

SURVEY PAPER

Open Access



# Survey on clinical prediction models for diabetes prediction

N. Jayanthi<sup>1,3\*</sup> , B. Vijaya Babu<sup>1</sup> and N. Sambasiva Rao<sup>2</sup>

\*Correspondence:  
jneelampalli.phd@gmail.com  
<sup>3</sup> Department of CSE,  
Institute of Aeronautical  
Engineering, Dundigal,  
Hyderabad, Telangana, India  
Full list of author information  
is available at the end of the  
article

## Abstract

Predictive analytics has gained a lot of reputation in the emerging technology Big data. Predictive analytics is an advanced form of analytics. Predictive analytics goes beyond data mining. A huge amount of medical data is available today regarding the disease, their symptoms, reasons for illness, and their effects on health. But this data is not analysed properly to predict or to study a disease. The aim of this paper is to give a detailed version of predictive models from base to state-of-art, describing various types of predictive models, steps to develop a predictive model, their applications in health care in a broader way and particularly in diabetes.

**Keywords:** Predictive analytics, Diabetes, Clinical prediction models, Traditional model, Hybrid model, Machine learning

## Introduction

Predictive analytics use statistical or machine learning method to make a prediction about future or unknown outcomes [1]. It uses text mining for unstructured data, answers the question “what is next step?” It uses historical and present data to predict future regarding activity, behaviour and trends. To do this it makes use of statistical analysis techniques, analytical queries and automated machine learning algorithms. Predictive analytics need experts to build predictive models. These models are used for prediction.

There are many applications of predictive analytics, out of which one is health care. A most common disease now a day's is diabetes. People are suffering with it and the patient number increases day by day. The World Health Organization (WHO) predicts that by 2030 there will be approximately 350 million people worldwide affected by diabetes [2, 3]. Mostly whatever food we eat is converted into glucose or sugar. Now, this glucose or sugar is used for energy. Glucose is transported to body cells through insulin. If the body does not produce sufficient insulin or does not make proper use of insulin then it leads to diabetes.

There are four types of diabetes which are TYPE 1, TYPE 2, GESTATIONAL, PRE DIABETES. TYPE 1 diabetes is also known as insulin dependent diabetes [4] where the pancreas does not produce the hormone insulin. TYPE 2 diabetes is also known as non-insulin dependent diabetes [4] where adequate insulin is produced but the body cannot make use of insulin. Gestation diabetes is a type of diabetes which occurs during pregnancy [5]. Pre diabetes refers to a situation where blood glucose levels are higher than normal but not so high to diagnosis as diabetes [6]. Diabetes is a disease in which

blindness, nerve damage, blood vessel damage, kidney disease and heart disease can be developed [7]. By the use of predictive analytics in the field of diabetes, diabetes diagnosis, diabetes prediction, diabetes self-management and diabetes prevention can be achieved as per the literature survey.

Future the paper is organized into three sections. “Related work” gives the background of predictive analytics. “Clinical prediction model” describes different predictive models used in health care particularly for diabetes, followed by the summary of predictive models and ends with a conclusion and future work.

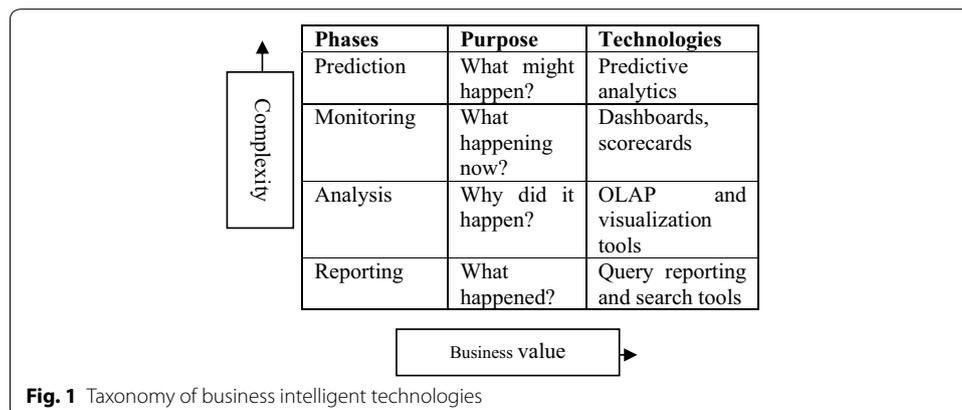
**Related work**

**Predictive analytics**

As per literature, there are many forms of describing predictive analytics. It is inductive [8]. It doesn’t expect anything about data but it allows the data lead the way. It uses statistics, machine learning, neural computing, robotics, computational mathematics and artificial intelligence to explore all data and find meaningful relationships and patterns. Predictive analytics is a set of business intelligence (BI) technologies that uncover relationships and patterns within large volumes of data that can be used to predict behaviour and events. To be more clear see the Fig. 1. Machine learning used by predictive analytics is a technique to train algorithm which can predict an output based on some input value. This leads to correlations and not to conclusions [9].

