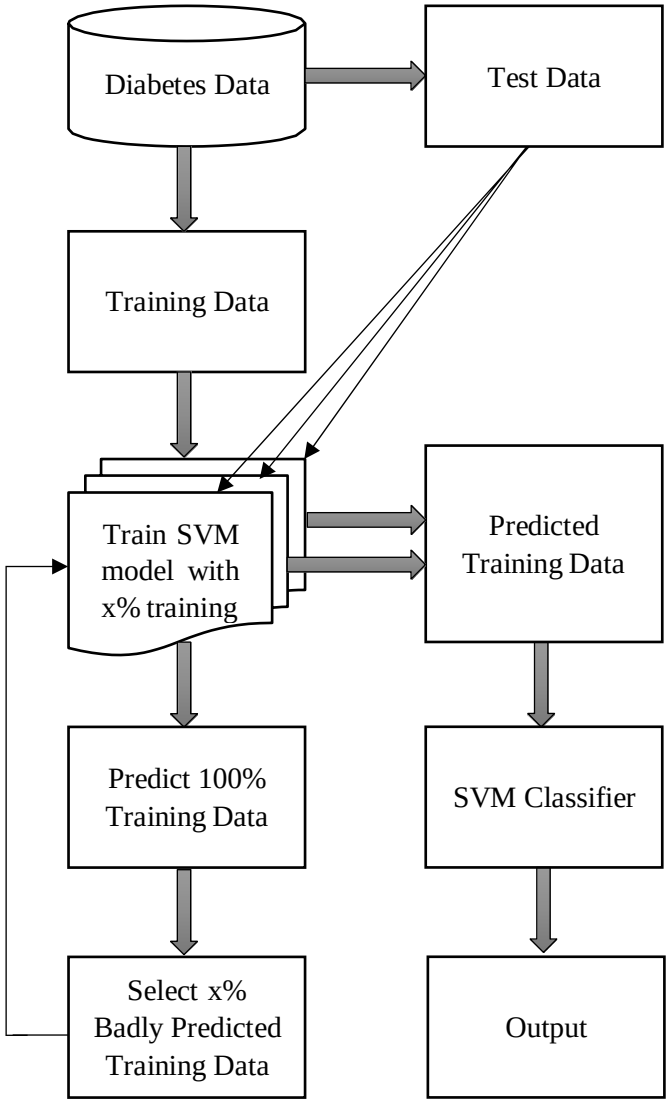

Fig 4.1: Architecture of Proposed System

4.2 DATA FLOW DIAGRAM

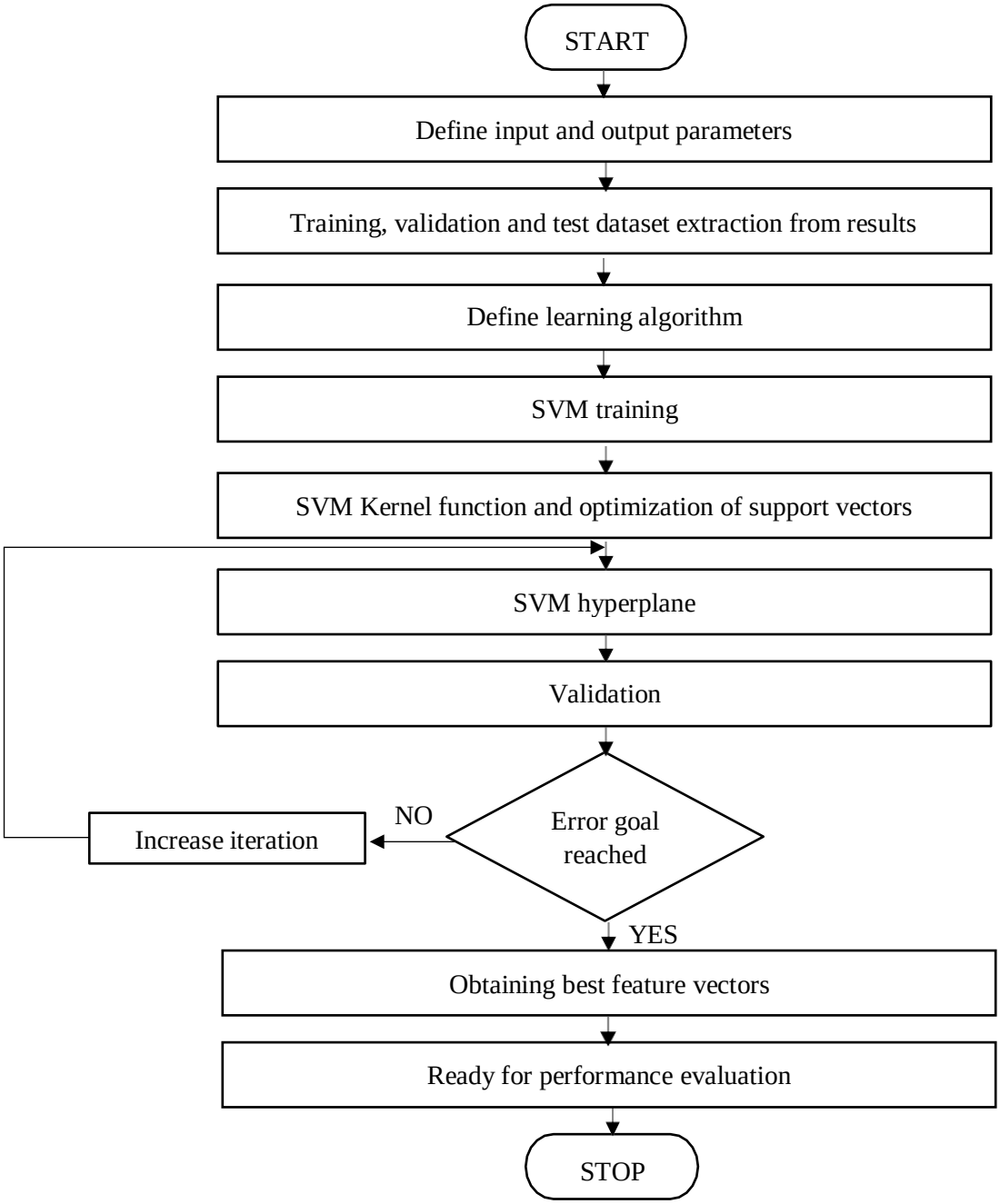

Fig 4.2: Data Flow Diagram

4.3 DISEASE CLASSIFICATION USING SVM

4.3.1. Experimental Setup

The SVM models for classification have been developed for the classification of diabetes dataset. The experiments are conducted on Matlab R2010a. The datasets are stored in MS Excel documents and read directly from Matlab. The diagnostic performance of the developed models is evaluated using Receiver Operating Characteristic (ROC) curve. In ROC curve the true positive rate (sensitivity) is plotted in function of the false positive rate for different cut-off points. Each point on the ROC plot represents a sensitivity/specificity pair corresponding to a particular decision threshold.

