Wireless Personal Communications (2022) 127:465–478
https://doi.org/10.1007/s11277-021-08274-w

Predictive Analysis and Prognostic Approach of Diabetes Prediction with Machine Learning Techniques

J. Omana[1] · **M. Moorthi**[2]

Accepted: 4 February 2021 / Published online: 18 February 2021

Abstract
Medical experts indulge in numerous strategies for efficient and predictive measures to model the health status of patients and formulate the patterns that are formed in test results. Most patients would dream of their betterments of their health conditions and thus preventing the progression of any disease. When diabetics is considered in the model, or highly intervening methodology would be required for pre-diabetic individuals. Hidden Markov models have been modified into variant models to derive predictions that accurately produce expected results by investigating patterns of clinical observations from a detailed sample of patient's dataset. There are yet unanswered and concerning challenges to derive an absolute model for predicting diabetes. The datasets from which the patterns are derived from, still holds levels of in completeness, irregularity and obvious clinical interventions during the diagnosis. The Electronic Medical Records are not furnished with all requisite information in all conditions and scenarios. Due to these irregularities prediction has become highly challenging and there is increase in misclassification rate. Newton's Divide Difference Method (NDDM) is a conventional model for filling the irregularity in electronic datasets through divided differences. The classical approach considers a polynomial approximation approach, thus leading to Runge Phenomenon. If the interval between data fields id higher, severity of finding the irregularities is even higher. By using this type of technique it helps in improving the accuracy thereby bringing in high level prediction without any error and misclassification. In this technique proposed, a novel approximation technique is implemented using the Euclidean distance parameter over the NDDM approximation to predict the outcomes or risk of Type 2 Diabetes Mellitus among patients. Real world entities in CPCSSN are considered for this study and proposed method is tested. The proposed method filled the irregularity in the data components of EMR with better approximations and the quality of prediction has improved significantly.

Keywords Prognostic modelling · Prediction · Automated modelling · Type 2 diabetes mellitus · Sparse data handling · Approximation · Machine learning algorithm

✉ J. Omana
omanajayakodi123@gmail.com

Extended author information available on the last page of the article

1 Introduction

Type 2 Diabetes Mellitus is a potential health concern with a signification impact over the epidemic population throughout the world. The disease also termed to be hyperglycemias is considered to be a metabolic syndrome. This category of diabetes will result I permanent failure of organs, damage to internal organs, blindness ultimately leading to fatality. The target area of the disease is kidney, eyes, blood vessels and heart. The diabetic patients are prone to peripheral, cardio vascular dysfunction which finally leads to death [1]. There are severe forms of diabetes that affect a huge population irrespective of the age.

Diabetes progress as an endocrine drive to lead to a numerous side effects which welcomes multiple other devices. In developing countries like India, this disease is increasingly affecting the population at an alarming rate. WHO have ranked the diabetes to be the fifth reason that leads to fatality of patients among other diseases. The International Diabetic Foundation have analysed a found a shocking number that diabetes leads to one death in every six second. Moreover 425 million adults were affected by diabetes in 2017 and it is expected that it may significantly extend to affect 629 million adults by the year 2045. To add more, economic and social impact of health care industry expenditure constitutes to a higher level, for diabetes affected population in developing countries. With the available information, it is estimated that a whooping total of 12% of total healthcare expenditure was spent on treating diabetes in 2017 alone. A notable and serous percent of productivity was affected with patients with this serous health condition working on the premises. Since the disease does not exhibit any visual symptoms, the patients remain unaware of the condition.

The factors are shockingly high and thus need some incredible research contributions to mitigate the effects of diabetics and its variants [2, 3]. The implementations are tested for different scenarios and has to be proven efficient in terms of controlling the progress of the disease. The potential individuals are identified by a number of parameters and their lifestyles. Methods of controlling the progress have to be effective to limit the side effects and other complications. Hence the methodology was implemented with Machine Learning Techniques in this research article. This is a domain in which an intensive research has been invested and has assisted early and advanced detection strategies from available pool of information. The information available throughout the world has been digitized in the form of Electronic Medical Records (EMR) or Electronic Health Records (EHR) to provide a detailed and comprehensive representation of longitudinal information about the diseases, medications, their progress or regress from the treatment plans. Their advanced form of illustrations facilitated the overall progress of deriving the new and accurate history sheets, futuristic analysis of Type 2 Diabetes Mellitus [4]. Thus, the Electronic Health Records have indispensable and plays a pivotal role in data analytics especially in the health industry. This also facilitated on time medical assistance, promising healthcare solutions and round the clock support from medical practitioners. The intention of regulating such information is to provide an end to end support in healthcare and to ensure universal and standard diagnosis to patents globally.

In a conventional method of monitoring diabetes, the patients are tested for their glycaemic values for four or more instances on a single day. Advancements in the sector have been simplifying the entire process through non-intrusive and painless analysis, using a fingerstick analysis of blood glucose and aided by occasional insulin injections. The method is conventional and has its own disadvantages since the fluctuating hyperglycaemia will not derive a standard value. This raised a bar and may be due to intense physical activities,

food intake, or emotional stress factor. The insulin injections are often under or over the prescribed limit. The answer to this limitation is through smart healthcare devices where these technologies are for low cost and immediately available which solves the problem intelligently. It is also presumed that Continuous Glucose Monitoring Systems are implemented to deliver on time and periodical insulin treatments [2, 5]. There are sensors deployed onto the skin of patients, these monitoring systems will keep a track of glycaemic values by sampling one instance every minute. The same data is stored onto the Electronic Health Records and the amount of data generated throughout the medical institution is considerably high. The information will be used collectively to provide a prediction of future values and to concur with a proper diagnosis if the current if the treatment is not working out. This model also allows remote monitoring of patients, equipped with alerting mechanisms and hold a central storage space to hold information to simplify the retrieval of medical history of patients universally. Along with the aforementioned benefits, the system delivers a promising preventive measure in detection of hypoglycaemic or hyperglycaemic conditions and infers the level of insulin that is required at the moment.