**Taxonomy of predictive analytics**

From previous work, there are two major types of predictive analytics such as supervised learning and unsupervised learning [8, 10]. Supervised learning is a process of creating predictive models using a set of historical data and produce predictive results. Examples are classification, regression and time-series analysis where as in Unsupervised learning does not use the previously known result to train its models. It uses descriptive statistics. It identifies clusters or groups [8]. Further classification of predictive models are of nine types business rules, classification and decision trees [11–13] naive Bayes, linear regression [10–12], logistic regression [11, 12, 14], neural networks (NNs) [11–13], machine learning, support vector machines (SVMs), natural language processing (NLP) [11]. In paper [13] the author has described seven types of regression models, each holding importance of its own. He listed seven types of regression models as linear regression,



**Fig. 1** Taxonomy of business intelligent technologies

logistic regression, polynomial regression, stepwise regression, ridge regression, Lasso regression, elastic net regression. More versions of predictive models are described in two ways Smooth Forecast Model which describe smooth variable outcome for example profit and Scoring Model which describe binary outcome for example whether the blood report indicates disease or normal condition [9]. Another list of predictive models is linear models, decision trees, neural networks, clusters models, support vector machines, expert systems. The difference between simple linear model and generalized linear model is presented in Table 1 [10]. Furthermore, a list of models are Time-series analysis, Monte Carlo simulations, Statistics for spatial data [15].

**Steps to develop predictive model**

It was listed that there are six steps for developing predictive models. Which are listed as follows project definition, exploration, data preparation, model building, deployment, model management [9].

Usage of predictive model by Organizations is summarized in paper [16] which is listed in Table 2.

A basic architecture about building predictive model consisting of three layers is presented in Fig. 2. What is important is how to implement model inputs and outputs, how model show growth when it is turned on or off, when to upgrade or replace according to ongoing time [11].

**Selecting model according to situation**

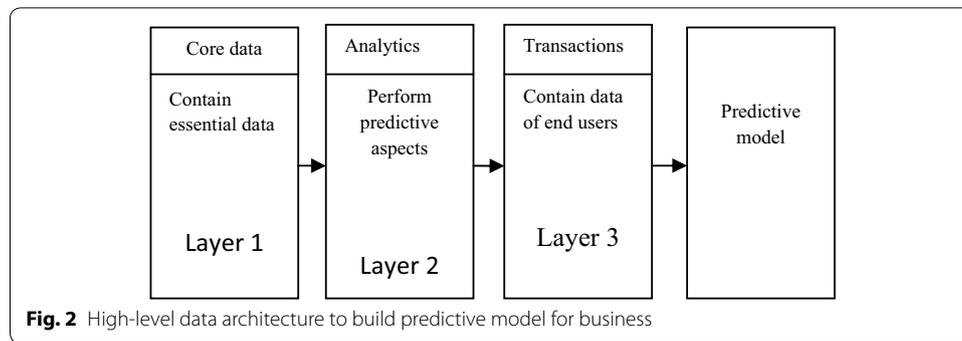
Depending on situation what model has to be selected is described as for segmentation use clustering algorithm, for developing recommender system use classification algorithm, Use decision tree when linear decision boundary is used, for predicting next

**Table 1 Difference between linear model and generalized model**

S. no	Simple linear model	Generalized linear models
1	$\mu = E(Y) = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \dots + \beta_n * X_n$	$g(\mu) = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \dots + \beta_n * X_n$
2	Target variable Y does not depend on the value of Y for any other record, only the predictors	Target variable Y does not depend on the value of Y for any other record, only the predictors
3	Y is normally distributed	Distribution of Y is a member of the exponential family of distributions(normal, Poisson, gamma, binomial, negative binomial, inverse Gaussian)
4	Mean of Y depends on the predictors, but all records have the same variance	Variance of Y is a function of the mean of Y
5	Y is related to predictors through simple linear function	$g(\mu)$ is linearly related to the predictors. The function $g$ is called the link function

**Table 2 Percentage of usage of prediction model in organizations**

Percent of model (%)	Model purpose
65	Use to guide decision and plans
52	To score records
41	Import models into BI tools or reports
36	Scores to create or augment rules
33	Embed rules or models in applications to automate or optimize processes



outcome of time driven events use regression algorithms, to predict continuous values use regression, use naïve Bayes when features are conditionally independent, Machine learning is used for classifying text problems with ensemble model sometimes [11].

**Deploying the predictive model**

Deployment means using a model for the intended purpose. Most predictive models are built depending on the decisions made by the organization [17]. Various ways for deploying the predictive model are sharing the model, score the model, incorporate the model in a BI report, embed the model in application [8].

