RESEARCH ARTICLE OPEN ACCESS

Expert System Based Diabetes Diagnoses Using Ensemble Model

Pappu Chandra Roy[1], Sudhir Kumar Mishra[2]
[1] *Student M. Tech, ECE, Chandigarh University, Punjab, pappuchandra5876@gmail.com*
[2] *Associate Professor, ECE, Chandigarh University, Punjab*

ABSTRACT:
The title "Expert Systems" was given by John McCarthy, in which he entitled his study as "Applied Science and Technology" of making quick witted devices". The development of 'thinking' computer systems popularly known as "Expert Systems" furthermore; also known as "Synthetic Intelligence" which explains the idea of learning these algorithms so that they can act intelligently. The mindset of software programs is basically to execute their activities intelligently like humans. The designs that allow software programs to work in such a way which frames its society as the ability to act intelligently known as Expert systems. The ability of these systems in order to use relevant connection between database which can be used in primary level screening for various disease, proper medication as well as estimating various results. In today's growth following approach is beingused in all useful domains which are beneficial for the society so that complex issues could be solved on time. This domain was being setup on the increasing demand the central attribute of human – mind can be imitated via these Expert systems.The work executed in this domain depicts how these expert systems could work intelligently provided if the data fed to this system is preciselyaccurate and tested by cross validation. To carry out this research total fourteen classifiers are being studied out of which we find four algorithms more suitable for this research namely Artificial Neural Network, K-Nearest Neighbor, Support vector machine, Naïve bayes. On top of it, to make our results more accurate and efficient, ensemble technique *is* used which predicts the output bytaking the majority number of votes from the results predicted by these algorithms; in order to make this tool more efficient and accurate. The software tools which are being used are to execute this research are matrix laboratory and weka 3.6.13. After detailingof medical history, we prepared a rich dataset of around 400 people which is taken in the form of questionnaire from different sections of the society on the basis of ten physiological parameters. This enriched data has been tested and validated in terms of accuracy and prediction of correct data. This trained dataset is fed to graphical user interface and is tested so that it can predict correct output on the given input data provided to it.The values which are provided to the interface are five numeric values and restare nominal values. The fig which illustrates the working of this tool is shown in fig 3. The foremost aim of this proposed work is to build an intelligent tool which is capable of predicting accurate results and provide correct data so that it could prove helpful in medical domain and can be used for initial screening of patients at an earlier stage so that proper treatment can be followed by them.Talking about the privacy of these algorithms out of all algorithms ANN outperformed with accuracy of 97.5% and our proposed ensembling technique assured the accuracy of 98%.

Keywords: GUI based diagnostic tool, Ensemble method, MATLAB, Diabetes, WEKA 3.6.13, Classifiers and Expert Systems.

Date of Submission: 10-06-2023 Date of acceptance: 24-06-2023

I. INTRODUCTION

The use of its algorithms in medical domain is gracefully wide spreading its work in the world. It can be used for primary level screening of peoples so that they can get proper medication and treatment on time. It can be really beneficial for the people which are living in rural areas or urban areas where the medical attention is still far away. By using these expert based systems in rural areas dispensary or local clinics they can be pre-diagnosed with the disease at an early stage and also can be diagnosed with further medications on time so that the increasing death rate due to lack of medical facilities can be reduced considerably. The coding done in these algorithms is used to recognize the complex patterns from training phase and then assort them in testing phase which means to decide in which problem it would be more suitable in order to predict a useful and good outcome. Intelligence based expert systems prepare their database from various questionnaires which are being used for data collection, various medical treatment records pertaining to related disease which is easily available in medical hospitals, private clinics or local dispensaries, various online medical cases record available in different websites given by UCI repository that plays a pivotal role in the training phase of these systems. Many hospitalizations provide a well- resourced tool from which a useful and good approximation can be generated. The patient record, which are stored in computers in order to maintain medical records, can also be used in various study, in investigating the data and in health information management system [1]. Artificial Intelligence, Intelligent system, artificial networks as well as diverse systems can be produced which can be in the form of stimulation-reaction pairs. Among these frameworks none of them can act intelligently until a learning and training phasesapplied on them.

DIABETES- The term diabetes as the name depicts tells us about miscellaneous disarray that influence the ability of ourbody system employ the vitality of food supplies. In today's progressing generation diabetes is a very common issue that is being faced by almost in 1 out of 3 every person. In additionto it, it is wide- spreading more due to lack of healthy diet andour daily exercise routine in our day-to-day lives. The ratio of persons suffering from diabetes from different classes and different regions of the society is also widely increasingshowed in review taken by IDF. Figures shows that people living in Near East as well as the people situated near geographical area were at the majority population suffering from the disease. On the contrary, United Kingdom of Britain, Pitcairn Islands, Northern Ireland, China were the countries in which people suffering with diabetes found in majority of adults. Hyper glycaemia which is considered in type-2 of diabetes is found maximum in females especially which are pregnant and the risk of having same disease to their children is increased considerably which can increase the risk of developing disease genetically. As a result of which many upcoming generations could suffer from the same disease in future [2]. t is basically a long-lived disease which lasts longer until the proper medication and treatment. It identified to individual at their exact time otherwise, it will keep rising under different sections of the society [3]. Frequent symptoms are excessive thirst, urination, Hunger and blurred visions, Fatigue etc.

Cause - It normally starts from blood sugar obstruction, which is a state where some parts of body cells are unable for the intake of the sugar properly due to which different body parts wants more sugar or glucose so that they would be ableto enter body cells. These pancreases itself develops a hormone called "insulin". In this process, the cells change inthe form of glucose which is consumed by persons in the formof meal that is taken by person. Diabetic persons usually cannot develop the sugar in them neither their body cells cancreate it, as they should. This state of disease doctors named them as pre-diabetes. Few major causes of diabetes are discussed as under-.

