Volume 8, No. 4, May-June 2017
ISSN No. 0976-5697
International Journal of Advanced Research in Computer Science
RESEARCH PAPER
Available Online at www.ijarcs.info

Artificial Intelligence based Ensemble Model for Diagnosis of Diabetes

Harsha Sethi
Student M.Tech, ECE,
RIEIT, Railmajra, Punjab, India

Anudeep Goraya
Associate Professor,
ECE, RIEIT
Railmajra, Punjab, India

Vinod Sharma
Professor,
Dept. of Computer Sc. & IT
University of Jammu, Jammu
J&K, India

Abstract: The work done in this paper exhibits an expert system based ensemble model in diagnosing diabetes. Diabetes Mellitus is a disease with high mortality rate that affects more than 60% population. The mindset of this task is to analyze various machine learning techniques for binary classification concerning with illness i.e. to diagnose whether a subject is suffering from disease or not. There are in total fifteen classifiers considered and out of them five major techniques namely: ANN, SVM, KNN, Naive Bayes and Ensemble are used. For achieving the desired goals, the tools that were employed namely matrix laboratory (MATLAB) and WEKA 3.6.13. In Ensemble method the predictive potentials of various individual classifiers are fused together. Using Ensemble method, it increases the performance by combining the classifying ability of individual classifiers and the chances of misclassifying a particular instance are reduced significantly, this provides a greater accuracy to the overall classification process. It is the enhancing technique that does the majority voting and gives us the percolated results. The medical database analysed in this study includes a rich database of about 400 people from across a wide geographical region and ten physiological attributes. Furthermore, this diagnostic tool is examined by verifying denary cross attestation; on top of that the outcome has been confronted along the truly existing real interpretation about the cases. A GUI based diagnostic tool founded upon ensemble classifier is developed in such a manner it would be able to predict whether a patient is enduring against the disease or not when it is fed with all the 10 attributes from user through a user friendly GUI (Graphical User Interface). Out of 10 parameters that the user needs to enter as input in GUI based diagnostic tool five are numeric and the rest are nominal values. The diagnostic tool in execution is demonstrated below in fig 3. The main objective of this manuscript is to propose an intelligent framework that will act as a useful aid for doctors for correct & timely biopsy can be done at early stage. The result indicated that ensemble technique assured an accuracy of 98.60% that clubs the predictive performance of multiple AI based algorithms and are superior in comparison with all other individual counterparts. The algorithms with better exactness than others are followed by Artificial neural network (ANN), Naïve Bayes, Support Vector Machine (SVM), K-Nearest Neighbor (*K*-NN).

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

1. INTRODUCTION

Diabetes is a generalized term used for heterogeneous disorders that affects the capability of body in order to utilize the energy source found in food.The global prevalence of diabetes in adults (20-79 years old) according to a report published in 2013 by the IDF was 8.3% (382 million people), with 14 million were more men than women [1] (198 million men *vs* 184 million women), the majority between the ages 40 and 59 years and the number is expected to rise beyond 592 million by 2035 with a 10.1% global prevalence. With 175 million cases still undiagnosed, the number of people currently suffering from diabetes exceeds half a billion. An additional 21 million women are diagnosed with hyperglycemia during pregnancy [1]. When a person takes its meal which include carbohydrates, that splits into the abdomen along with digestion in shape with simple sugar that itself acts as an essential power rise in human body and as well found in many carbohydrate products. Carbohydrates containing food are like starchy foods, sugary foods, fruits, milk and some diary. These carbohydrates are then converted in the form of glucose although our human body senses whether the sugar level is rising or not. During this process, the pancreas starts releasing insulin; that insulin helps human to get energy but if a person has diabetes; then the human body is not able to produce sufficient insulin [2]. Diabetes is a chronic condition that occurs only when the body cannot produce enough or cannot effectively use insulin. Diabetes can mainly be of 3 types: type-1 diabetes, type-2 diabetes and gestational diabetes. The frequent symptoms of diabetes are increased thirst, frequent urination, increased Hunger, visions, Fatigue, family history etc. Being a lifestyle disorder; diabetes is usually controllable by medicines, by managing the eating habits and by doing regular physical activity. The conventional way for diagnosis of diabetes is by pathological test of checking the glucose level in the blood. If the glucose level in the human body remains continuously above the normal range as it is required, it is confirmed that the person is suffering from this disease. Diabetes is an internationally recognized public health problem affecting nearly 60% of the world population.