**Assessment of predictive model**

Predictive models are commonly assessed using c-statistics. This measure indicates whether the predictive model predicts positive or negative outcomes. If model c-statistics exceeds 0.7 it is considered as an acceptable prediction, if c-statistics is 0.5 then model prediction is not good [18]. c-statistics is equal to the area under ROC curve. Grading based on c-statistics is shown as in Table 3. Another way to assess the prediction model is Analysis of variance (ANOVA) when data is categorical [11]. The measures are listed in Table 4.

**Table 3 Acceptance of predictive model based on c-statistics**

Range	Grade
0.9–1.0	Excellent
0.8–0.9	Good
0.7–0.8	Acceptable
0.6–0.7	Poor
0.5–0.6	Fail

**Table 4 Measures used model assessment when data is categorical**

Measure	Significance
R square	Project the efficiency of the model in terms of independent variables
Significance F	To check whether results are reliable or not if F less than 0.5 model is OK else stop using those independent variables
Coefficients	Regression line $Y = \text{intercept} + A * X1 + -B * X2$ A and B are coefficients these are useful for forecasting
Residuals	These are used to show how far the actual data points are from predicted data points

Some metrics that are considered for successful model are

1. Uplift from model
  - a. Compare the performance of predictive model against random results with lift charts and decline tables.
  - b. Evaluate the validity of the discovery with target shuffling.
  - c. Test predictive model consistency using bootstrap sampling.
2. Use empirical measures of accuracy such as confidence levels or other statistical quantities if the aim of models is to provide highly accurate predictions or decisions.

### **Applications of predictive analytics**

Predictive analytics have huge applications in various fields like Homeland security, Crime prevention, Infrastructure management, Cyber security, Intelligent Transportation, Health care and bioinformatics, Text mining, Fraud detection, Social media and decision support [1], Credit scores, Credit card, fraud detection, Mail Sorting, Weather prediction, Hot dogs and hamburgers [16], not only about resource allocation but also about where or how should to allocate resource, what to expect from outcomes of model, how to manage the key drivers of the economic model for better outcome? [19]. There are four ways to monetize predictive models. The first way is for saving cost. The second way is using a predictive model for improving revenue. The third way is for returning investment and lastly is for risk management [20].

### **Predictive modeling tools**

Predictive modeling tools [19] are summarized in Table 5.

### **Existing predictive models**

From paper [16] author have listed existing predictive models and their usage

1. Optum predictive model
 

This model is used to predict employees of the company, who are interested to join optum health management programs.
2. Netflix
 

This algorithm determines which movies a customer is likely to enjoy.
3. Match.com
 

This is a behaviour modeling algorithm as it learns from the behaviour of similar users and factors and recommends those sites to new users who for the same topic on the web.
4. Santa Cruz's predictive policing program
 

This algorithm will analyse and detect patterns of past years crime data and predicts areas and windows of time that are high at risk.
5. Harrahs casino
 

It predicts the pattern of gambling players based on their level of dissatisfaction and intervene.
6. Fighting medicare fraud

**Table 5 Predictive modeling tools**

Predictive modeling tools		
Risk groupers	These tools are best used for	Actuarial Underwriting Profiling perspectives
Statistical models	These tools require lots of historical data	Linear regression Logistic regression Anova Time series Trees Non-linear regression Survival analysis
Artificial intelligence models	These are new methods	Fuzzy logic Neural networks Genetic algorithm Nearest neighbor pairing Conjugate gradient Rule induction Principal component analysis Simulated annealing Kohonen network

Centers for Medicare and Medicaid Services (CMS) is poised to begin using predictive modeling technology to fight Medicare fraud. This algorithm automatic alerts and risk scores for claims.

#### 7. VISA

This algorithm predicts people behaviour.

#### 8. TCS diabetes Readmission predictive analytics model

This predictive model predicts patient readmission rates using a statistical model [21].

### Clinical prediction model

#### Clinical model

The clinical prediction model is very valuable as it can be applied to a various scenario like screening, prediction, medical decision making and education in health. In the medical field, the prognosis is considered as a fundamental component. In paper [22] a process of developing clinical prediction model is explained with five steps:

- i. Preparation for establishing clinical prediction models.
- ii. Dataset selection.
- iii. Handling variables.
- iv. Model generation.
- v. Model evaluation and validation.

Predictive analytics uses regression models on available data for predicting outcomes mostly in the medical field. In past medical data was collected manually by

handwritten, dictated or incomplete, this data was small for predictive modeling. But now a day because of EMRs, diverse and large digital data of health care is available as listed in Table 6. Natural processing languages are used to access unstructured data. This increases the quality of data and also quality prediction. The relative predictive power of a statistical model increases exponentially when using millions of patients instead of hundreds of patients. To discover relevant predictive variables clinical, claims, socioeconomic and care management data should be integrated to form one dataset [18].