1. **HEREDITARY-** Many researchers study shows that several pieces of genes influence person's bodyto stop from making glucose which increase the riskof disease among them.
2. **BEING OBESE**- When a person has extra pounds around the middle. It affects kids, teens as well as adults mainly due to overweight.
3. **METABOLIZE DISORDER-** People which are diagnosed with pre-diabetes usually suffer fromvarious states which includes increased blood-
sugar, being obese, increased blood pressure etc.
4. **DEVELOPING EXCESS SUGAR**- In this state, if the insulin is down, then the kidney usually startsmaking as well as sending the glucose.
5. **UNABLE TO SEND OR RECEIVE SIGNALS-** More often, it happens when our body parts cells are unable to communicate with each other or theyreceive.
6. **DAMAGED BODY CELLS-** In this, the body cells which are able to produce glucose in human's body more often they deliver incorrect measure of sugar at incorrect duration which leads to breakage of communication between the cells.

HAZARDS- Some identified risk factors of diabetes given asunder: -

- **Time Period: - 40 or above**
- **Background History**- Any person in the familysuffering from disease.
- **Being overweight or obese**
- **Hypertension**- It is a highly hazardous, even if it istreated and under control.
- **Acanthosis nigricans-** It is a state in which there aredark rashes near neck or armpits.

Frequent Indications- Most common signs of a person suffering from the pertaining disease could be so moderate that someone can't even detect it in their daily routine. Even according to IDF about 8 million people who have it don't know it. Some symptoms of diabetes are as under-Feeling thirsty, Frequent urination, Blurred vision,Tingling or numbness in hand or feet, whether the person smokes or not,fatigue.

Problem Statement- The diagnosis of Diabetes involves a number of clinical tests of the patient under consideration.The diagnosis can be divided into two phases:

1. Diagnosing whether a person is suffering fromDiabetes or not.
2. To determine the stage of Diabetes of a patient.

Number of Diabetic patients in lower class and mid class countries in terms of earnings is way more than high wage countries. the reason being that people in high income countries go through regular health screening. Moreover, in lower income countries health infrastructure is poor, the numbers of doctors are very few and have to share the burden of addressing the needs of a large number of patients.

The use of machine learning methods could assist a doctor in diagnosing disease in less time with high accuracy with the minimum number of tests required could be of great help. Doctors can easily cross check their findings in relatively short time with the help of such a technique.

The current study is focused on to analyze and employ machine learning classification techniques to performdiagnosis of Diabetes expressly to find if a diseased person is diabetic or non-diabetic.

II. RELATED WORK

Russell et. al. [1995] [4] have proposed a book called ‘Artificial Intelligence: A modern approach’. In this manuscript, it is subdivided into many parts so that the reader can acquire useful and complete knowledge. Few are discussed below- i) It clearly aims at the use of expert systems in various fields whether we talk about the area of gaming, automation and mobiles, satellites and space, medical domain. It tells that how these expert-based systems can be used by means of programming which can help various researchers and doctors for the welfare of the society. ii) Next topic comes of "Problem Solving," which tells us about various stepwise approaches or procedures that should take during the programming of these systems so that chances of error should approach to zero and chances of success should approach to infinity. It basically talks about the idea that machines also do think and can take intelligent decisions likehumans. iii) "Knowledge and Reasoning,” as the name suggests, it focuses on the ability of machines that can give explanation for their taken decision if required and also has the power to reason strongly for the decisions taken by thesesystems. iv) "Acting logically," that tells us about using of logical reasoning in the designing of these systems like- to solve puzzles or any other game in intelligence is required like chess or quiz. v) “Uncertain Knowledge and Reasoning," it concentrates on reasoning and decision-making.

Dreiseitl et.al. [2002] [5] have proposed a system named ‘Reverting and machine learning classifiers.’ Logistic regression is a paradigm in which it has a dependent variable- a variable which can predict only two possible outcomes. In this method the output is generally predicted in discrete values i.e., ‘0’ or ‘1’ which ‘yes’ or ‘no’. Recent study shows that Logit model can be used in various fields like science field, computer research fields. In field of medicines etc. This type of models usually learns from past experiences data which helps them to modify themselves accordingly. It workson some set of rules also they compare themselves from other data mining classifiers so that they can do the needed changes which will help them to work better than others. After training their data, rigorous testing is done, and graphical representation is generated which is considered as results of this regression. For instance, if there more than two outcomes it is considered as multiple regression. In this article, logistic regression and artificial neural networks plays the same role statistical pattern recognition, also it is observed that ANN isjust a stimulation of Logistic regression as both work in the same way. These both techniques have their roots in two different communities (statistical and computer science) but have many similarities. Furthermore, both techniques are compared with different data mining classifiers like Boosted trees, Random Forest, Ada Boost, Perceptron etc. to check their accuracy and performance. In this literature, 72 papers were sampled comparing the two classifiers i.e., logistic regression and neural networks in the field of medical. The results showed that at present, Log-it model and ANN both are most popular classifiers which are widely used in area of biomedicine, according to numerous research papers reviewed in Med line- 29,800 for regression model, 9500 for ANN, 4200 for Random Forest, 1200 for Random trees, 850 for Ada boost and 100 for SVM.

Temurtas,Hasan et.al. 2006 [6] have discussed ‘Comparative study on diabetes disease diagnosis using neural networks. The work done in this study gives a review on Diabetes foundation of India which is diagnosed with the help of multiple layers of neural networks which was first train by using Levenberg Marquardt algorithm and by calculating the probability of these neural networks. In this case study, as a result of which two neural networks and results were matched with Pima Indian diabetes. The classifiers which were being used in this study are types of neural networks classification to which training is being done on neural networks by using Levenberg Marquardt technique. This classifier is used in many software developing tools usually for the solving the problem of generic curve-fitting. The standard way of learning and testing phase is by applying deanery validation to enhance the accuracy of the classifier. After testing phase, the outputs are matched with these neural networks classifiers to reduce the chances of faults producedin the outcomes. Finally, outputs are being matched with research reviewed Pima diabetes disease prognosis by employing the same datasets. Reports shows that the databases of Pima Indian diabetes is divided into two parts with nearly about 785 cases. The group is divided as Part i) - typical (600) ii) Pima Indian diabetes (278). All such cases have eight attributes, the attributes which are taken are- Attribute 1: Total number of times pregnancy occurs. Attribute 2: Testing of plasma glucose concentration a 2-hit person's body. Attribute 3: - Measure of the pulse pressure (mmHg). Attribute 4: Arm skin fold breadth. Attribute 5: 2-h plasma glucose (lU/). Attribute 6: BMI-body mass index. Attribute 7: Diabetic extraction work. Attribute Feature 8: Age of a person (in years). The out turn of this research paper was evaluated with the calculation past research reported aiming on diabetes mellitus by using UCI data knowledge discovery of data as the conclusion, It has observed that these neural networks can prove as helpful and useful in the prognosis of pre-diabetes.