Artificial intelligence (AI) is a subpart of computer science, concerned with, how to give computers the sophistication to act intelligently, and to do so in increasingly wider realms. It is the name of the academic field of study which studies how to create computers and computer software that are capable of exhibiting intelligent behaviour. It is usually defined as "*the study and design of intelligent agents*", in which an intelligent agent is a system that perceives its environment and takes actions that maximize its chances of success [3]. The field of AI research was founded at a

conference on the campus of Dartmouth College in 1956. The attendees, including John McCarthy, Allen Newell, Arthur Samuel, and Herbert Simron, became the leaders of AI research for many decades. They and their understudies composed projects that were astonishing to the vast majority of people: PCs were winning at checkers, taking care of word issues in polynomial math, demonstrating sensible hypotheses and communicating in English. The field was established on the claim that the main property of humans, Intelligence—can be mimicked by a machine. Artificial intelligence has now days turn into an integral part of the industry, providing most effective solutions for most difficult problems from all walks of daily life. The general aim of AI is simulating Human Intelligence. This aim can be broken down into a number of specific, sub-specific aims / domains; depending upon the particular traits or capabilities that researchers would like an intelligent system to display.

In this work we have developed an expert system called as "Diabetes Diagnoser" which is powered by four Artificial Intelligence based algorithms. These algorithms have been trained, tested and validated using a primary database of 400 people from different sections of the society. We have done a rigorous testing of this expert system using ten-fold cross validation and compared its results with the actual diagnosis of the patients. The results obtained indicate that diabetes diagnoser is proficient tool for prognosis of diabetes and can be used as an effective aid in primary level screening of diabetes. Among all the algorithms implemented for the problem under consideration, the artificial neural network outperformed. In addition to this, we have also proposed a hybrid ensemble technique that clubs the predictive performance of multiple artificial intelligence based algorithms and performs better than all other individual counterparts.

The "Diagnostic Tool" presented in this paper is basically a branch of applied artificial intelligence (AI), developed by the artificial intelligence (AI) community in the mid-1960s with an aim to transfer the expertise of a human into a computer. Characteristically, an expert system integrates an inference engine i.e. a set of rules in the form of a program for applying the knowledge obtained from a knowledge base that contains the accumulated experience. The expert systems these days are embedded with machine learning algorithms that allow them to learn from past experience just as is done by humans and thus improving their working efficiency with time [4].

2. RELATED WORK

Oguz karan et.al. designed a system named 'Diagnosing diabetes using neural networks on small mobile devices' [5] in which Pervasive is being used in improving healthcare. In this system a novel approach is used for diagnosing diabetes with the help of neural networks and pervasive healthcare computing technologies. In this paper, neural networks based classifiers are being used for primary screening of patient's disease for which communication is set between the patient's hand-held small mobile devices (PDA) and powerful desktop PC wireless network for real time use of pervasive healthcare services depending on the nature of illnesses. The client mobile application tries to make its ANN and other complex calculations locally and shows the results to the patient without contacting to the server for a range of illnesses including diabetes. T. Manju et al. designed "Heart Disease Prediction System Using Weight Optimized Neural Network" [6]. In this work, the multi-layer feed forward neural network (MLFFN) and genetic algorithm (GA) were used for assisting medical doctors in predicting the heart disease. The cardiac arrest (heart attack) is a major cause of death in the world, its major causes are smoking, high blood pressure, unhealthy diet, obesity and diabetes. The data set used in the study was collected from university of California at Irvine (UCI) repository and consists of data of 270 patients. The ANN is trained using back propagation and feed forward neural network. Weight optimization is done using genetic algorithm. The weights are associated with each connection in the neural network nodes. The accuracy of the system on training dataset came out to be 79.7% and on testing accuracy 89.67%. Babak Sokouti et.al. presented the work named "A framework for diagnosing cervical cancer disease based on feed forward MLP neural network and Thin Prep histopathological cell image features" [7]. In this work, Levenberg–Marquardt feed forward MLP neural network (LMFFNN) was being used in order to classify cervical cell images obtained from 100 patients including healthy, low-grade intraepithelial squamous lesion and high-grade intraepithelial squamous lesion cases. This neural network along with extracted cell image features is a new model for cervical cell image classification. The semi-automated cervical cancer diagnosis system is composed of two phases: image pre-processing /processing and feed forward MLP neural network. In the first stage, image pre-processing is done to reduce the existing noises without lowering the resolution. After that, image processing algorithms were applied to manually cropped cell images to achieve a linear plot which includes real components, were used as LMFFNN inputs for classification of cervical cell images. Based on the results, cervical cell images were classified successfully with 100 % correct classification rate using the proposed method. Moreover, the rates of sensitivity and specificity were calculated as 100 % using LMFFNN method. It was shown there was a good agreement between the expert decision and values gained from the ANN model. Yasodha et al. proposed a system called "Analysis of a population of diabetic patient's databases in WEKA tool"[8]. In this, the study concerning with a database regarding diabetes mellitus (DM) disease patients. The authors studied various machine learning based classifiers like REP Tree, Naïve Bayes, decorate as well as Multi class classifiers in order to study plus also matched with the desired outcome. Prime aim concerning with the research in order to create a Diagnostic tool based on artificial intelligence; inputs being person's day-to-day blood sugar levels as well as doses of insulin; in such a way, that this diagnostic tool can predict the patient's insulin dosage for the next day. Igor Kononenko proposed a system called "Machine learning for medical diagnosis: history, state of the art and perspective"[9] presented a view on the use of Machine learning techniques 1) in the past for the interpretation of medical data 2) For intelligent analysis of medical data in the current scenario and 3) for assistance of physicians in diagnosis of medical disorders, in the future. Integration of machine learning techniques with the existing instrumentations for the acceptance of machine learning in medicine is suggested by the authors. Kemal