#### Different prediction models used for diabetes

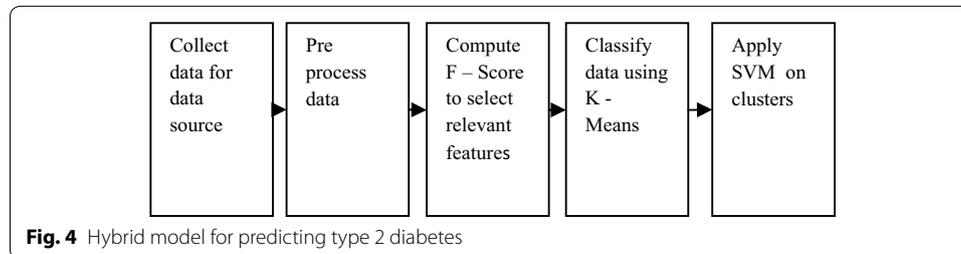
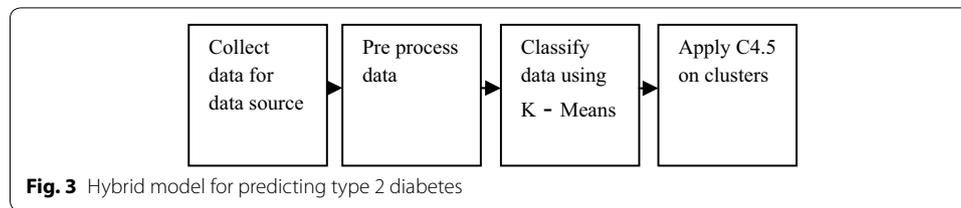
A multi stage adjustment model with low misclassification rate which predicts which persons are most likely to develop diabetes is built by using KoGES dataset [23]. A physiological model which can predict the blood glucose level 30 min in advance was developed using five patients data by training SVR with physiological features. This helped in producing best results than doctors [24]. Another type of predictive model is sparse factor graph model. By using which diabetes complications are not the only forecast but also can discover the underlying associations between diabetes complications and lab test types. All algorithms were implemented in C++, and all experiments were performed on a Mac running Mac OS X with Intel Core i7 2.66 GHz and 4 GB of memory. The data set used for the experiment is collected from a geriatric hospital. The data set contain 1-year span data with 181,933 medical records, 35,525 patients data and 1945 types of lab tests. 60% of data was chosen for training the model and the rest for testing. The proposed model addresses two challenges feature sparseness and knowledge skewness [25].

A hybrid model has been developed to predict whether the diagnosed patient may develop diabetes within 5 years or not. A tool used for this purpose is WEKA and the data set was PIMA Indian diabetes data set. This hybrid model has achieved 92.38% accuracy [26]. The details of the hybrid model are shown in Fig. 3.

Another hybrid prediction model helps in producing optimal feature subset. This helps in detecting diabetes with high accuracy. To implement the model WEKA tool is used on PIMA Indian diabetic dataset. The proposed models have given an accuracy of 98.9247%. The procedure adopted by authors in developing predictive model is first pre-processed the dataset, then compute F-score values of features, select features with high F-score as discriminative features, then k-means algorithm is used to select feature subset that gives minimum clustering error and finally SVM is used to classification [27] as shown in Fig. 4.

**Table 6 Predictive analytics in health now a day**

Model	Data		
	Size	Sources	Quality
Old	Limited data	Claims data Inpatient data only	Poor Unstructured Data can not be accessed
Morden	Large data	Emr + claims + socioeconomic + care management Inpatient + outpatient + ED	Excellent Unstructured data can be accessed



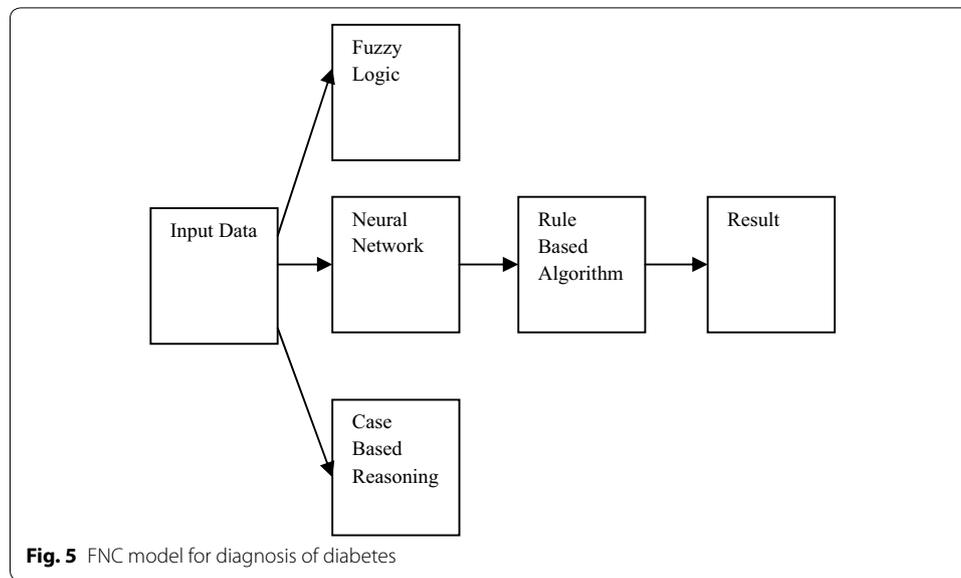
In paper [28] the authors used two different types of neural networks to express which will output the accurate classifier in predicting diabetes. The two neural network models are multilayer neural network and probabilistic neural network. The dataset contains Pima Indian diabetes, having two classes and 768 samples. 576 samples were used for training and 192 were used for testing. The proposed methods were proved to be better when compared with other previous methods.