Tomar, Divya et.al. [2013] [7] have proposed a system named 'A review of Data Processing in Healthcare domain'.This survey expands the research area of Artificial Intelligence methods like classifying the data, clusters the datasets, making use of logistic model. Artificial Intelligenceis one of the most popular research areas and is encouraging research fields of artificial intelligence. It advantages of extracted useful and valuable data through bulk volumes of datasets makes it unique from other software. In today's world, the use of AI is wide spreading in many fields especially if the area is biomedicine and medical domain. The use of Ai in medical domain has many pros such as by using this tool it removes the necessity of unnecessary tests which were taken during patient's diagnosis, availability of this tool in local dispensaries ' or clinics of people residing in rural areas so that they can be diagnosed with disease initially, making it affordable and at low prices, detection of the fraud in health insurance, recognizing the risk factorsand symptoms of disease earlier so that medical treatment could be started timely. Throughout this survey, the achievement of AI was noticed in medical domain. In this survey, the datasets produced by the health corporation was very wide and compound due to which it becomes hard to inspect the dataset. Different types of classifiers which were used to detect the patient disease includes- k- Nearest Neighbor (KNN), Decision Tree (DT), Support Vector Machine (SVM), Regression, Neural Networks (NN), Bayesian Methods, A prior Algorithm, Frequent Pattern Tree algorithm. In classification process the dataset is categorized in two phases- training and testing by which prediction is being made for each dataset after doing testing on these values. For instance, the person can be predicted as "high risk" or "low risk" which totally based on these patterns thatwere observed during training and testing. Some of the algorithms are discussed below- a) By applying these algorithms, there is a demand to identify the redundant unseemly parameters as the parameters starts behaving as a noise and give wrong results which ultimately effects the speed of job b) We analyzed that only when more than two classifiers are used in designing of any tool it can generate better results which is not possible if only an individual classifier is used. There is a fundamental requirement of Development of thinking computer systems to recognize thecomplex patterns. These classifiers provide a novel data concerned with medical domain which can be turned useful in the sense of taking managerial and medical realm decisions like evaluation of medical staff, resolving policies of medical insurances, selection of suitable treat which can be proven helpful for person who is ailing with disease, initial screening of disease.

Abid Sarwar et al. [2015] [8] presented a system called 'Hybrid ensemble learning technique for screening of cervical cancer using Papanicolaou smear image analyses. In this work, ensembling methodology is used which clubs theresults of four machine learning algorithm and thus by taking majority votes it predict whether the person is suffering fromdisease or not. This technique is used conductive to enhance the exactness and effectiveness of diagnosis so that a propertreatment can be taken by a person timely. Artificial intelligence which can be used in the screening of cervix tumor after doing a not only evaluation but also its categorization of Pap smear images. These samples of human's cells were spotted during Pap smear technique by which these cells were examined analyzed that during test the existence of uncommon growths were found in these testing cells which shows that there could be a possibility of having malignant tumor and potentially precancerous changes. In case of unusual findings, are seen then the case is aided for the diagnostic procedure. When these cells are tested to check whether any unusual or strange developments are growing in the cells or not, so that an immediate action could be taken to stop these abnormalities from further increasing and early medical treatment could be started in order to reduce the increasing death rate which is due to cervical cancer. This ensemble method is a very well-known technique which is widely known for screening of cervical cancer. Furthermore, in this work the comparable estimation with concerning various machine learning classifiers which are useful in primary screening of cervix cancer. In order to evaluate the results predicted by these techniques calculationdone by root mean square and correctly classified percentage is taken as primary source for analysis. Thus, it can be competently used in prognosis of cervical cancer.

Out of various classifiers which are used in this study, the hybrid ensemble method shows better results in comparison with all other counterparts with an accuracy of 96% in case of two-class problem and about 79% for seven-class problem. These predicted values when made in comparison with other techniques it is observed that this technique is considerably better for both 2-class and 7-class problems.
Victor Chang et al. [2016] [9] focused on the creation a machine learning algorithm-based method for detecting heart illness. He illustrated how artificial intelligence may be used to predict whether or not person is ailing with the concerned disease. On top of it, a scripting language software was built for research investigations since it is more reliable and makes it simpler to track and set up various medical monitoring equipment. They show data processing by transforming categorical columns and dealing with categorical data. They also talked about the important stages of software development, such as data gathering, logistic regression, and feature evaluation. To better identify cardiac problems, a random forest classifier method was developed. Such a solution, which was judged significant because to itsabout 83% rate of precision across training cases, necessitated data analysis. The random forest classifier algorithm, as well as the tests and outcomes, will be detailed next. This process increases the precision of research diagnosis. The aims, restrictions, and research contributions of the work were based on previously available information.

Rati Goel et al. [2021] [10]**,** the Cardiac muscle is important in the human body. Greater precision and accuracy arerequired for heart disease diagnosis and analysis. Cardiovascular disease is a potentially fatal disorder. This disease arises as a result of a variety of problems with the human body, including excessive blood pressure, sugar levels, hypertension, lipids, and so on. This work was investigated, and heart disease was predicted using Python and sophisticated analytics. The author claims that the data set's multiple properties may be utilized to forecast this sickness. They had obtained a data set of 13 factors and 383 individual values to evaluate the patients' performance. The article's major purpose was to increase the accuracy of heart disease diagnosis using ML algorithms.