Polat et.al. have proposed a system "An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease" [10] for diagnosis of diabetes using principal component analysis and adaptive neuro fuzzy inference system. The authors have used the dataset obtained from UCI machine learning data repository. The system works in two phases, in first phase feature selection algorithm is used to reduce the size of the data set for analysis from eight attributes to four attributes; In contrast with it, the other stage collected data base with four attributes is fed to neuro fuzzy inference system for artificial intelligence based diagnosis of diabetes. They have reported that their system has presented an efficiency of 89.47% which is better than the other systems reported in the literature. Chad Ton Su et al. have proposed a system "Data mining for the diagnosis of type II diabetes from three-dimensional body surface anthropometrical scanning data"[11] for prediction of diabetes type-II using anthropometrical body. They have used four data mining approaches including dignitary, logistic regression, neural networks and rough set for feature selection and reducing the size of data base for analysis. The results of their study have indicated height, weight, head circumference, volume of right arm, surface area of right arm In addition to these measurements, the subjects' age and gender are the factors /parameters that play key role in the manipulation of disease. They have also reported that the classification by dignitary and rough set is much better as compared to the classification of back propagation and logical regression. Mohammad Amine chikh have proposed a system "Diagnosis of diabetes diseases using an artificial immune recognition system2 (AIRS2) with fuzzy k-nearest neighbor"[12]. It is a modified version of Artificial Immune Recognition system-2 (AIRS-2) which they have called MAIRS-2. In this modified version they have replaced the KNN nearest algorithm with fuzzy k nearest neighbor algorithm with an aim to improve the diagnostic accuracy of diabetes. They have used the Pima Indian diabetes dataset obtained from UCI machine learning repository. They have evaluated the performance of MAIRS-2 using ten-fold cross validation on the performance parameters like sensitivity, accuracy and specific values. The authors have reported that the MAIRS-2 presented an accuracy of 89.10% as compared to AIRS-2 which presented an accuracy of 82.69%. T Jayalaxmi has proposed a system named "A novel classification method for diagnosis of diabetes mellitus using artificial neural networks." [13] stressed on management of missing values in various dataset used for analysis of medical disorders. She has also investigated the impact of data pre-processing techniques on various classification algorithms. She has used some pre-processing techniques on Pima Indian diabetes dataset and reported that the performance of algorithms has improved considerably after pre- processing. Seppo Puuronen et.al. designed an algorithm named "A dynamic integration algorithm for an ensemble of classifiers." [14] which can be used for ensemble of classifiers by integrating. As we know in data mining techniques, the selection of an algorithm can be performed by two methods viz. static and dynamic. The latter one is done by taking new instances which usually results in higher accuracy. Thus by further enhancing it a dynamic integration algorithm is being used for the ensembling of classifiers. In this paper, results states that by using three machine learning dataset and by using a dynamically integrated algorithm have achieved higher accuracy as compared to other counterparts. Humar kahramanli et.al have designed a system called "Design of a hybrid system for the diabetes and heart diseases" [15]. In this paper, data were obtained from the University of California at Irvine (UCI) machine learning repository. To enrich the presentation done in this paper, many measures has been taken under consideration which serves the clinical class of society. By using the suggested techniques, presented in this paper accuracy of dataset was achieved through N-cross validation; According to Pima Indian Diabetes Dataset and Cleveland Heart Disease Dataset published in UCI websites, proposed method achieve accuracy values 84.24% and 86.8% respectively [15]. Nahla Barakat et.al suggested "Intelligible support vector machines for diagnosis of diabetes mellitus"[16]. In this system type II diabetes typically accounts for more than 90-95% after the age of 35 but in the case of type I diabetes, it has been observed that, it is more prevalent in children than in adults. It has been observed that the people whose sibling or parents has/had suffered from diabetes. In this support vector machines (SVM) is used which turns into 'black box' model of an SVM into an intelligible representation of SVM diagnostic. Results on real- life dataset show that intelligible SVM provide a promising tool for prediction of diabetes, with prediction accuracy of 94%, sensitivity of 93% and specificity of 94% [16].