In paper [29] the author developed a prediction model based on Hybrid-Twin Support Vector Machine (H-TSVM), which predicts whether a new patient is suffering from diabetes or not. They used Pima dataset for conducting an experiment. The factor that keeps this proposed method different from others is kernel function. The classifier produces an accuracy of 87.46%.

In paper [30] the author proposed a predicting model that classifies type 2 diabetic treatment plans into three groups such as insulin, diet and medication. The dataset used for developing the model was JABER ABN ABU ALIZ clinic centre which contains 318 medical records. The model was developed using WEKA tool by applying J48 classifier and it has produced an accuracy of 70.8%.

In paper [31] the author developed a prediction model which predicts what are different types of disease a diabetic patient can develop. To develop the model a data set of 3 years span is collected from AR hospital with 739 patient details and 31 attributes. The pre-processed data after deleting outliers by using distance-based outlier detection (DBOD), is given as input to a logistic regression model which was built by Bipolar Sigmoid Function that is calculated using Neuro-based Weight Activation function. The model produced a prediction accuracy of 90.4%.

In paper [32] a tool FNC was developed that can be used for diagnosing diabetes as shown in Fig. 5. The proposed model is developed by incorporating three techniques: fuzzy logic, neural network, and Case-based Reasoning with 200 patient details having 16 input attributes. Fuzzy logic and neural networks are implemented by using Matlab, Case-based Reasoning is implemented by using MyCBR plug-in. After obtaining the



result from three approaches rule based algorithm was applied to all three techniques to improve the accuracy. Finally, the best accuracy was obtained for case based reasoning.

In paper [33] author developed a hybrid model KSVM. The important criteria that make this model different from other methods are feature selection algorithm. PIMA data set was utilized to do experiments and results were produced. It was shown that diagnosis results using K-SVM are 99.74, 99.78, and 99.81 for learning experiments with amount 50, 60, and 70% data respectively, and 99.82, 99.85, and 99.90 for testing experiments with amount 50, 60, and 70% data respectively.

In paper [34] authors developed a prediction model that would predict whether a person would develop diabetes by considering daily lifestyle activities. To build prediction model PIMA diabetes data set was used and CART (Classification and Regression Trees) machine learning classifier was applied. The proposed model could provide an accuracy of 75%.

In paper [35] authors developed a prediction model that would predict whether a person develops diabetes or not. To achieve this PIMA diabetes dataset was used. In the proposed method first controlled binning technique is applied then multiple regression was used to improve the accuracy of the model. After incorporating all techniques an accuracy of 77.85% was achieved. The controlled binning technique which is innovative thought in this paper is calculated by using the Eqs. (1) and (2)

$$\text{Bin size} = (\text{Loss percentage}) * \text{total number of transactions} \quad (1)$$

$$\text{Loss on each data element} = (\text{loss\%} * \text{value of data element}) \quad (2)$$

In paper [36] authors developed a decision tree model for the diagnosis of type 2 diabetes. They used Pima Indian diabetes dataset. Pre-processing techniques like attributes identification and selection, handling missing values, and numerical discretization was used to improve the quality of data. Weka tool was used, J48 decision tree classifier

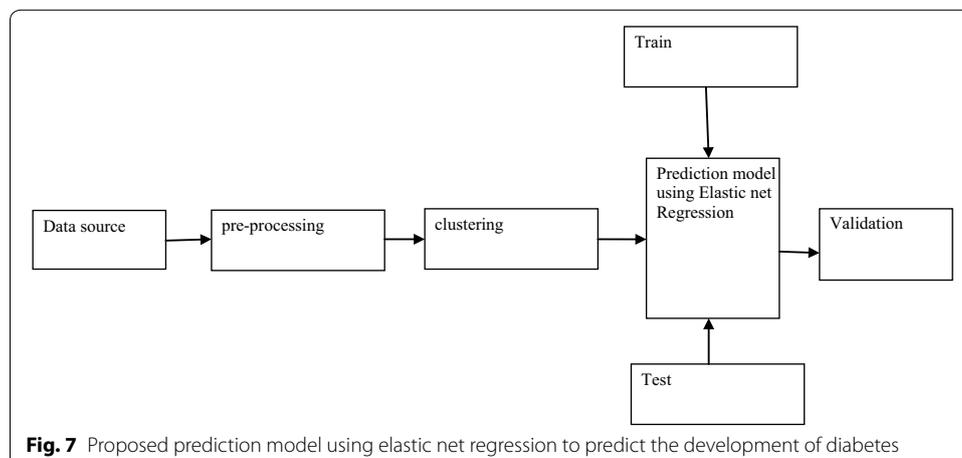
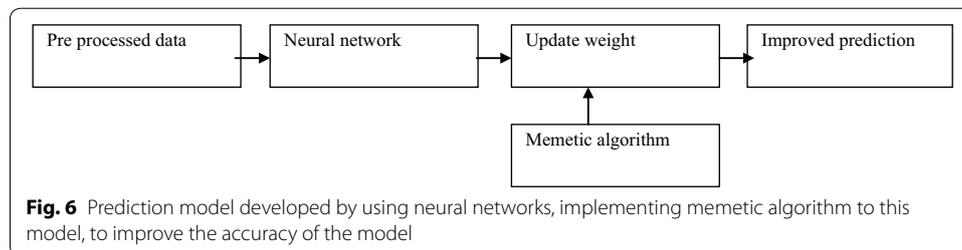
was applied to construct the decision tree model. The model produced an accuracy of 78.17%.