4.3.2. Diabetes Disease Dataset

The Pima Indian diabetes dataset, donated by Vincent Sigillito, is a collection of medical diagnostic reports from 768 records of female patients at least 21 years old of Pima Indian heritage, a population living near Phoenix, Arizona, USA. The binary target variable takes the values “0” or “1” while “1” means a positive test for diabetes, “0” means a negative test. There are 268 cases in class “1” and 500 cases in class “0”. The significance of the automatically selected set of variables was further manually evaluated by fine tuning parameters. The variables included in the final selection were those with the best discriminative performance.

There are eight numeric variables: (1) Number of times pregnant, (2) Plasma glucose concentration a 2h in an oral glucose tolerance test (3) Diastolic blood pressure (mm Hg) (4) Triceps skin fold thickness (mm) (5) 2-hour serum

insulin (mu U/ml) (6) Body mass index (7) Diabetes pedigree function (8) Age (years). Although the dataset is labeled as there are no missing values, there were some liberally added zeros as missing values. Five patients had a glucose of 0, 28 had a diastolic blood pressure of 0, 11 more had a body mass index of 0, 192 others had skin fold thickness readings of 0, and 140 others had serum insulin levels of 0. After the deletion there were 460 cases with no missing values.

4.3.3. Dataset Evaluation

To evaluate the robustness of the SVM models, a 10-fold cross-validation was performed in the training data set. The training data set is first partitioned into 10 equal-sized subsets. Each subset was used as a test data set for a model trained on all cases and an equal number of non-cases randomly selected from the remaining nine datasets. This cross-validation process was repeated 10 times, and each subset serve once as the test data set. Test data sets assess the performance of the models.

4.4 TRAINING PHASE

The Training Phase is used for train the data from dataset for carrying out the prediction which is needed. It basically trains the algorithm to create the correct output. The training set should be labelled correctly, so that the predictionwould be more accurate. 80% of data from dataset is used for training phase. Thisphase includes 3 stages namely pre-processing, feature selection and instance evaluation.

4.4.1 Pre-processing

Pre-processing is used for the transformation of the data before loading into the algorithm. The raw data is not suitable for processing. So, it is pre-processed

into understandable dataset. An integral step in Machine Learning is Data pre-processing which provide quality of data and the useful data can be analyzed from it directly disturbs the ability of the model to learn. Therefore, it is extremely important to pre-process the data before providing it into our model. The mathematically feasible format through Data Pre-processing (DP) involves the presence of data is an application of Machine Learning algorithm. Data reduction, data projection and missing-data treatment are involved in pre-processing.

a) Data Reduction

The size of the datasets can be reduced by Data Reduction technique by means of Feature Selection (FS) or Case Selection (CS). Dimensionality reduction refers the techniques that reduce number of input variables in a dataset.It is the corrected, ordered, and simplified form from the transformation of numerical or alphabetical digital information derived empirically or experimentally. The basic concept is the reduction of large amount of data to the meaningful parts. It is also called as dimensionality reduction.

b) Data Projection

Data projection aims to change the presence of the information from database, e.g., scaling, which scales all features and provide into a predefined same range. This process includes sorting the data in instances by age, removing duplicate data, missing data imputation and normalization. Missing data imputation means replacing the missing data values with comparing relevant features. The normalization is done for values with different ranges to change them into a common scale without distorting the original range. The major stage in ML methods is scaling. Scaling is used in measurements conducted for particular and repeatable conditions. In ML, scaling transforms feature values are based on to a defined rule so that all scaled features have the same degree of effect and turn into

immune. Normally, the intervals of [0,1] and [-1,1] are used in scaling to predict the target.

c) Removing Duplicate Data

Duplicate data are distinct as data in which selected feature share the same values. Duplicate inputs results in some distribution in output (Khan et al. 2012). The algorithm k-means clustering is used for removing the duplicate data. After predicting the duplicate data cluster application is used which is able to proceed and examine flows pairs within the cluster and determine whether the data is duplicated or not by check their feature values. Duplicate data is reduced from the cluster when similar data is determined (Wu et al. 2018). The duplicate data can be categories into three as follow.

Fully Duplicated Data: In this data having two identical rows representing the same values for the particular features.

Erroneous Duplicate Data: It is the duplicate data records, which seems to be different. It is due to the data entry operators. This type of duplicate data is challenging to predict and remove from data set. Sorting techniques are used to predict this type of duplicate data.

Partially Duplicate Data: This type has partial duplication and it is different. This type of duplicate data can be easily predicted and removed from dataset.

d) Missing-data treatments

Missing-data treatments (MDTs) include removing missing values and changing them with the estimates. The logarithmic transformation is applied often for linear regression to keep the normality assumption for the correct application

of linear regression. To ensure the normality of the residual missing data treatment is an important step for regression models. In ML studies, logarithm does not frequently appear. In deletion methods, Listwise Deletion (LD) is broadly used as an approach for treating with missing values which result in removal of large portions of datasets and introducing biasness (Nikfalazar et al. 2020). Another solution of MDT, more extensive, complicated statistical, computational analysis and natural prediction error requires Missing data imputation as the solution for biasness. The missing value can be imputed using Mean Imputation (MI) by calculating the mean of observed values and preserves the information of data. The simplest imputation method may cause to reduce the variance of variables(Jing et al. 2016).

e) Normalization

Normalization is used in pre-processing of data from database. It is used to arrange the data in same range of value in which it is used for feature selection and testing of diabetes mellitus. The normalization of trained data is used for testing the instance for prediction.

4.5 TESTING PHASE

To evaluate the performance of the model a set of observations is performed in testing phase by using some performance metric. The observations taken from the training set are involved in the testing set. In testing phase, the parameter involved are also in training phase. Data pre-processing such as removal of duplicate data, missing data imputation and normalization are performed which is same as that of training phase. The testing of diabetes is measured by using accuracy, precision and recall. 20% of data from dataset is used in testing phase.