3. MATERIAL & METHODS

3.1 Database used for study

In order to perform the research reported in this manuscript, we first studied the medical literature of diabetes and the related research work being carried out in this domain [17]. We also consulted the medical experts of the concerned domain and discussed with them the problem under consideration. After a detailed conversation, we identified 10 attributes that perform a pivotal task during manipulation connected with diabetes mellitus. The attributes selected are illustrated in table 1, which also presents an interval-frequency data about the various parameter values in the database. On the basis of these parameters authors prepared a rich database of about 400 people from across a wide geographical region. During the preparation of database, it was ensured to have diversity and variety in the database in terms of all the parameters under consideration. The database included both categories of people i.e. diabetic and non-diabetic. To populate the database with variety of instances we collected data from different sections of society i.e. people residing in urban areas, rural areas, upper class, lower class, people from different age groups, people with different eating habits, cultures, smokers, non-smokers, drinkers, non-drinkers etc. In this database, least age of the person considered is 5 years to the top age of the person considered is 78 years. While constructing a databank care was taken to assign discrete values used for analysing the data in order to hold the uniformity in the record taken. Fig. 1 depicts about the dataset used for study.

Physiological 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 smoker or 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 person drinks	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 mass for 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 kg Overweight-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

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 1: Screen shot of a sample of database used for training of Algorithms

To carry out this study which is presented therein paper, we adapted the information comprises of 400 different people from community in such a manner that a breed of evidences possibly be assured. Data set we chosen involves 10 physiological parameters which descripts indispensible role in the declaration of diabetes. The values contain in data is taken after the detailed discussion with expert consultants we come to an end with this statistical analysis given above.

Among the various parameters selected for study; Age has a pivotal role to play. According to analysis type II diabetes typically accounts for more than 90-95% after the age of 35 but in the case of type I diabetes, it has been observed that, it is more prevalent in children than in adults. It has been observed that the people whose sibling or parents has/had suffered from diabetes, are 30-35 % times more susceptible to diabetes than a person without such family history. People who suffer from diabetes are not able to manage and maintain the levels of the sugar in their blood, as such usually report an increased desire for thirst. There is also a change in the body weight of the persons suffering from diabetes [18]. According to a report of the American diabetes association around 50% of men and 70% of women become obese at onset of diabetes after bariatric surgery [19]. It has been observed that women are more at risk of developing diabetes as compared to men; this is attributed to usually lesser amount of the physical activity in women than in men. It has been observed that people who do not smoke or use tobacco are at lower risk as compared to the other counterparts. Many studies have reported that people who moderated intakes liquor are associated along with inflated susceptibility towards insulin's and lowers down this hazard of developing diabetes mellitus. India being one of the countries sharing a highest amount world's diabetic population, so the collection of data from the subjects for study was easy.

3.2 Ensemble technique

The ensemble classification technique works by constructing a large number of classifiers at training time and producing the class that is the mode of the classes output by individual classifiers [20]. The ensemble method proposed in this work presents a commendable level of accuracy, better than all the individual classifiers. It is an enhancing technique applied to the results of different algorithms to achieve better accuracy in which the final classification is better than individual classifiers. By using this method, the likelihood of misclassifying a particular instance are reduced considerably, and same process is repeated for all instances, final decision class for an instance is obtained by considering the maximum voting by putting the algorithms results in MS EXCEL and then majority voting is so constructed. The methods used in this work are K-Nearest Neighbour, Naive Bayes, Artificial Neural Networks, and Support vector machine.