III. MATTER & METHODS

1. Dataset used for this proposed work.

For preparing the useful and valuable database given in this report, the literature survey of diabetes was analyzed and further research that can enhance the performance is being carried out in this related field [11]**.** In order to carry forward this research and to enhance its performance we seek advice from concerned diabetologist and discuss our work. After detail study, we come to conclusion that there are 10 parameters which plays vital role in detection of disease. By reason of these attributes a rich dataset has been prepared of around four hundred people across wide geographical. While preparing this dataset both the quality of data and assortment of dataset has been taken care of. This dataset is divided into two parts- diabetic and non-diabetic. The dataset presented inthis paper has been taken from society consisting urban, rural areas majorly upper & lower class. The people from various age sections, with adjacent eating habits, copious smokers, non-smokers category, drinkers and non-drinkers etc. Age group variation is from minimum 5 years and maximum age of 78 years. On top of it, in order to evaluate these results discrete values are assigned 0 and 1 for analyzing the result and to maintain the uniformity in the record taken.

After a detailed literature study about the selected disease followed by consultation with the medical expert, ten different physiological parameters were identified that act as significant tasks pertaining to diabetes. These parameters were Age, Family history, Weight, Gender, Drinking, Smoking, Thirst, Frequency of urination, Height and Fatigue. On basis of selected medical disorder, the primary data was obtained from multiple healthcare center and by door-to-door collection using questionnaires in order to perform various tasks viz. learning and cross validation phases which are obtained by using expert systems and its techniques.
The parameters selected, are given in table 1 as shown below, also represents a data of interval-frequency aboutnumerous parameter values.

parameter	Description	Range of values	Analysis of data
Age	Age of the Person	5 to 78	Age 5 to age 20: 30 Age 21 to age 35: 131 Age 36 to age 50: 142 Age 51 to age 78 : 97
Gender	Gender of the person	0 or 1	Male: 190 (represented by 1) Female: 210 (represented by 0)
Smoking	Whether the person is smokeor not	0 or 1	Smokers: 68 Non-Smokers: 332
Drinking	Drinker or non drinker	0 or 1	Drinkers: 79 Non-Drinkers: 321
Urination	How many times person urinates in day	1-15Times	1-5: 195 6-10: 153 11-15: 52
Thirst	How many times persondrinks	1-15Times	1-5: 112 6-10: 196 11-15: 92
Height	Height of a person	60-185 cm	60 – 95: 7 96 – 125: 9 126 – 155: 119 156 – 185: 265
Fatigue	Healthy levels of fat massfor a fit person	0 or 1	Fatigue(Yes): 276 Fatigue(No): 124 Min-5% in men, 12%in women Max-25% in men, 32%in women Average-15 to 18% in men, 22 to 25% in women
Weight	Weight of the person	15 to 96	15 – 36: 13 37 – 56: 110 57 – 76: 244 77 – 96: 33 Average weight-62 kgOverweight-34.7 %
Family History	Any person in family is diabetic or not	0 or 1	Family History(Yes): 116 Family History(No): 284
Diabetic	If a person is diabetic or not	0 or 1	Diabetic: 149 Non-Diabetic: 251

Table 1: Details of the various parameters used and their analysis

Out of all attributes which are considered for this research, Age of the person plays the foremost role in analyzing as it had been observed in past records that the people whose ageis ranging between 30-35 are at high risk of type II diabetesbut it had been seen that people whose age is above 35 are prone to type I diabetes. Furthermore, due to unhealthy lifestyle in today's generation it has been seen that it is morespreading in children than in adults. Another parameter- Family history also plays an important role in detecting the disease, as it has been frequently seen that if any person has family background or there is heredity in their family whether a child or an adult of that family is more prone to have that disease. Generally, people are ailing from disease are unable to maintain healthy lifestyle as a result of which their pancreas becomes incapable of producing enough insulin in their body required to support the body's glucose quantity. As a result of which, that person feels tired, change in weight, rise of frequent thirst are the common symptoms which a diabetic person feels. The research in this related medical domain shows that various methods have been proposed which shows that if any person has much more intake of meal than required and has increased hunger or

3.2 Considered algorithms for study.

The diabetes diagnoses design, it is powered by using four algorithms namely.

- Artificial Neural Networks
- K-nearest neighbor
- Naïve Bayes
- Support of vector machine.

Artificial Intelligence algorithms can be used in various domains like medical domain, in designing AI based games such as 'Alpha Go', AI based vehicle like 'Spirit, Robotic surgeon, AI based chess player 'Deep Blue'. All these sections of society in which AI has been used shows an incredible performance which reflects that machine can

3.2.1 Artificial Neural Networks

Artificial neural network is inspired genetic technology network of neurons. This algorithm is proven great in modeling the methodological data, which is efficient way of encapsulating, reprehending and capable of mimicking complicate datasets among entries as well as the possible outcomes by acting as a numerous analogous estimation. In Artificial Neural network, the networks combined via various appetite, generally has high level of ammonia in valproic acid therapy [12]. The study shows that the persons which are suffering from excess obese or being excessively overweight which is clear evidence of diabetes. However, by applying various procedures of bariatric surgery on person it has been observed by the researchers that the number of persons which undergo this surgery not only lose weight but also kept it more than three years and more. This incredibly method can also be used for the curing of diabetic patients [13] think and can take intelligent decisions like humans [14]. The brief narration of four algorithms which are described above are elaborated below-surgeon, AI based chess player 'Deep Blue'. All these sections of society in which AI has been used shows an incredible performance which reflects that machine can think and can take intelligent decisions like humans [14]. The brief narration of four algorithms which are described above are elaborated below-layers out of which one is input, second is hidden and third yields the outcome to achieve desired result. In this method, the neurons weights are adjusted repeatedly until the desired outcome is achieved. While Figure 3 represents the structure of a typical artificial neuron**.**