This is the final stage in testing phase. The entropy based SVM Classifier algorithm and Adam Optimizer is used for classification in this model. The classification is based on the binary value 0 and 1. If the binary value is 0 then diabetes is negative in and the binary value is 1 then diabetes is positive.

4.5.1 Feature Selection

Feature selection (FS) is the method for choosing only the most related feature for processing. The subset of features is selected that have significant and similar impacts to the target value. FS is suggested to improve the accuracy of classifier in supervised learning algorithm. FS techniques is categorized as wrappers and filters. The wrappers convolve with predictors, using cross-validation to predict the advantages of adding or removing a feature from the feature subset.

The number of dimensions by selecting most informative features based on some statistical score is reduced using Feature selection. The performance of classification is estimated on the reduced diabetes dataset. The classification is less accurate if the diabetes dataset contains irrelevant and redundant features. Thebest features selection for prediction of diabetes is done by Modified Squirrel Search Feature Selection algorithm. The 8 features selected in this approach are as follows:

- Number of times the person had been pregnant
- Glucose level in blood
- Diastolic blood pressure
- BMI

• Thickness of skinfold (mm)

• Insulin level in a period of 2 hours

• Pedigree Function (hereditary factor)

• Age of the person

These 8 attributes should be in numeric format, which is then loaded into the algorithm and are trained to make decisions whether one reads diabetes positive or negative.

4.5.2 Modified Squirrel Search Algorithm

Squirrel search algorithm (SSA) resembles the southern flying squirrel with its dynamic foraging behavior. It is an active method used by small mammals to travel long distance. They use a special method of locomotion and flying squirrel do not fly which is dynamically cheap and allowing small mammals to cover large distances quickly and efficiently. The food resources are optimally use by squirrels which shows the dynamic foraging behavior. Likewise, it optimally selects the feature set.

The rate of convergence of SSA depends on exploration and exploitation abilities. The SSA dynamically adjusts the iteration process which introduced to provide an adaptive strategy of predator. The premature convergence is discouraged using this strategy and the intensive improves the search ability of the algorithm, so that a balance can be managed between the exploration and exploitation capabilities. For each iteration, the better quality of result can be achieved using dimensional search enhancement strategy

In this feature selection method initial population is randomly created with number of flying squirrels at each iteration and other parameter are also initialized. Then evaluation and selection are performed. Evaluation is done according to a fitness function and sorting the fitness value of squirrel in increasing order. The minimum fitness values provide the best feature which is used for feature selection. Then update squirrel location which based on the occurrence of probability of the predictor. The fitness valued of flying squirrel at each iteration is analyzed and location is updated until it reached to the maximum number of iterations. Then best feature is extracted by using squirrel search algorithm.

The modified squirrel search algorithm generates a possible random initial clarification and applies diverse strategy operation for updating squirrel algorithm to create a new population throughout the exploration for global solution. This algorithm provides a better feature selection in option in which it predicts the suitable feature for diabetic prediction.

Algorithm 1: Modified Squirrel Search

```
Initialize attributes
For each attribute
    {
      do until maximum iteration
        {
          Calculate data value
          If
            {
              Fitness is better
              Sort data according to fitness
```

Else

Check for better fitness

}

}

}

minimum fitness is selected

The squirrel search algorithm is modified to determine the features from dataset. The fitness is obtained for every attribute in the dataset and then sorted the data based on the fitness. The data with minimum fitness is selected for testing the diabetes. The squirrel location is updated in which each attribute is processed and fitness is calculated.

4.6 SUPPORT VECTOR MACHINE

SVM is a model-free method that provides efficient solutions to classification problems without any assumption regarding the distribution and interdependency of the data. In epidemiologic studies and population health surveys, the SVM technique has the potential to perform better than traditional statistical methods like logistic regression, especially in situations that include multivariate risk factors with small effects (e.g., genome-wide association data and gene expression profiles), limited sample size, and a limited knowledge of underlying biological relationships among risk factors.

Table 4.1: Performance of SVM model

Model	Dataset	Sensitivity	Specificity	PPV	NPV	AUC
Classification Scheme I	Test	0.7715	0.7503	0.4926	0.9127	0.8347

	Training	0.7938	0.7169	0.4550	0.9211	0.8383
	10-fold cross-validation	0.7765	0.7027	0.4388	0.9310	0.8242
Classification Scheme II	Test	0.7359	0.6254	0.5061	0.8195	0.7318
	Training	0.7092	0.6590	0.6729	0.8087	0.7393
	10-fold cross-validation	0.7059	0.6589	0.5293	0.8054	0.7357

This is particularly true in the case of common complex diseases where many risk factors, including gene-gene interactions and gene-environment interactions, have to be considered to reach sufficient discriminative power in prediction models. Our work provides a promising proof of principle by demonstrating the predictive power of the SVM with just a small set of variables. This approach can be extended to include large data sets, including many other variables, such as genetic biomarkers, as data become available.

Support vector machine modelling is a promising classification approach for detecting a complex disease like diabetes using common, simple variables. Validation indicated that the discriminative powers of our two SVM models are comparable to those of commonly used multivariable logistic regression methods. Our Diabetes Classifier tool, a web-based tool developed for demonstration purposes only, illustrates a potential use of the SVM technique: the identification of people with undetected common diseases such as diabetes and prediabetes.

CHAPTER 5

SYSTEM DESCRIPTION

5.1 Hardware Requirements

System	:	Pentium IV 3.5 GHz or Latest Version.
Hard Disk	:	40 GB.
Monitor	:	14’ Color Monitor.
Mouse	:	Optical Mouse.
Ram	:	1 GB.

5.2 Software Requirements

Operating system	:	Windows 10
Coding Language	:	Python Programming
Tools used	:	Jupyter Notebook

5.2.1 Operating System: Windows 10

Microsoft Windows is a group of several graphical operating system families, all of which are developed, marketed, and sold by Microsoft. Each family caters to a certain sector of the computing industry. Active Windows families include Windows NT and Windows Embedded; these may encompass subfamilies, e.g., Windows Embedded Compact (Windows CE) or Windows Server. Defunct Windows families include Windows 9x, Windows Mobile and Windows Phone.