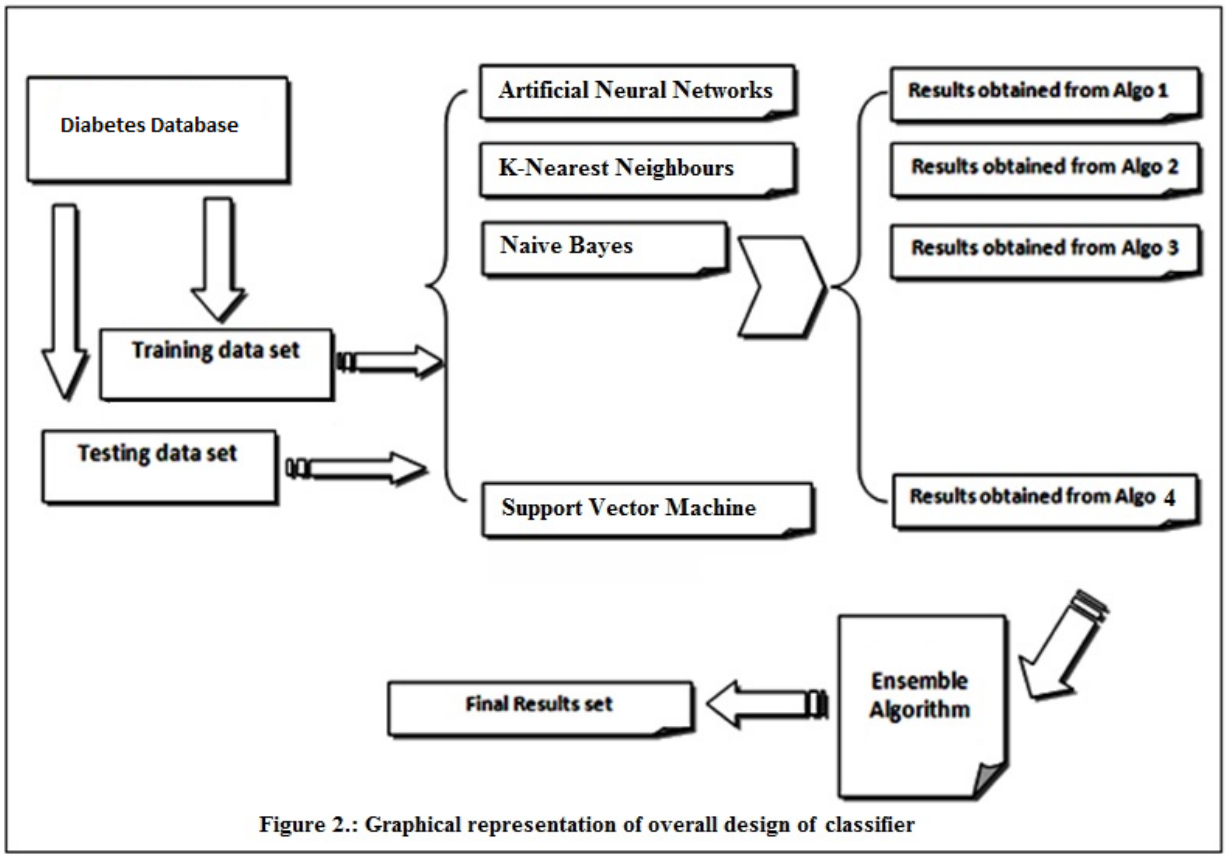

Figure 2.: Graphical representation of overall design of classifier

Fig. 2 shows the overall design of ensemble classifier. The primary database of diabetes was divided into two parts viz. Training dataset and testing dataset in the ratio of 80:20. Using the training dataset the four contributing algorithms were trained. Once the algorithms were trained with considerable level of accuracy, there performance was evaluated using the testing dataset. The testing dataset contains the instances which were unseen to the algorithms. For evaluating new cases, the test case is subjected to classification by all the four algorithms and the results from all the four algorithms are obtained. This is followed by subjecting the results to pass from an ensemble function which obtains the votes from all the classifiers and classifies the test instance into the class having majority votes.

3.3 Algorithms used for study

For designing the diabetes diagnoser we have used four algorithms namely Artificial Neural Networks, K-nearest neighbor, Naïve Bayes and support vector machine. These algorithms have achieved remarkably better accuracy as compared to other classifiers as well as popular among related literature work in medical domain [21]. A brief narration of these selected algorithms is given under.

3.3.1 K-nearest neighbor

K nearest neighbor is one of those algorithms that are very simple to understand as well they works astonishing in practice. In this algorithm the units are located next to each other and respond to input vector. For one or two dimensional it is easy to visualize the data using Self organizing maps that groups the input data into clusters. KNN can be used for both classification and regression predictive problems like if given N training vectors, K-NN algorithm identifies the K nearest neighbors of the test data regardless of labels, and sorts trial evidences hooked among possible classes by taking the votes from all the k-nearest neighbors. It has been observed that K-NN is being increasingly used in estimating statistical data and recognizing the data patterns. K-NN is termed as lazy and non-parametric learner [22] because it only stores the trained database and no general model from training dataset is constructed as is in the case of eager learners like decision trees.

3.3.2 Naïve Bayes

Naive Bayes algorithm is based on three concepts- antecedent, feasibility and prediction where antecedent means some past information about the incident, feasibility means chances of that event to occur in future and prediction means some forecasting made about the occurrence of that event on basis of first two concepts. The relation between the three variables is given by: -

Prediction = Antecedent * Feasibility / Corroboration

Mathematically the above relation can be represented as: -

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

Naïve Bayes classifier greatly simplify learning by assuming that features are independent given class [23]. These classifiers are statistical classifiers which use Bayes theorem as underlying principle. They are based on an assumption about mutual independence of attributes and this assumption is far from being true and this is the reason why this method is called naive. According to Bayes theorem the probability of a hypothesis H can be calculated on the basis of the hypothesis H and evidence about the hypothesis X according to the following formula:

$$P\ (H|X)\ = \frac{P\ (X|H)\ P\ (H)}{P(X)}$$

P (H): Prior probability of hypothesis H.
P(X): Prior probability of training data X.
P (H/X): Probability of H given X.
P(X/H): Probability of X given H.