In paper [37] the authors developed a prediction model by using neural networks to classify and to diagnose onset and progression of diabetes. They have used 545 patients' data from a diabetes clinic. First, they trained and tested neural networks with a different number of neurons and found a neural network with seven neurons has produced highest accuracy. The memetic algorithm is used to update weights which improved the accuracy of the model from 88.0 to 93.2%. this model was compared with other models too. But a neural network with seven neurons and application of memetic algorithm is observed as the best model (Figs. 6, 7).

In paper [38] the authors have developed an expert healthcare predictive decision support system that predicts diabetes. This model is trained on Pima diabetes dataset. Decision tree and K-nearest neighbor algorithms are used to develop the model and found that C4.5 algorithm has achieved 90.43% accuracy.

In paper [39] the authors have developed a prediction model using Chi squared test to find not only dependencies between factors but also independences. Then CART is applied to build a prediction model which has 75% accuracy. Data was collected through questioners from 200 people and model was built using R tool.

In paper [40] authors developed an elastic net model which improves the accuracy for estimating glucose. The authors have collected 45 experimental sessions data set from diabetic patients. The data was collected from a noninvasive glucose device i.e., a blood sample is not taken. Three models were constructed using regularized methods LASSO,



**Table 7 Summary of different prediction models used for diabetes**

Paper no	Dataset	Prediction model	Technique	Tool	Outcome	Accuracy
11	Koges	Multi stage adjustment model	Not mentioned	Not mentioned	Which person is most likely to develop diabetes	Not mentioned
12	Five patients data	Physiological model	Svr	Not mentioned	Predicts blood glucose level 30 min in advance	Not mentioned
17	Geriatric Hospital	Sparse factor graph model	Not mentioned	Not mentioned	Forecast diabetes complications and uncover underlying relationship between diabetes and lab reports	Not mentioned
18	Pima	Hybrid model to predict	Clustering + C4.5	Weka	Predict whether the diagnosed patient may develop diabetes within 5 years or not	92.38%
19	Pima	Hybrid prediction model	Clustering + SVM	Weka	Optimal feature subset which helps in detecting diabetes with high accuracy	98.9247%
20	Pima	Neural networks	Multilayer neural network and probabilistic neural network	Not mentioned	Output the accurate classifier in predicting diabetes	
21	Pima	Hybrid-twin support vector machine	Kernel functions	Not mentioned	Predicts whether a new patient is suffering from diabetes or not	87.46%
22	Jaber Abn Abu Aliz	Prediction model	J48 classifier	Weka	Classify type 2 diabetic treatment plans	70.8%
23	Ar Hospital	Logistic regression model	Bipolar sigmoid function that is calculated using neuro based weight activation function	Not mentioned	Predicts what are different types of disease a diabetic patient can develop	90.4%
24	Not mentioned	Fnc model	Fuzzy logic, neural network, case based reasoning, rule based algorithm	Matlab and Mycbr plug-in	Used for diabetes diagnosing	Not mentioned
25	Pima	Ksvm	Feature selection algorithm	Not mentioned	Used for diabetes diagnosing	99.82–50, 99.85–60, and 99.90–70% of data
26	Manual collection	Cart		Manual	Used to predict whether a person would develop diabetes or not	75%

**Table 7 continued**

Paper no	Dataset	Prediction model	Technique	Tool	Outcome	Accuracy
27	Pima	Correlation analysis	Multiple regression	Manual	Predicts whether patient develops diabetes or not	77.85%
36	Pima	CART	J48	weka	Predicts whether patient develops diabetes or not	78.17%
37	Manual	Neural networks	Memetic algorithm	Not mentioned	Classify and diagnose onset and progression of diabetes	93.2%
38	Pima	Prediction model	C4.5 and KNN	Not mentioned	Predicts diabetes or not	93.43%
39	Questioner	Prediction model	CART	R	Predicts whether a person fall into diabetic in future	75%
40	Manual	Prediction model	LASSO, ridge and elastic net regressions	R	Predicts glucose level accurately	Not mentioned
Current work	Pima	Prediction model	Elastic net regression	R	Predicts whether a person develops diabetes or not with in 6 months	To be worked out

Ridged and Elastic net model. The elastic net model has compared with LASSO, ridged and partial least square regression and found Elastic net model is best.