Microsoft introduced an operating environment named Windows on November 20, 1985, as a graphical operating system shell for MS-DOS inresponse to the growing interest in graphical user interfaces (GUIs). Microsoft Windows came to dominate the world's personal computer (PC) market with over

90% market share, overtaking Mac OS, which had been introduced in 1984. Apple came to see Windows as an unfair encroachment on their innovation in GUI development as implemented on products such as the Lisa and Macintosh (eventually settled in court in Microsoft's favor in 1993). On PCs, Windows is still the most popular operating system. However, in 2014, Microsoft admitted losing the majority of the overall operating system market to Android, because of the massive growth in sales of Android smartphones. In 2014, the number of Windows devices sold was less than 25% that of Android devices sold. This comparison however may not be fully relevant, as the two operating systems traditionally target different platforms. Still, numbers for server use of Windows (that are comparable to competitors) show one third market share, similar to for end user use.

Fig 5.1: Windows OS

Microsoft, the developer of Windows, has registered several trademarks each of which denote a family of Windows operating systems that target a specific sector of the computing industry. As of 2014, the following Windows families are being actively developed.

5.2.2 Jupyter Notebook

Project Jupyter is a project and community whose goal is to "develop open-source software, open-standards, and services for interactive computing across

dozens of programming languages". It was spun off from IPython in 2014 by Fernando Pérez. Project Jupyter's name is a reference to the three core programming languages supported by Jupyter, which are Julia, Python and R, and also a homage to Galileo's notebooks recording the discovery of the moons of Jupiter. Project Jupyter has developed and supported the interactive computing products Jupyter Notebook, JupyterHub, and JupyterLab.

Project Jupyter's operating philosophy is to support interactive data science and scientific computing across all programming languages via the development of open-source software. According to the Project Jupyter website, "Jupyter will always be 100% open-source software, free for all to use and released under the liberal terms of the modified BSD license".

Jupyter Notebook (formerly IPython Notebooks) is a web-based interactive computational environment for creating Jupyter notebook documents. The "notebook" term can colloquially make reference to many different entities,mainly the Jupyter web application, Jupyter Python web server, or Jupyter document format depending on context. A Jupyter Notebook document is a JSONdocument, following a versioned schema, containing an ordered list of input/output cells which can contain code, text (using Markdown), mathematics, plots and rich media, usually ending with the ".ipynb" extension.

Fig 5.2: Jupyter Notebook

A Jupyter Notebook can be converted to a number of open standard output formats (HTML, presentation slides, LaTeX, PDF, ReStructuredText, Markdown, Python) through "Download As" in the web interface, via the nbconvert library or "jupyter nbconvert" command line interface in a shell. To simplify visualisation of Jupyter notebook documents on the web, the nbconvert library is provided as a service through NbViewer which can take a URL to any publicly available notebook document, convert it to HTML on the fly and display it to the user.

Jupyter Notebook provides a browser-based REPL built upon a number of popular open-source libraries:

- IPython
- ØMQ
- Tornado (web server)
- jQuery
- Bootstrap (front-end framework)
- MathJax

Jupyter Notebook can connect to many kernels to allow programming in different languages. By default, Jupyter Notebook ships with the IPython kernel.

As of the 2.3 release (October 2014), there are currently 49 Jupyter-compatible kernels for many programming languages, including Python, R, Julia and Haskell.

The Notebook interface was added to IPython in the 0.12 release (December 2011), renamed to Jupyter notebook in 2015 (IPython 4.0 – Jupyter 1.0). Jupyter Notebook is similar to the notebook interface of other programs such as Maple, Mathematica, and SageMath, a computational interface style that originated with Mathematica in the 1980s. According to The Atlantic, Jupyter interest overtook the popularity of the Mathematica notebook interface in early 2018.

The Jupyter Notebook has become a popular user interface for cloud computing, and major cloud providers have adopted the Jupyter Notebook or derivative tools as a frontend interface for cloud users. Examples include Amazon's SageMaker Notebooks, Google's Colaboratory and Microsoft's Azure Notebook.

Colaboratory (also known as Colab) is a free Jupyter notebook environment that runs in the cloud and stores its notebooks on Google Drive. Colab was originally an internal Google project; an attempt was made to open source all the code and work more directly upstream, leading to the development of the "Open in Colab" Google Chrome extension, but this eventually ended, and Colab development continued internally. As of October 2019, the Colaboratory UI only allows for the creation of notebooks with Python 2 and Python 3 kernels; however, an existing notebook whose kernelspec is IR or Swift will also work, since both R and Swift are installed in the container. Julia language can also work on Colab (with e.g., Python and GPUs; Google's tensor processing units also work with Julia on Colab).

5.2.3 Python Programming:

Python is an interpreted, high-level and general-purpose programming language. Created by Guido van Rossum and first released in 1991, Python's design philosophy emphasizes code readability with its notable use of significant whitespace. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects. Python is dynamically typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly, procedural), object-oriented, and functional programming. Python is often described as a "batteries included" language due to its comprehensive standard library. Python interpreters are available for many operating systems. A global community of programmers develops and maintains CPython, a free and open-source reference implementation. A non-profit organization, the Python Software Foundation, manages and directs resources for Python and CPython development.

Python is a multi-paradigm programming language. Object-oriented programming and structured programming are fully supported, and many of its features support functional programming and aspect-oriented programming (including by metaprogramming and metaobjects (magic methods)). Many other paradigms are supported via extensions, including design by contract and logic programming.

Fig 5.3: Python Programming

Python uses dynamic typing and a combination of reference counting and a cycle-detecting garbage collector for memory management. It also features dynamic name resolution (late binding), which binds method and variable names during program execution.

Python's design offers some support for functional programming in the Lisp tradition. It has filter, map, and reduce functions; list comprehensions, dictionaries, sets, and generator expressions. The standard library has two modules(itertools and functools) that implement functional tools borrowed from Haskell and Standard ML.

The language's core philosophy is summarized in the document The Zen of Python (PEP 20), which includes aphorisms such as:

- Beautiful is better than ugly.
- Explicit is better than implicit.
- Simple is better than complex.
- Complex is better than complicated.
- Readability counts.

Rather than having all of its functionality built into its core, Python was designed to be highly extensible. This compact modularity has made it particularly popular as a means of adding programmable interfaces to existing applications. Van Rossum's vision of a small core language with a large standard library and easily extensible interpreter stemmed from his frustrations with ABC, which espoused the opposite approach.