These probabilities may be estimated from the given data. Naïve Bayes classifier outclassed among numerous other classifiers in terms of performance, reliability, correctly classified instances and accuracy. Bayesian method based on supervised learning includes two phases: namely learning and testing phase. In learning phase estimation is made from the attributes applied to it which maintains the record of these attributes and classifies its features. The other is testing phase, in which when a new test data is tested prediction is being made on the basis of learning phase and probability is calculated for the desired outcome. These attributes provide a self- sufficient benefaction in the prodigy of concluding outcome.

3.3.3 Artificial neural networks

Artificial neural network is a machine learning method inspired and evolved from the structure, function and working of human brain. It is a powerful data-modeling tool capable of capturing, representing and simulating complex relationship between inputs and outputs by performing multiple parallel computations. These are analytical tools which try to emulate ‘‘learning’’ process of the cognitive system and the neurobiological functions of the human brain. Learning is achieved by repeatedly adjusting the numerical weights associated with the interconnecting edges between different artificial neurons. In addition to this, an activation function is used that converts a neuron’s weighted input to its output activation where output regarding neural network relies on the cooperation of the individual neurons within the network to operate. These processing elements of a neural network are organized into a sequence of layers with full or random connections between various layers.

Mathematically, Let I= (I_1, I_2..... I_n) represent the set of inputs presented to the unit U. Each input has an associated weight that represents the strength of that particular connection. Let W= (W1, W_2..... W_n) represent the weight vector corresponding to the input vector X. By applying to V, these weighted inputs produce a net sum at U given by S = SUM (W_I *I_N).

These processing elements are usually organized into a sequence of layers with full or random connections between various layers. In neural network, an activity pattern (input vector) is applied to sensory nodes of the network and its effect propagates through the network layer by layer. Finally, a set of outputs is produced as the actual response of the network. During the forward pass, the synaptic weights of the networks are all fixed. In the backward pass, on the other hand, the synaptic weights are all adjusted in accordance with an error-correction rule. Specifically, the actual response of the network is subtracted from desired (target) response to produce an error signal. The synaptic weights are adjusted to make the actual response of the network move closer to the desired response in a statistical sense [24]. Input layer is not the neural computing layer because the node doesn’t have the input weights and also they don’t have any activation function. The top layer is the output layer that presents the response for the input fed to the network. The other layers are called the hidden or intermediate layers as they don’t have any connection with the outside world.

3.3.4 Support Vector Machine

Support vector machine (SVM) is a supervised machine learning algorithm and is used to perform both classification and regression. These classifiers are based on structural risk minimization principal and statistical learning theory with

an aim of determining the hyperplane (decision boundaries) that produce the efficient separation of classes. The underlying algorithm is Support Vector Classification (SVC) and it revolves around the perception of a "margin"- either side of a hyperplane that divides two data classes. Maximizing the margin creates the largest possible distance among the hyperplane and the instances on either side of the hyperplane reduce an upper bound on the anticipated generalization error. It works on two types of data i.e. linearly separable data and linearly Non-separable data. In case of linearly separable data only one hyperplane is needed for separating the data but in the case of latter more than one hyperplanes are needed. It uses kernel trick approach, implicitly mapping their inputs into high-dimensional feature spaces. Although SVMs handle non-linear decision boundaries of arbitrary complexity, we limit ourselves, in this paper, to the linearity of SVM. This paper presents a survey of machine condition monitoring and fault diagnosis using support vector machine (SVM). It attempts to summarize the review of recent research and developments of SVM in machine condition monitoring and diagnosis. SVM has excellent performance in generalization so it can produce high accuracy in classification for machine condition monitoring and diagnosis. Until 2006, the use of SVM in machine condition monitoring and fault diagnosis is tending to develop towards expertise orientation and problem-oriented domain. Finally, the ability to continually change and obtain a novel idea for machine condition monitoring and fault diagnosis using SVM will be future works [25].

4. DEVELOPMENT OF DIABETES DIAGNOSER AND COMPARISON WITH OTHER ALGORITHMS

For realization of algorithms used for the development of Diabetes diagnoser, authors wrote program in MATLAB (Matrix Laboratory) in such a way that all the four algorithms work on a centralized primary database prepared for study. The interface of the diabetes diagnoser has been so developed, as it is easy and intuitive to use. The interface of diabetes diagnoser contains multiple drop down menus and text boxes for facilitating the entry of ten physiological parameters that have been considered for the study. There is a sub-module for conversion of the height of the subject from foot to centimeter. The text data fed by the user are converted into corresponding coded integer values for all the possible values in the domain. Once all the values are fed in the interface of the expert system, they are converted into a matrices-vector representing the subject in physiological characteristics. This vector is fed to the trained algorithms that work as the backbone of the expert system for the prognosis of diabetes. The diabetes diagnoser presents the result to the user in terms of Yes / No in the interface of the system. Figure 3 shows the screenshot of the matlab program of diabetes diagnoser in execution.