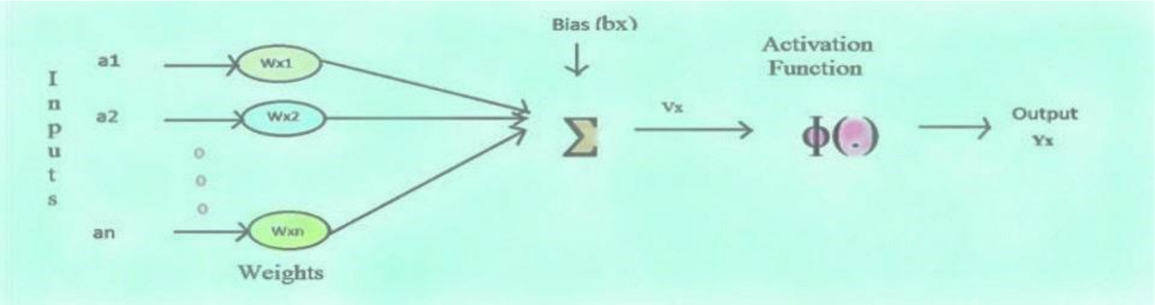


FIG 1: - A Artificial Neuron

A neural network can be single layer or multilayer i.e., it consists of one or more hidden layers. Figure 4 depicts a neural network with two hidden layers; the input layers nodes are passive doing nothing but simply forwarding the values from input to multiple outputs whereas the hidden and output layer nodes are active nodes and do actual processing.

Weights which are considered as entries moves in forward direction i.e., moving from input to hidden layer by adjusting weights and then yields the output. The ANN performance is improved by utilizing error propagation algorithm. In this error back propagation learning is applied into two phases

3.2.2 Naïve bayes

In this type of algorithm, it is being founded on three basic concepts namely- Naive Bayes algorithm is based on well- known 3 concepts-preceding, possibility and prediction where preceding means record of past data, which is attained during the incident, preceding means the probability of that incidence to see if the same can occur in future or not.

Prediction= Preceding * Feasibility / possibility
Mathematically interpreted as-

(Probability) (B Given A) = (Prior Probability) * Probability(A and B)/ Probability (A)

In this algorithm, learning is done by presuming that theseattributes are separate in given class. [25]. This algorithm is founded on basis of supposition closeto the independence regarding parameters and for thisbecause it is called as naive. In accordance with the theorem,this assumption of

N could be analyzed on the groundwork ofN also for the fact of assumption made by K. Given by the formula-

$$P\ (N|K) = \frac{P\ (K|N)\ P(N\ )}{P\ (K)}$$

P (N): Past likelihood assumption for N.

P (k): Past likelihood for unclassified data K.

P (N/K): Assumption of N given by K. P (K/N): Assumption of K given by N.

This estimated given data may be possibility. This enhanced classifier, outperformed among all other classification methods in terms of its feasibility, efficiency, accuracy, correctly classified and incorrectly classified data. Naive bayes algorithm is based on supervised learning which means learning with teacher by using past data record the future prediction is being made on basis of past incidences [15] .

3.2.3 k-NEAREST NEIGHBOR

KNN algorithms are very easy to acknowledge as KNN algorithm works astonishing in practice. During this method, two groups are formed which are located next to each other and respond to input vector. In case of one or two dimension it becomes simple to respond for unknown dataset by Self organizing maps that groups the input data into clusters. This algorithm is used for classifying the unknown data and can predict problems related to regression for eg- if there are N no. of training vectors, by using this algorithm and by taking the value of k (say k=3), it chooses the first 3 closest neighbor which are close to unclassified dataset and by taking majority votes it predict the class of test data. This algorithm is also known as lazy learner, which is being used for training the database and then stored on querying similar data between test data and training data records. KNN model considerably enhance the ability of this technique by using only few attributes in order to classify the data. Recent study illustrates that mean rate is decreased by 92.34% by using this model [16]. The figure 1 depicts the flow diagram of k-nearest neighbor, where 'k ' is the number of test data. Some of the commonly used distance metrics are Euclidean, Manhattan, Minkowski. But in case of this concerned topic, Euclidean distance is mainly used, higher the value of K it becomes hard in order to recognise among the classified and unclassified datasets.

3.2.4 Support vector machine

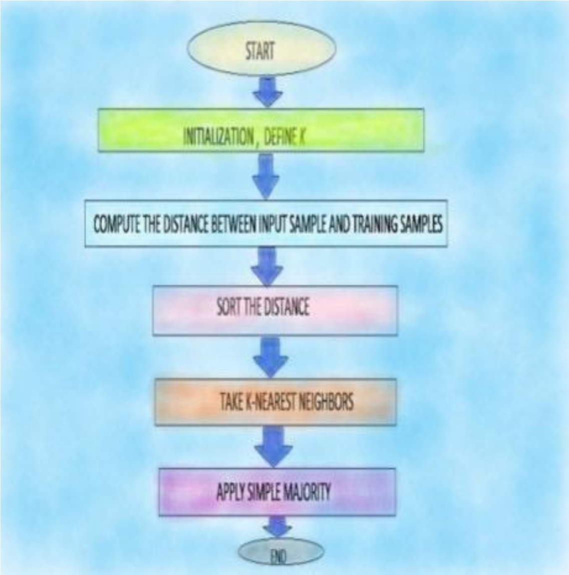


FIG 2: - KNN FLOW CHART

Support Vector Machine is a type of machine learning algorithm which is supervised type used for grouping of different type of datasets. These algorithms are founded on systematic risk by minimizing principal and with help of studying statistical learning theory, a hyperplane is used to divide two different classes of datasets. For resolving this purpose SVC and SVC rotates around the perception of a “margin”- i.e. a line which dissect the training and testing data by implying a margin on both its sides. If the margin is increased, it creates maximum possible length among the line and instances are marked on either side of hyperplane that be used to reduce an bound on the error. Data working on two types i.e. linear separable and linear non-separable data. Likewise, in former case, only one hyperplane is needed for separating the data but in the case of latter more than one hyperplane are needed. Though, in case of SVM there is nonlinearity in boundaries of arbitrary complexity, we limit ourselves, in this paper, to the linearity of SVM. The study presented in this paper, is taken by surveying of machine learning algorithms and diagnosing the disease with support vector machine. In this work, research has been reviewed which shows that this algorithm shows a good performance in medical domain and can provide more accurate accuracy in comparison with other algorithms of Artificial intelligence which are being used in medical domain. In previous years i.e., before 2005, this algorithm wasn’t being used in most of the fields the reason behind this was due it its incapability in predicting the accurate results. But the recent study, it has been observed that support vector machine has shown a tremendous growth in all traits whether it is in medicine, automation, image, sensors, in the design of various games, aerospace. This algorithm is now being used in all traits of research field and also play an important role in it. Due to its accurate predictions, it has proved beneficial in various traits and would be useful for future research due to its outstanding capability in predicting the estimated outcomes [17].