Python strives for a simpler, less-cluttered syntax and grammar while giving developers a choice in their coding methodology. In contrast to Perl's "there is more than one way to do it" motto, Python embraces a "there should be one—and

preferably only one—obvious way to do it" design philosophy. Alex Martelli, a Fellow at the Python Software Foundation and Python book author, writes that "To describe something as 'clever' is not considered a compliment in the Python culture."

Python's developers strive to avoid premature optimization, and reject patches to non-critical parts of the CPython reference implementation that would offer marginal increases in speed at the cost of clarity.[60] When speed is important, a Python programmer can move time-critical functions to extension modules written in languages such as C, or use PyPy, a just-in-time compiler. Cython is also available, which translates a Python script into C and makes direct C-level API calls into the Python interpreter.

An important goal of Python's developers is keeping it fun to use. This is reflected in the language's name—a tribute to the British comedy group Monty Python—and in occasionally playful approaches to tutorials and reference materials, such as examples that refer to spam and eggs (from a famous Monty Python sketch) instead of the standard foo and bar.

A common neologism in the Python community is pythonic, which can have a wide range of meanings related to program style. To say that code is pythonic is to say that it uses Python idioms well, that it is natural or shows fluency in the language, that it conforms with Python's minimalist philosophy and emphasis on readability. In contrast, code that is difficult to understand or reads like a rough transcription from another programming language is called unpythonic.

CHAPTER 6

SYSTEM IMPLEMENTATION

Implementation includes all those activities that take place to convert from the old system to the new. The old system consists of manual operations, which is operated in a very different manner from the proposed new system. A proper implementation is essential to provide a reliable system to meet the requirements of the organizations. An improper installation may affect the success of the computerizedsystem.

6.1 IMPLEMENTATION METHODS

There are several methods for handling the implementation and the consequent conversion from the old to the new computerized system.The most secure method for conversion from the old system to the new system is to run the old and new system in parallel. In this approach, a person may operate in the manual older processing system as well as start operating the new computerized system. This method offers high security, because even if there is a flaw in the computerized system, we can depend upon the manual system. However, the cost for maintaining two systems in parallel is very high. This outweighs its benefits.

Another commonly method is a direct cut over from the existing manual system to the computerized system. The change may be within a week or within a day. There are no parallel activities. However, there is no remedy in case of a problem. This strategy requires careful planning.A working version of the system can also be implemented in one part of the organization and the personnel will be piloting the system and changes can be made as and when required. But this method is less preferable due to the loss of entirety of the system.

6.2 IMPLEMENTATION PLAN

The implementation plan includes a description of all the activities that must occur to implement the new system and to put it into operation. It identifies the personnel responsible for the activities and prepares a time chart for implementing the system.

The implementation plan consists of the following steps.

- List all files required for implementation.
- Identify all data required to build new files during the implementation.
- List all new documents and procedures that go into the new system.

The implementation plan should anticipate possible problems and must be able to deal with them. The usual problems may be missing documents; mixed data formats between current and files, errors in data translation, missing data etc.

CHAPTER 7

SYSTEM TESTING

System testing is a critical aspect of Software Quality Assurance and represents the ultimate review of specification, design and coding. Testing is a process of executing a program with the intent of finding an error. A good test is one that has a probability of finding an as yet undiscovered error. The purpose of testing is to identify and correct bugs in the developed system. Nothing is complete without testing. Testing is the vital to the success of the system.

In the code testing the logic of the developed system is tested. For this, every module of the program is executed to find an error. To perform specification test, the examination of the specifications stating what the program should do and how it should perform under various conditions. Unit testing focuses first on the modules in the proposed system to locate errors. This enables to detect errors in the coding and logic that are contained within that module alone. Those resulting from the interaction between modules are initially avoided. In unit testing step each module has to be checked separately.

System testing does not test the software as a whole, but rather than integration of each module in the system. The primary concern is the compatibility of individual modules. One has to find areas where modules have been designed with different specifications of data lengths, type and data element name. Testing and validation are the most important steps after the implementation of the developed system. The system testing is performed to ensure that there are no errors in the implemented system. The software must be executed several times in order to find out the errors in the different modules of the system.

Validation refers to the process of using the new software for the developed system in a live environment i.e., new software inside the organization, in order to find out the errors. The validation phase reveals the failures and the bugs in the developed system. It will be come to know about the practical difficulties the system faces when operated in the true environment. By testing the code of the implemented software, the logic of the program can be examined. A specification test is conducted to check whether the specifications stating the program are performing under various conditions. Apart from these tests, there are some special tests conducted which are given below:

Peak Load Tests: This determines whether the new system will handle the volume of activities when the system is at the peak of its processing demand. The test has revealed that the new software for the agency is capable of handling the demands at the peak time.

Storage Testing: This determines the capacity of the new system to store transaction data on a disk or on other files. The proposed software has the required storage space available, because of the use of a number of hard disks.

Performance Time Testing: This test determines the length of the time used by the system to process transaction data.

In this phase the software developed Testing is exercising the software to uncover errors and ensure the system meets defined requirements. Testing may be done at 4 levels

- Unit Level
- Module Level
- Integration & System
- Regression

7.1 UNIT TESTING

A Unit corresponds to a screen /form in the package. Unit testing focuses on verification of the corresponding class or Screen. This testing includes testing of control paths, interfaces, local data structures, logical decisions, boundary conditions, and error handling. Unit testing may use Test Drivers, which are control programs to co-ordinate test case inputs and outputs, and Test stubs, which replace low-level modules. A stub is a dummy subprogram.

7.2 MODULE LEVEL TESTING

Module Testing is done using the test cases prepared earlier. Module is defined during the time of design.

7.3 INTEGRATION TESTING

Integration testing is used to verify the combining of the software modules. Integration testing addresses the issues associated with the dual problems of verification and program construction. System testing is used to verify, whether the developed system meets the requirements.

7.4 REGRESSION TESTING

Each modification in software impacts unmodified areas, which results serious injuries to that software. So, the process of re-testing for rectification of errors due to modification is known as regression testing.

Installation and Delivery: Installation and Delivery is the process of delivering the developed and tested software to the customer. Refer the support procedures.