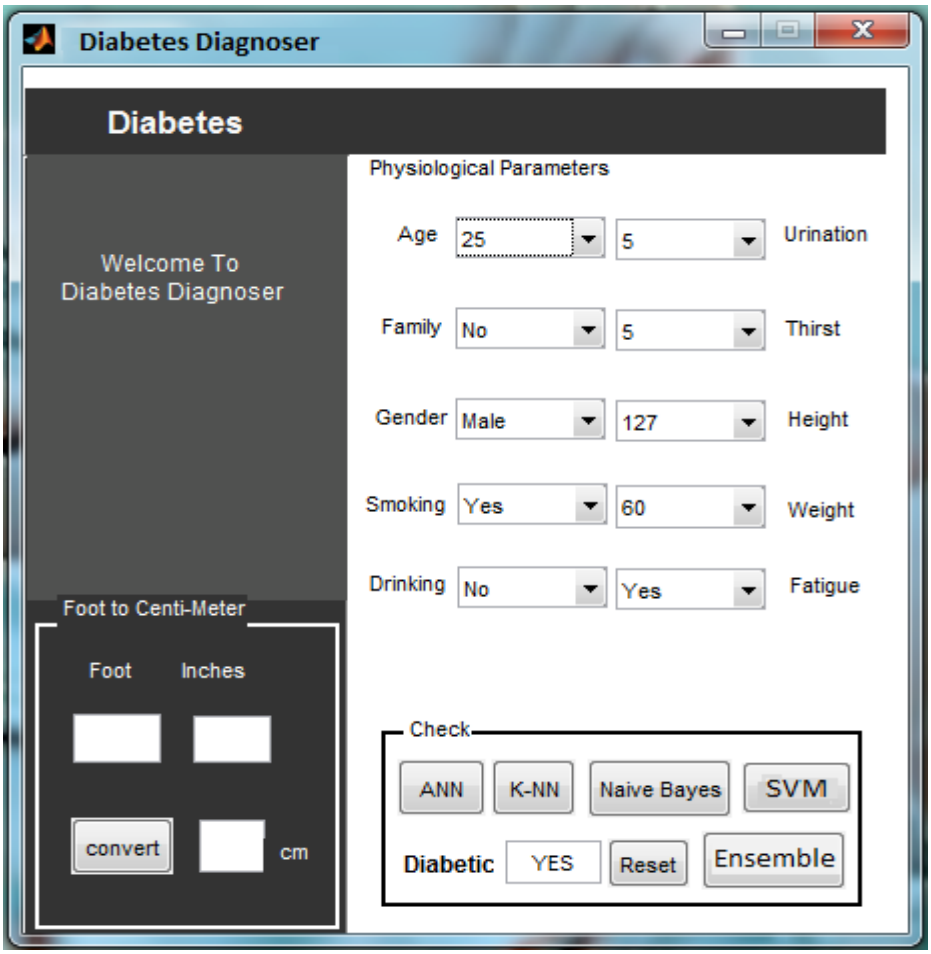

Fig 3. Diabetes diagnoser in execution

Authors have also compared the working efficiency of the Diabetes Diagnoser with ten other well known classifiers in the domain of artificial intelligence. For realization of machine learning based classifiers Waikato Environment for knowledge analysis (Weka) suite is used. It is an application written in Java & used machine learning community for

various research goals. The tool facilitates the application of expert systems supported classifiers in order to analyze data and designing of various predictive models as solution for some real world problems. The dataset was first converted into the necessary stable version i.e. an ASCII text file, for feeding into the WEKA workbench. For training and testing connected with numerous algorithms the dataset was divided in 80:20 ratio (311 instances for training and 89 instances for testing). The final results were evaluated on the basis of tenfold cross validation.

5. RESULTS AND DISCUSSION

In this paper authors have presented an artificial intelligence based ensemble tool for diagnosis of diabetes and compared its efficiency with multiple machine learning techniques. The efficiency of the system has been evaluated using tenfold cross validation technique and the results have been calculated using percentage of correctly classified cases and percentage of correctly incorrectly classified cases. The expert system presented in this work can be used as a tool for initial screening of people who suffer from this disease and as such can be an effective aid for mitigating the mortality due to this disease. Results indicate that the proposed 'Diagnostic tool' is competent during the diagnosis of patients suffering with disease named Diabetes Mellitus. The result indicated that ensemble technique assured an accuracy of 98.60%. The results obtained are illustrated in table 2.