Attributes can be plotted each side of plane which lessens upper portion of class and calculate the error. Fig 5 depicts the hyperplane which divides the two sections of different classes.
In this algorithm, it is subdivided into two planes- hyper- plane and line. In hyper plane or let say in case of first scenario thumb rule is being used so that it can recognize correct line which means selecting a plane to classify star, circle and maximizing the farness between the imminent data point which is called as Margin/ line. For improving the model performance of SVM some parameters are used suchas "Kernel”, “Gamma” and “C”.

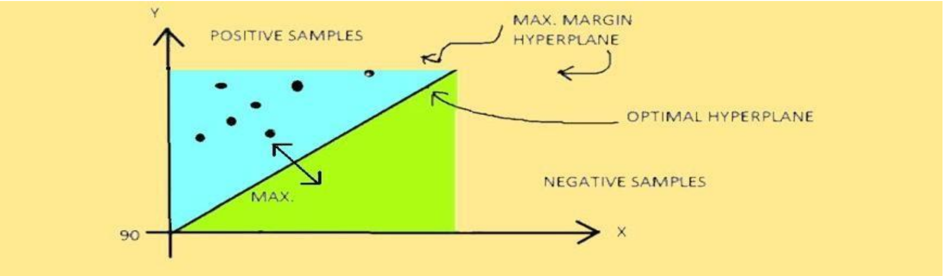


Fig 3- An problem example of two class with one hyperplane separation

3.3 PROPOSED MODEL

This model represents the proposal of implementing the ensemble methodology i.e. a novel method used for improvement andpredict the performance of an Expert Systems used for diagnosis of diabetes.

Ensemble Method

In Ensemble method, the possible outcome of each algorithm is taken and by means of taking majority votes we can predict the data. By using this technique, it increases the chances of accuracy and efficiency of each classifier in sucha way that it can be proved as an efficient tool used for prognosis of diabetes. In this method, different algorithms results are club together to predict the result in interface. Letus assume, in case if a particular result gets incorrectly classified by an individual algorithm, then in that case the error is easily rectified by other algorithms by taking the major ensemble classification method task is completed by constructing a broad numbers of data processing techniques during insight phase & predicting the outcome by calculating the mode between the algorithms used in it. Thistechnique would enhance the results by combining the results of other four algorithms any taking maximum voting itredicts its output [18]. rity votes which are taken by individual classifiers.

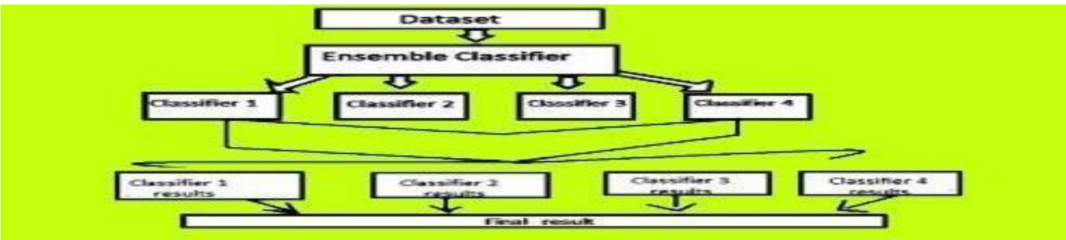


FIG 4- REPRESENTATION OF PROPOSED MODEL

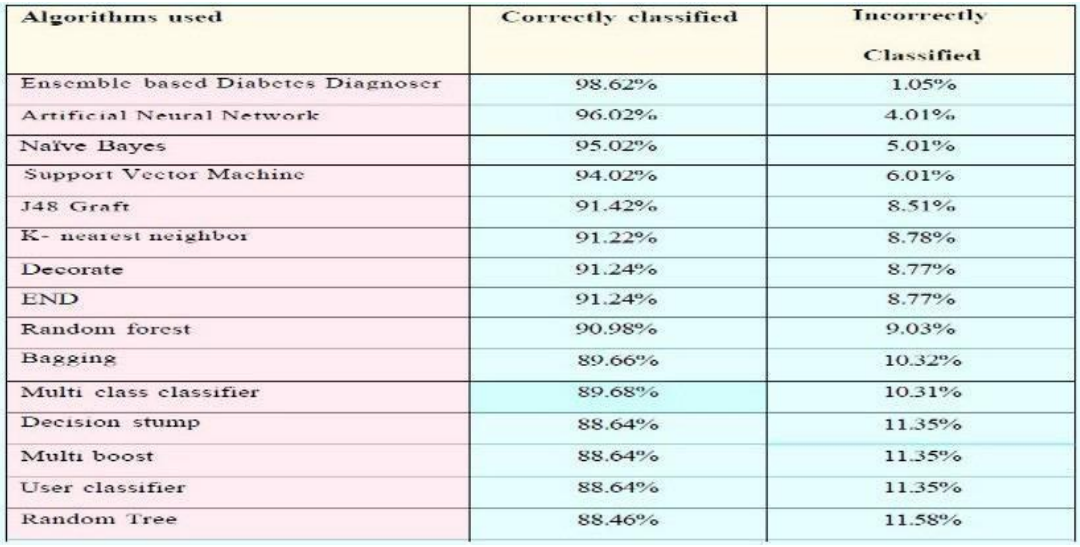
OBJECTIVES& THEIR DESIRED RESULTS

The objectives of the proposed study are as under:

1. To study all the existing methods and algorithms in machine learning, understand their working, advantages, disadvantages and applications so that we can analyze which helps in recognizing these attributes and decide which attribute is more suitable in the field of medical diagnosis.

DESIRED RESULT

This expert system based proposed ensemble model which is being used for prognosis of diabetes. When comparison is made it works efficiently with all AI algorithms. By using this proposed model the accuracy and efficiency is tested by using 10 fold cross validation. In addition to this, for choosing correct classification and incorrect classification WEKA tool is being used for better results.