Acceptance and Project Closure: Acceptance is the part of the project by which the customer accepts the product. This will be done as per the Project Closure, once the customer accepts the product, closure of the project is started. This includes metrics collection, PCD, etc.

CHAPTER 8

RESULTS AND DISCUSSION

3]:

	num_preg	glucose_con	Diastolic_bp	Thckness	Insulin	BMI	Diab_pred	Age	Target
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
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Fig 8.1 : Types of diabetes prediction

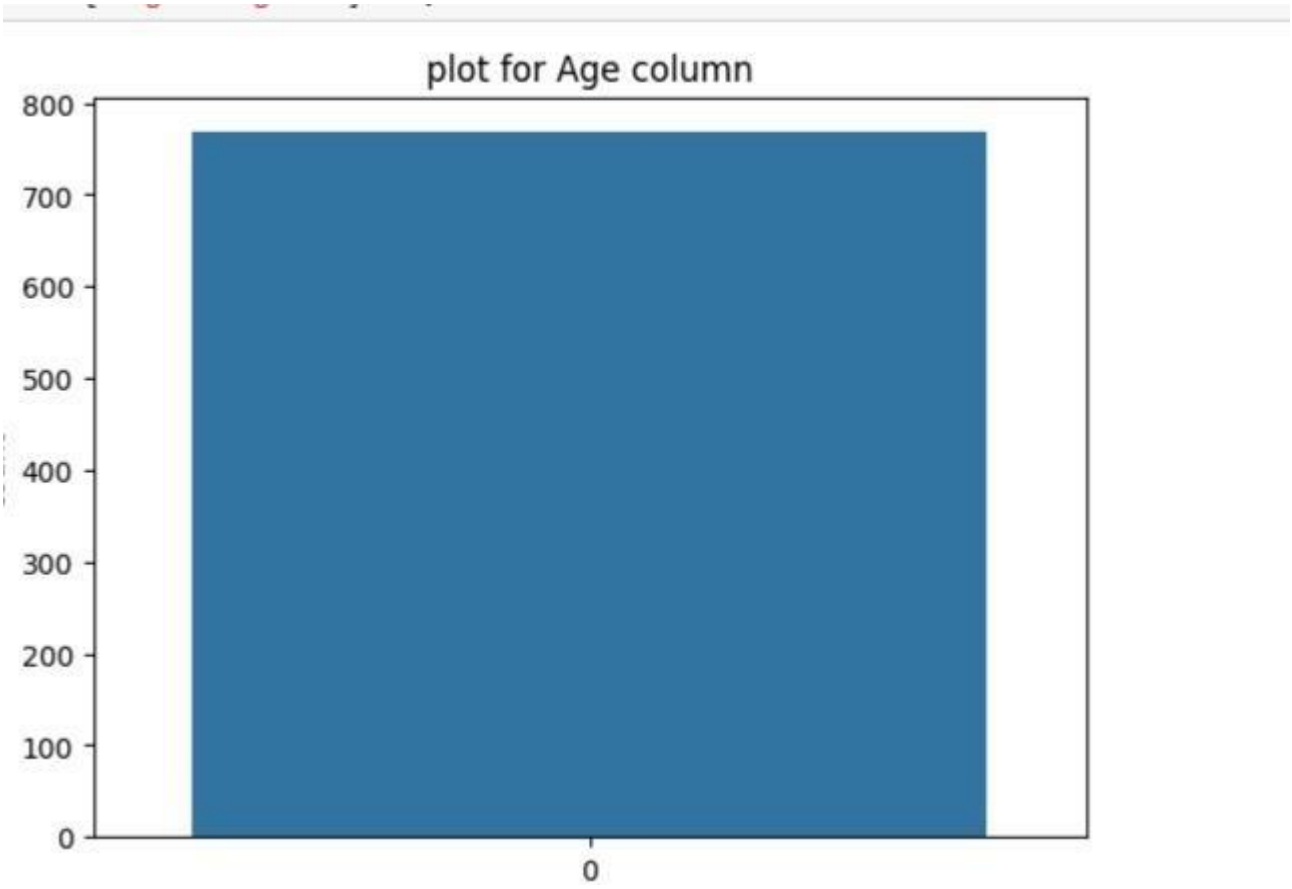

Fig 8.2 : Plot for age column

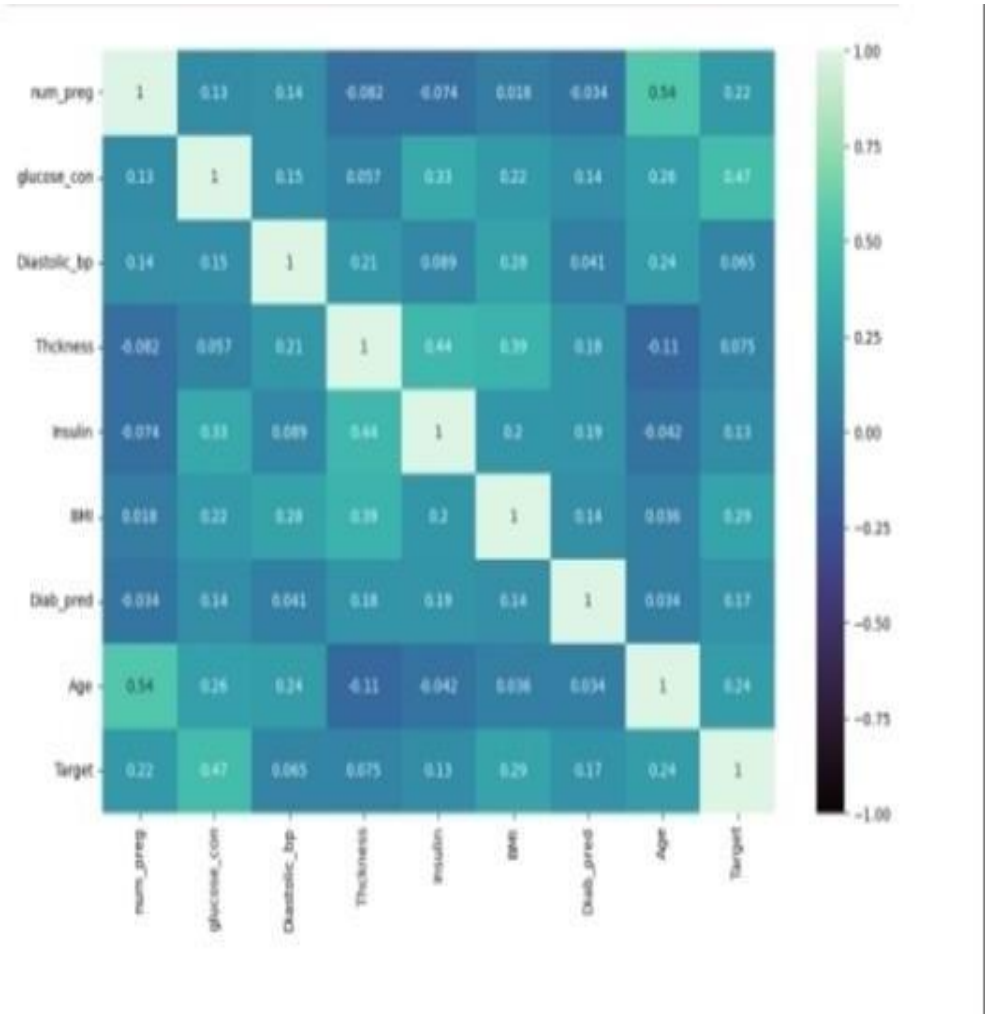

Fig 8.3 : Correlation matix

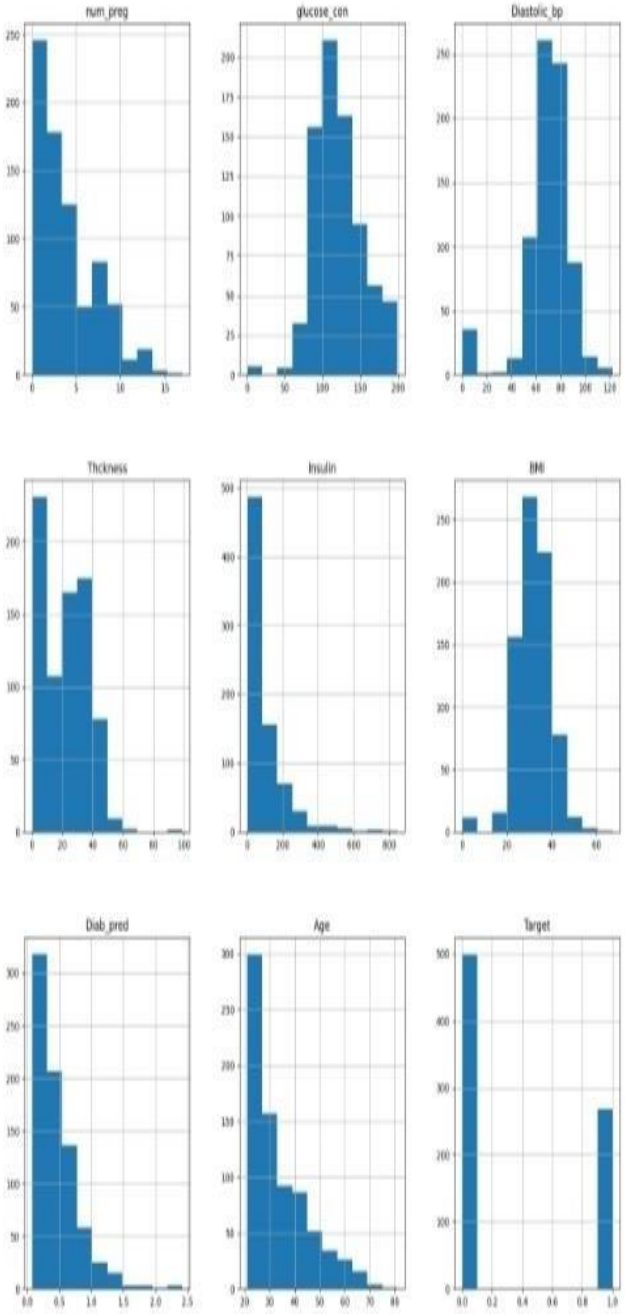

Fig 8.4 : Plots for different causes of diabetes

CHAPTER 9

CONCLUSION AND FUTURE ENHANCEMENT

9.1 CONCLUSION

Machine learning has the great ability to revolutionize the diabetes risk prediction with the help of advanced computational methods and availability of large amount of epidemiological and genetic diabetes risk dataset. Detection of diabetes in its early stages is the key for treatment. This work has described a machine learning approach to predicting diabetes levels. The technique may also help researchers to develop an accurate and effective tool that will reach at the table of clinicians to help them make better decision about the disease status.In this paper, we have used PIMA datasets for diabetes disease from the machine learning laboratory. All the patients‟ data are trained by using SVM. The choice of best value of parameters for particular kernel is critical