From all of the techniques and prediction models discussed above, we want a prediction model that predicts diabetes of a diagnosed person. Since this output can be obtained depending on the time we would lie to use regression model. Of all regressions, Elastic Net is most useful as categorical, numerical and image or signal form data can be given as input to the model. The elastic net regression model is a combination of LASSO (Least Absolute Shrinkage And Selection Operator) and Ridged Regressions. Thus elastic net regression support shrinkage of coefficients as well as grouping effect. One more interesting point is numerical, categorical and image form data can be given as input to the model.

### Summary

The survey below summarizes the previously developed systems. Different datasets, tools, techniques used previously are listed in Table 7.

### Conclusions

In this paper a detail description of predictive modeling is presented, a combination of tradition and hybrid prediction models Modeling, This paper showed that hybrid models produce more accuracy than traditional models. A researcher who is willing to do research in developing clinical prediction model would be benefited by this paper. There is a wide range of scope for the development of clinical prediction models especially for diabetes as this is a modern disease in developing countries like India.

As per the survey of above papers we can find many gaps that are to be filled, which are usage of larger dataset [23, 34], outlier detection [35], improving prediction model [34], integration of optimization techniques to hybrid prediction model [33], implementation of prediction models for other diseases on android mobile [31], development of prediction model that include type 1 treatment plans with more attributes [30], usage of datasets of multiple classes [4].

### Future work

The problems that are studies in the above papers for improving accuracy for prediction and, diagnosis of diabetes would be worked out further using elastic net regression. Elastic net regression is a combination of LASSO and Ridged Regression techniques to which categorical, numeric and image form data can be given to the regression.

### Authors' contributions

NJ has contributed for acquisition of data, analysis and interpretation of data, Drafting of the manuscript. VB has served as scientific advisors in study conception, design and for critical revision. SR has critically reviewed the study proposal. All authors read and approved the final manuscript.

### Author details

<sup>1</sup> Department of CSE, K L University, Guntur, Andhra Pradesh, India. <sup>2</sup> Sumathi Reddy Institute of Technology for Women, Warangal, India. <sup>3</sup> Department of CSE, Institute of Aeronautical Engineering, Dundigal, Hyderabad, Telangana, India.

### Acknowledgements

I would like to show our gratitude to Dr. Vasantha Vedachalam for sharing her pearls of wisdom with me during the course of this research. Dr. Subba Laxmi for providing necessary information. Dr. P.V. Rao for his comments that greatly improved the manuscript. I thank Dr. Madhu Bala for her insight and expertise that greatly assisted the research. I am also immensely grateful to Ms. B. Padmaja for her comments and interpretations on an earlier version of the manuscript, which was helpful for improvement of the manuscript.

**Competing interests**

The authors declare that they have no competing interests.

**Availability of data and materials**

PIMA Indian data set—UCI Repository.

**Consent for publication**

Not applicable.

**Ethics approval and consent to participate**

Not applicable.

**Funding**

I certify that no funding has been received for the conduct of this study and/or preparation of this manuscript.

**Publisher's Note**

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 25 May 2017 Accepted: 22 June 2017