Algorithms used	Correctly classified	Incorrectly Classified
Ensemble based Diabetes Diagnoser	98.62%	1.05%
Artificial Neural Network	96.02%	4.01%
Naïve Bayes	95.02%	5.01%
Support Vector Machine	94.02%	6.01%
J48 Graft	91.42%	8.51%
K- nearest neighbor	91.22%	8.78%
Decorate	91.24%	8.77%
END	91.24%	8.77%
Random forest	90.98%	9.03%
Bagging	89.66%	10.32%
Multi class classifier	89.68%	10.31%
Decision stump	88.64%	11.35%
Multi boost	88.64%	11.35%
User classifier	88.64%	11.35%
Random Tree	88.46%	11.58%

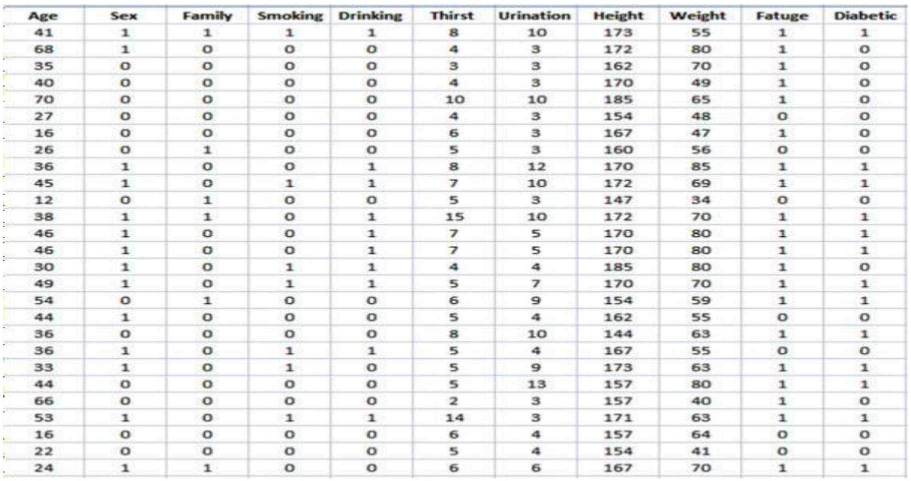
Table 2: Obtained outcome from various algorithms of performance metrics

The performance of the expert based systems is analyzed by calculating the average faults between the output data and the desired output data which is done in testing phase. Out of these four classifiers which are used in this manuscript, Artificial Neural Networks outperformed by predicting the most accurate results approximately of 96.00 % continued in order as naïve bayes (95.00%), SVM (94.00%), J-48 Graft (91.49%), KNN (91.23%), END (91.23 %), Decorate (91.23%), Random forest (90.97%),Bagging (89.69%), Multiple class classifier (89.69%), Multi- boosted Classifier (88.65%), User Classifiers (88.65%), Decision Stump (88.65%) and Random tree(88.40%). The performance of these algorithms can be increased if no. of incidence in these datasets are increased and by involving numerous objective that plays an important role in diagnosing the disease like diabetes.

2. To perform a detailed study of diabetes and its types to study which algorithm is best suited to prognosis of diabetes after collection of relevant data for training and testing and to analyse, understand and compare these algorithms on different parameters such as performance, reliability, and validity across different datasets etc.

DESIRED RESULT

According to the survey which is carried in field of medical domain and the concerned work in this domain. We seek advice from diabetologist and discuss the problem with them we come to conclusion that there are ten parameters which plays main role in detection of diabetes and have importanceused for the manipulation of disease.



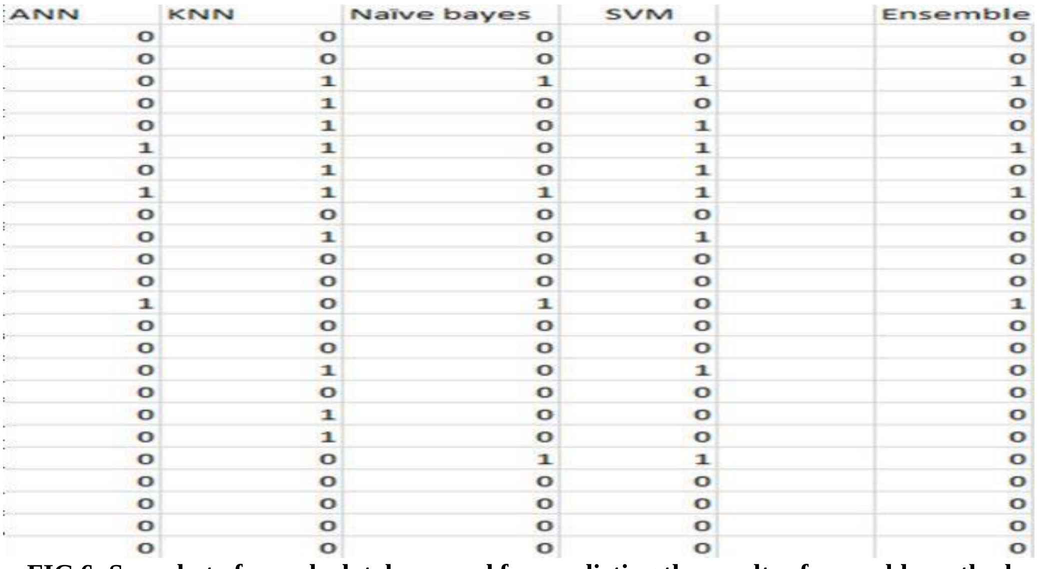
Age	Sex	Family	Smoking	Drinking	Thirst	Urination	Height	Weight	Fatuge	Diabetic
41	1	1	1	1	8	10	173	55	1	1
68	1	0	0	0	4	3	172	80	1	0
35	0	0	0	0	3	3	162	70	1	0
40	0	0	0	0	4	3	170	49	1	0
70	0	0	0	0	10	10	185	65	1	0
27	0	0	0	0	4	3	154	48	0	0
16	0	0	0	0	6	3	167	47	1	0
26	0	1	0	0	5	3	160	56	0	0
36	1	0	0	1	8	12	170	85	1	1
45	1	0	1	1	7	10	172	69	1	1
12	0	1	0	0	5	3	147	34	0	0
38	1	1	0	1	15	10	172	70	1	1
46	1	0	0	1	7	5	170	80	1	1
46	1	0	0	1	7	5	170	80	1	1
30	1	0	1	1	4	4	185	80	1	0
49	1	0	1	1	5	7	170	70	1	1
54	0	1	0	0	6	9	154	59	1	1
44	1	0	0	0	5	4	162	55	0	0
36	0	0	0	0	8	10	144	63	1	1
36	1	0	1	1	5	4	167	55	0	0
33	1	0	1	0	5	9	173	63	1	1
44	0	0	0	0	5	13	157	80	1	1
66	0	0	0	0	2	3	157	40	1	0
53	1	0	1	1	14	3	171	63	1	1
16	0	0	0	0	6	4	157	64	0	0
22	0	0	0	0	5	4	154	41	0	0
24	1	1	0	0	6	6	167	70	1	1