for a given amount of data SVM approach can be successfully used to detect a common disease with simple clinical measurements, without laboratory tests. In the proposed work, SVM with Squirrel Search is used for classification. The performance parameters such as the classification accuracy, sensitivity, and specificity of the SVM have found to be high thus making it a good option for the classification process.

9.2 FUTURE ENHANCEMENT

In future the performance of SVM classifier can be improved by feature subset selection process.

REFERENCES

[1] D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar, "Big data in health care: using analytics to identify and manage high-risk and high-cost patients," Health Affairs, vol. 33, no. 7, pp. 1123–1131, 2014.

[2] M. Chen, Y. Ma, Y. Li, D. Wu, Y. Zhang, C. Youn, "Wearable 2.0: Enable Human-Cloud Integration in Next Generation Healthcare System," IEEE Communications, Vol. 55, No. 1, pp. 54–61, Jan. 2017.

[3] M. Chen, Y. Ma, J. Song, C. Lai, B. Hu, "Smart Clothing: Connecting Human with Clouds and Big Data for Sustainable Health Monitoring," ACM/Springer Mobile Networks and Applications, Vol. 21, No. 5, pp. 825C845, 2016.

[4] L. Qiu, K. Gai, and M. Qiu, "Optimal big data sharing approach for tele-health in cloud computing," in Smart Cloud (SmartCloud), IEEE International Conference on. IEEE, 2016, pp. 184–189.

[5] M. Qiu and E. H.-M. Sha, "Cost minimization while satisfying hard/soft timing constraints for heterogeneous embedded systems," ACM Transactions on Design Automation of Electronic Systems (TODAES), vol. 14, no. 2, p. 25, 2009.