Algorithms used	Correctly classified	Incorrectly Classified
Ensemble based Diabetes Diagnoser	98.60%	1.04%
Artificial Neural Network	96.00%	4.00%
Naïve Bayes	95.00%	5.00%
Support Vector Machine	94.00%	6.00%
J48 Graft	91.49%	8.50%
K- nearest neighbor	91.23%	8.77%
Decorate	91.23%	8.76%
END	91.23%	8.76%
Random forest	90.97%	9.02%
Bagging	89.69%	10.30%
Multi class classifier	89.69%	10.30%
Decision stump	88.65%	11.34%
Multi boost	88.65%	11.34%
User Classifier	88.65%	11.34%
Random Tree	88.40%	11.59%

Table 2: Results obtained from various algorithms in terms of performance metrics

Among all the algorithms, the result indicated that ensemble technique assured an accuracy of 98.60% that clubs the predictive performance of multiple artificial intelligence based algorithms and performs better than all other individual counterparts followed by Artificial Neural Network (ANN), Naïve Bayes, Support Vector Machine (SVM), K-Nearest Neighbor (*K*-NN), J-48 Graft, END, Decorate, Random forest, Bagging etc.The working efficiencies of these algorithms can be further increased by increasing the number of instances in the database and by including various clinical features that play some role in diagnosis of diabetes.

6. REFERENCES

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

[2] Olokoba, Abdulfatai B., Olusegun A. Obateru, and Lateefat B. Olokoba. "Type 2 diabetes mellitus: a review of current trends." Oman Med J 27.4 (2012): 269-273.

[3] Deepa, S. N., and B. Aruna Devi. "A survey on artificial intelligence approaches for medical image classification." Indian Journal of Science and Technology 4.11 (2011): 1583-1595.

[4] Michalski, Ryszard S., Jaime G. Carbonell, and Tom M. Mitchell, eds. Machine learning: An artificial intelligence approach. Springer Science & Business Media, 2013.

[5] Karan, Oğuz, et al. "Diagnosing diabetes using neural networks on small mobile devices." Expert Systems with Applications 39.1 (2012): 54-60.

[6] Manju, T., K. Priya, and R. Chitra. "Heart Disease Prediction System Using Weight Optimized neural Network." International Journal Of Computer Science and Management Research 2.5 (2013).

[7] Sokouti, Babak, Siamak Haghipour, and Ali Dastranj Tabrizi. "A framework for diagnosing cervical cancer disease based on feedforward MLP neural network and ThinPrep histopathological cell image features." Neural Computing and Applications 24.1 (2014): 221-232.

[8] Yasodha, P., and M. Kannan. "Analysis of a population of diabetic patients databases in WEKA tool." International Journal of Scientific & Engineering Research 2.5 (2011).

[9] Kononenko, Igor. "Machine learning for medical diagnosis: history, state of the art and perspective." Artificial Intelligence in medicine 23.1 (2001): 89-109.

[10] Polat, Kemal, and Salih Güneş. "An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease." Digital Signal Processing 17.4 (2007): 702-710.

[11] Su, Chad-Ton, et al. "Data mining for the diagnosis of type II diabetes from three-dimensional body surface anthropometrical scanning data." Computers & Mathematics with Applications 51.6 (2006): 1075-1092.

[12] Chikh, Mohamed Amine, Meryem Saidi, and Nesma Settouti. "Diagnosis of diabetes diseases using an artificial immune recognition system2 (AIRS2) with fuzzy k-nearest

neighbor." Journal of medical systems 36.5 (2012): 2721-2729.

[13] Jayalakshmi, T., and A. Santhakumaran. "A novel classification method for diagnosis of diabetes mellitus using artificial neural networks." Data Storage and Data Engineering (DSDE), 2010 International Conference on. IEEE, 2010.

[14] Puuronen, Seppo, Vagan Terziyan, and Alexey Tsymbal. "A dynamic integration algorithm for an ensemble of classifiers." Foundations of Intelligent Systems (1999): 592-600.

[15] Kahramanli, Humar, and Novruz Allahverdi. "Design of a hybrid system for the diabetes and heart diseases." Expert systems with applications 35.1 (2008): 82-89.

[16] Barakat, Nahla, Andrew P. Bradley, and Mohamed Nabil H. Barakat. "Intelligible support vector machines for diagnosis of diabetes mellitus." IEEE transactions on information technology in biomedicine 14.4 (2010): 1114-1120.

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

[18] El-Khatib, Firas, et al. "Valproate, weight gain and carbohydrate craving: a gender study." Seizure 16.3 (2007): 226-232.

[19] 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.

[20] Dietterich, Thomas G. "Ensemble methods in machine learning." International workshop on multiple classifier systems. Springer Berlin Heidelberg, 2000.

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

[22] Guo, Gongde, et al. "KNN model-based approach in classification." OTM Confederated International Conferences" On the Move to Meaningful Internet Systems". Springer Berlin Heidelberg, 2003.

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

[24] Dao, Vu NP, and V. R. Vemuri. "A performance comparison of different back propagation neural networks methods in computer network intrusion detection." Differential equations and dynamical systems 10.1&2 (2002): 201-214.

[25] 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.