FIG 5 - Figure showing sample analyzed database used for algorithm training.

3. A proposed prognostic framework can be developed that could aid a medical doctor in diagnosis of diabetes. To test the proposed framework and verify its authenticity and validity can be done by using 10-fold cross validation.

DESIRED RESULT

The desired result like efficiency, accuracy has been achieved by using this methodology and it is observed that this is a beneficiary tool which can be used for initial levelof screening.



ANN	KNN	Naïve bayes	SVM		Ensemble
0	0	0	0		0
0	0	0	0		0
0	1	1	1		1
0	1	0	0		0
0	1	0	1		0
1	1	0	1		1
0	1	0	1		0
1	1	1	1		1
0	0	0	0		0
0	1	0	1		0
0	0	0	0		0
0	0	0	0		0
1	0	1	0		1
0	0	0	0		0
0	0	0	0		0
0	1	0	1		0
0	0	0	0		0
0	1	0	0		0
0	1	0	0		0
0	0	1	1		0
0	0	0	0		0
0	0	0	0		0
0	0	0	0		0
0	0	0	0		0

FIG 6- Snapshot of sample database used for predicting the results of ensemble method

Future Scope

This work can be further enhanced by including some clinical and genetic features also as parameters. Relevant data collection about clinical and genetic features can be made and optimize database can be used for training, testing and validation of intelligent system whose performance can be better than the system proposed. Expert system based on AI (as the one proposed in this thesis) should be encouraged and made easily available to people so that initial self-diagnosis can also be done by people who have symptoms relevant to diabetes

REFERENCES

[1]. Hogan, William R., and Michael M. Wagner. "Accuracy of data in computer-based patient records." Journal of the American Medical Informatics Association 4.5 (1997):

[2]. Kharroubi, Akram T., and Hisham M. Darwish."Diabetes mellitus: The epidemic of the century." Worldjournal of diabetes 6.6 (2015): 850.

[3]. Olokoba, Abdulfatai B., Olusegun A. Obateru, and Lateefat B. Olokoba. "Type 2 diabetes mellitus: a review of current trends." Omanmedicaljournal 27.4 (2012): 269..

[4]. Russell, Stuart, Peter Norvig, and Artificial Intelligence. "A modern approach." Artificial Intelligence. Prentice-Hall, EgnlewoodCliffs 25 (1995): 27.

[5]. Dreiseitl, Stephan, and Lucila Ohno-Machado. "Logistic regression and artificial neural network classification models: a methodology review." Journal of biomedical informatics 35.5 (2002): 352-359.

[6]. Temurtas, Hasan, Nejat Yumusak, and Feyzullah Temurtas. "A comparative study on diabetes disease diagnosis using neural networks." Expert Systems with applications 36.4 (2009): 8610-8615.

[7]. Tomar, Divya, and Sonali Agarwal. "A survey on Data Mining approaches for Healthcare." International Journal ofBio-Science and Biotechnology 5.5 (2013): 241-266.

[8]. Sarwar, Abid, Vinod Sharma, and Rajeev Gupta."Hybrid ensemble learning technique for screening of cervical cancer using Papanicolaou smear image analysis." Personalized Medicine Universe 4 (2015): 54-62.

[9]. Chang, Victor, Yen-Hung Kuo, and Muthu Ramachandran. "Cloud computing adoption framework: A security framework for business clouds." Future Generation Computer Systems57 (2016): 24-41.

[10]. Goel, Rati. "Heart Disease Prediction Using Various Algorithms of Machine Learning." Proceedings of the International Conference on Innovative Computing & Communication (ICICC).2021.

[11]. Arunachalam, Subbiah, and Subbiah Gunasekaran. "Diabetes research in India and China today: from literature-based mapping to health-care policy." Current Science 82.9 (2002): 1086- 1097.

[12]. El-Khatib, Firas, et al. "Valproate, weight gain and carbohydrate craving: a gender study." Seizure16.3 (2007): 226-232.

[13]. Buchwald, Henry, et al. "Weight and type 2 diabetes after bariatric surgery: systematic review and meta-analysis." The American journal of medicine 122.3 (2009): 248-256.

[14]. Vijiyarani, S., and S. Sudha. "Disease prediction in data mining technique–a survey." International Journal of Computer Applications & Information Technology 2 (2013): 17-21.

[15]. Rish, Irina. "An empirical study of the naive Bayes classifier." IJCAI 2001 workshop on empirical methods in artificial intelligence. Vol. 3.No. 22. IBM, 2001.

[16]. Guo, Gongde, et al. "KNN model-based approach in classification." CoopIS/ DOA/ ODBASE. Vol. 2003. 2003.

[17]. Widodo, Achmad, and Bo-Suk Yang. "Support vector machine in machine condition monitoring and fault diagnosis." Mechanical systems and signal processing 21.6 (2007): 2560-2574.

[18]. Dietterich, Thomas G. "Ensemble methods in machine learning." Multiple classifier systems 1857 (2000): 1- 15.

[19]. Patel, Vimla L., et al. "The coming of age of artificial intelligence in medicine." Artificial intelligence in medicine 46.1 (2009): 5- 17