[6] D. Tian, J. Zhou, Y. Wang, Y. Lu, H. Xia, and Z. Yi, "A dynamic and self-adaptive network selection method for multimode communications in heterogeneous vehicular telematics," IEEE Transactions on Intelligent Transportation Systems, vol. 16, no. 6, pp. 3033–3049, 2015.

[7] J. Wang, M. Qiu, and B. Guo, "Enabling real-time information service on telehealth system over cloud-based big data platform," Journal of Systems Architecture, vol. 72, pp. 69–79, 2017.

[8] J. Wan, S. Tang, D. Li, S. Wang, C. Liu, H. Abbas and A. Vasilakos,"A Manufacturing Big Data Solution for Active Preventive Maintenance", IEEE Transactions on Industrial Informatics, DOI: 10.1109/TII. 2017.2670505, 2017.

[9] K. Lin, J. Luo, L. Hu, M. S. Hossain, and A. Ghoneim, "Localization based on social big data analysis in the vehicular networks," IEEE Transactions on Industrial Informatics, 2016.

[10] Y. Zhang, M. Qiu, C.-W. Tsai, M. M. Hassan, and A. Alamri, "Healthcps: Healthcare cyber-physical system assisted by cloud and big data," IEEE Systems Journal, 2015.

APPENDIX

CODING:

```
import pandas as pd

df = pd.read_csv("pima-indians-diabetes.csv")


df.head(5)
df.tail(5)
df.info()
#checking null values
df.isnull().values.any()
#droping NaN,N/A values
df=df.dropna()
df.shape
import seaborn as sns
import numpy as np
from matplotlib import rcParams
import matplotlib.pyplot as plt
from matplotlib.cm import rainbow


sns.countplot(df.Age)
plt.title('plot for Age column')
rcParams['figure.figsize'] = 8,10
rcParams['figure.figsize'] = 15,20
df.hist()
corr = df.corr()

plt.figure(figsize=(12, 8))
sns.heatmap(corr, annot=True, vmin=-1.0, cmap='mako')
plt.show()
sns.countplot(df.Target)
plt.title('plot for age column')
rcParams['figure.figsize'] = 4,4
x_Data = df.drop(columns='Target').to_numpy()
y_Data = df['Target']
y_Data
from sklearn.ensemble import RandomForestClassifier as Featuree
```

```
from sklearn.preprocessing import MinMaxScaler
#normalizing features
scaler = MinMaxScaler(feature_range=(0,1))
train_samples = scaler.fit_transform(x_Data)
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(train_samples, y_Data, test_size =
0.30)
#Import svm model
from sklearn import svm
#Create a svm Classifier
clf = svm.SVC(kernel='linear') # Linear Kernel
#Train the model using the training sets
clf.fit(X_train, y_train)
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report,
f1_score
def print_score(label, prediction, train=True):
    target_names = ['Diabetes', 'No Diabetes']
    if train:
        clf_report = classification_report(label, prediction, target_names=target_names)
        print("Train
Result:\n================================================")
        print(f"Accuracy Score: {accuracy_score(label, prediction) * 100:.2f}%")
        print("_______________________________________________")
        print(f"Classification Report:\n{clf_report}")
        print("_______________________________________________")
        print(f"Confusion Matrix: \n {confusion_matrix(y_train, prediction)}\n")
elif train==False:
        clf_report = classification_report(label, prediction, target_names=target_names)
        print("Test
Result:\n================================================")
        print(f"Accuracy Score: {accuracy_score(label, prediction) * 100:.2f}%")
        print("_______________________________________________")
        print(f"Classification Report:\n{clf_report}")
        print("_______________________________________________")
        print(f"Confusion Matrix: \n {confusion_matrix(y_test, prediction)}\n")
y_train_pred = clf.predict(X_train)
y_test_pred = clf.predict(X_test)
print_score(y_train, y_train_pred, train=True)
print_score(y_test, y_test_pred, train=False)
cm = confusion_matrix(y_test, y_test_pred)
target_names = ['Diabetes', 'No Diabetes']
def plot_confusion_matrix(cm,
```

```
                      target_names,
                      title='Confusion matrix',
                      cmap=None,
                      normalize=True):
import matplotlib.pyplot as plt
   import numpy as np
   import itertools
accuracy = np.trace(cm) / float(np.sum(cm))
   misclass = 1 - accuracy

   if cmap is None:
      cmap = plt.get_cmap('Blues')
plt.figure(figsize=(8, 6))
   plt.imshow(cm, interpolation='nearest', cmap=cmap)
   plt.title(title)
   plt.colorbar()

if target_names is not None:
      tick_marks = np.arange(len(target_names))
      plt.xticks(tick_marks, target_names, rotation=45)
      plt.yticks(tick_marks, target_names)
   if normalize:
      cm = cm.astype('float') / cm.sum(axis=1)[:, np.newaxis]

  thresh = cm.max() / 1.5 if normalize else cm.max() / 2
   for i, j in itertools.product(range(cm.shape[0]), range(cm.shape[1])):
      if normalize:
         plt.text(j, i, "{:0.4f}".format(cm[i, j]),
                  horizontalalignment="center",
                  color="white" if cm[i, j] > thresh else "black")
  else:
         plt.text(j, i, "{:,}".format(cm[i, j]),
                  horizontalalignment="center",
                  color="white" if cm[i, j] > thresh else "black")
plt.tight_layout()
   plt.ylabel('True label')
   plt.xlabel('Predicted label\naccuracy={:0.4f}; misclass={:0.4f}'.format(accuracy,
misclass))
   plt.show()
plot_confusion_matrix(cm,target_names)
from sklearn import preprocessing
scaler = preprocessing.StandardScaler().fit(X_train)
```

```
X_train_transformed = scaler.transform(X_train)
clf = svm.SVC(C=1).fit(X_train_transformed, y_train)
X_test_transformed = scaler.transform(X_test)
clf.score(X_test_transformed, y_test)

from sklearn.ensemble import RandomForestClassifier
X_train, X_test, y_train, y_test = train_test_split(train_samples, y_Data, test_size = 0.30)
rf = RandomForestClassifier()
rf.fit(X_train,y_train)
rf.score(X_test,y_test)
rf.score(X_train,y_train)
y_test_pred = rf.predict(X_test)#testing the data
y_train_pred = rf.predict(X_train)#tesing the training data

print_score(y_train, y_train_pred, train=True)
print_score(y_test, y_test_pred, train=False)
```