Published online: 23 August 2017

**References**

- Brown DE, et al. Predictive analytics. Washington: IEEE Computer Society; 2015.
- <http://www.predictiveanalyticsworld.com/patimes/intro-to-machine-learning-algorithms-for-it-professionals-0620152/5580/>. Accessed 2 July 2017.
- [http://www.who.int/diabetes/publications/en/screening\\_mnc03.pdf](http://www.who.int/diabetes/publications/en/screening_mnc03.pdf). Accessed 29 Mar 2017.
- Sanakal R, Jayakumari T. Prognosis of diabetes using data mining approach-fuzzy C means clustering and support vector machine. *Int J Comput Trends Technol*. 2014;11(2):94–8.
- Lakshmi KR, Kumar SP. Utilization of data mining techniques for prediction of diabetes disease survivability. *Int J Sci Eng Res*. 2013;4(6):933–40.
- Repalli P. Prediction on diabetes using data mining approach. Stillwater: Oklahoma State University; 2011.
- Motka R, et al. Diabetes mellitus forecast using different data mining techniques. In: Computer and communication technology (ICCTT), IEEE, 4th international conference. New York: IEEE; 2013.
- Eckerson WW. Predictive analytics. *Tdwi Research*. 2006.
- <http://data-magnum.com/types-and-uses-of-predictive-analytics-what-they-are-and-where-you-can-put-them-to-work/>. Accessed 15 Apr 2017.
- [https://link.springer.com/chapter/10.1057%2F9781137379283\\_6#page-1](https://link.springer.com/chapter/10.1057%2F9781137379283_6#page-1). Accessed 5 July 2017.
- Kalechofsky H. A simple framework for building predictive models. 2016.
- Tevet D, et al. Introduction to predictive modeling using GLMs a practitioner's viewpoint.
- <https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/>. Accessed 20 Apr 2017.
- Gemson Andrew Ebenezer J. Big data analytics in healthcare: a survey. *ARNP J Eng Appl Sci*. 2015;10(8).
- <http://www.dummies.com/programming/big-data/data-science/data-science-for-dummies-cheat-sheet/>. Accessed 30 Mar 2017.
- Predictive modeling, Julie Chambers, the 56th annual Canadian reinsurance conference.
- Abbott Analytics. Strategies for building predictive models. 2014.
- Predictive analytics: poised to drive population health White Paper, Optum.
- Duncan I. Introduction to predictive modeling. 2015.
- <https://www.linkedin.com/pulse/4-types-predictive-analytics-models-mark-rabkin>. Accessed 3 July 2017.
- <http://234w.tc.tracom.net/healthcare/Pages/Diabetes-Readmission-Predictive-Analytics.aspx>. Accessed 25 Mar 2017.
- Lee YH, et al. How to establish clinical prediction models. Seoul: Korean Endocrine Society; 2016.
- Lee J, et al. Development of a predictive model for type 2 diabetes mellitus using genetic and clinical data. *Osong Public Health Res Perspect*. 2011;2(2):75–82.
- Plis K, et al. A machine learning approach to predicting blood glucose levels for diabetes management. Association for the Advancement of Artificial Intelligence. 2014.
- Yang Y, et al. Forecasting potential diabetes complications. In: Proceedings of the twenty-eighth AAAI Conference on artificial intelligence, Copyright c. Association for the Advancement of Artificial. 2014.
- Patil BM, et al. Hybrid prediction model for type-2 diabetic patients. *Expert Syst Appl*. 2010;37(12):8102–8.
- Sarojini Ilango, B. et al. A hybrid prediction model with F-score feature selection for type ii diabetes databases. In: *A2CWIC*. 2010.
- Temurtas H., et al. A comparative study on diabetes disease diagnosis using neural networks. *Expert Syst Appl*. 2009;36(4):8610–5.
- Divya et al. Predictive model for diabetic patients using hybrid twin support vector machine. In: Proc. of int. conf. on advances in communication, network, and computing, CNC. Amsterdam: Elsevier; 2014.
- Ahmed TM. Developing a predicted model for diabetes type 2 treatment plans by using data mining. *J Theor Appl Inf Technol*. 2016;90(2):181–7.
- Devi MN, et al. Developing a modified logistic regression model for diabetes mellitus and identifying the important factors of type II DM. *Indian J Sci Technol*. 2016; 9(4).

32. Thirugnanam M, et al. Hybrid tool for diagnosis of diabetes. *IIOAB J.* 2016;7(5).
33. Osman AH, et al. Diabetes disease diagnosis method based on feature extraction using K-SVM. *Int J Adv Comput Sci Appl.* 2017;8(1).
34. Anand A. Prediction of diabetes based on personal lifestyle indicators. In: 2015 1st international conference on next generation computing technologies (Ngct-2015) Dehradun, India, 4–5 September 2015.
35. Jakhmola S. A computational approach of data smoothing and prediction of diabetes dataset. New York City: ACM; 2015.
36. AlJarullah AA. Decision tree discovery for the diagnosis of type II diabetes. In: International conference on innovations in information technology. New York: IEEE; 2011.
37. Jahani Meysam, Mahdavi Mahdi. Comparison of predictive models for the early diagnosis of diabetes. *Healthc Inform Res.* 2016;22(2):95–100.
38. Hashi EK, et al. An expert clinical decision support system to predict disease using classification techniques. In: International conference on electrical, computer and communication engineering (ECCE), ©2017 IEEE, February 16–18, 2017, Cox's Bazar, Bangladesh.
39. Anand A, Shakti D. Prediction of diabetes based on personal lifestyle indicators. In: Next generation computing technologies (NGCT), 2015 1st international conference on 4–5 Sept. New York: IEEE; 2015.
40. Zanon M, et al. Regularised model identification improves accuracy of multisensor systems for noninvasive continuous glucose monitoring in diabetes management. *J Appl Math.* 2013;2013.

**Submit your manuscript to a SpringerOpen<sup>®</sup> journal and benefit from:**

- ▶ Convenient online submission
- ▶ Rigorous peer review
- ▶ Open access: articles freely available online
- ▶ High visibility within the field
- ▶ Retaining the copyright to your article

---

Submit your next manuscript at ▶ [springeropen.com](http://springeropen.com)

---