Ever since the automated glucose monitoring systems are introduced, various methods of the same genre have been surveyed and documented in the literature survey. On the other hand, from a generic point of view, all methods fall under a collective category named apriori method of glucose level prediction and the other is data driven approach [6, 7]. The apriori methodologies are also termed to be physiological methods where glucose kinetics are equated within mathematical equations to deduce the metabolic response of diabetes patients. The data driven approaches will derive the futuristic values of glucose obtained from analysis of machine learning techniques which worked upon original data from patients. These data driven approaches are not bound to any constraints thus providing a better flexibility, generalized methodology and have better strategies to predict glycaemic values among substandard data availability. This research method concentrates on a data driven approach with a data driven approach that enhances the quality of preprocessing, enriching the training datasets and deliver a better prediction analysis [3, 8]. The next section analyses different state of art methods of the same origin and followed by the contributions of this proposed method.

2 Background Study

Predicting the glucose level in the affected patients has been upgraded with various new parameters and refined over years of study. The existing values have been certainly helpful in guaranteeing the accuracy of the prediction models. These models incorporate the values retrieved from previous studies and treatments through the data collection devices. This literature survey details the approaches used in numerous researches and they are found to be in common. Multiple authors frequently implemented time-based series which are autoregressive or auto neural networks. In an autoregressive method, the results are analysed over the results produced by a first order polynomial methodology [9, 10]. An alternative theory which implements the Kalman filtering was proposed to predict the occurrence of hyperglycaemic spikes based on glucose level monitoring devices.

The methods are prone to multiple errors due to scarce and incomplete information in the datasets. Incomplete information also affects the quality and accuracy of prediction leading to misleading forecast strategies. The forecast windows of the mentioned methods are limited to a short span of 45 min max and methods implementing ANN procured up

to 3 h. When the accuracy horizon increases the results of predictions will decrease. The current methods despite implementing ANN or support regression approaches have shown better accuracy results. These works have highlighted patients with calibrated glucose levels. Currently observed glucose level should have increased from the recently registered glucose level and these patients were considered for studies. The ultimate benefit of this model is that it can be customized and applied for an individual patient, glucose recording devices and their observations. Applicability to individual patient characteristics have been increasingly available with the help of such models yet this carries certain disadvantages. Unless the calibration of individual characteristics is finalized, the same model cannot be applied for other patients, which restricts the extendibility of the model. Moreover, a strong dataset is required for completely adhering to a single patient. The next demerit of this system is that it is too specifically designed being unfit for generalized models for all patients. Any method should be able to accommodate a diverse range of patients, be stable in a dynamic and unpredictable environment. The same should adapt itself to the variations of values, recording devices, medications, diagnosis and longitudinal information of patients. Above all, the recording devices should be generalized and work on all patients immediately without additional or alterations of training dataset.

There are some articles in the survey which proposed a comprehensive model to monitor the glucose level applicable for heterogeneous patients. This was achieved by a diverse range of patients and extracting their details and adding it into a training dataset. An AR model incorporated fixed coefficients, applied a data filtering method along with Tikhonov regularization approach and compare it with three different configurations. The model was tested on different patients who were tested with same glucose monitoring systems and different glucose monitoring systems. Experimental results required a forecasting horizon of 30 min which is lesser than the other conventional methods[9]. Similarly, significant improvement was depicted in the prediction quality and accuracy. The methodology was extended in a feed forward ANN, recurrent neural network (RNN). The prediction quality of these generalized models was significant but uncertain in terms of training dataset. These datasets consist of number of patients from different age groupswhat resulted in poor outcomes.The observed methodologies included a shallow neural network architecture and a few others included deep learning strategies.

The deep learning methods are termed to be the Convolutional Neural Networks [11]. Such methods and networks are particularly needed in a universal model since the quantity and quality of training dataset and other computational resources. Yet these approaches did not possess enough parameters of evaluations in terms of CGM signals that are utilized for training the networks, patients, type of diabetes, types of recording diseases[12]. The existing approaches which are used for Electronic Health records in order to determine the prediction factors are not able to give out the accuracy in terms of misclassification of values. When there are factors that are comprehensive the generation of set of rules seems to be very large that it stops or disturbs the process interpretation.

3 Methodology

The proposed research method in this paper will extend the models which acquires learning from multiple patients, heterogeneous datasets and will utilize information to predict the glucose levels and categorizing it according to the diabetes. The dataset implemented in this research work is obtained from CPCSSN and RT_CGM used for further investigations.

The dataset concentrates on diabetes and eight other neurological diseases. The dataset is available open source and promotes research over the globe. All participating networks enrich the database with medical information about the diseases, their medications and progress of patients. CPCSSN comprises of 172,168 patients, comprises of 812,007 medical records and records over a span of 15 years. Signs, symptoms, diagnosis, plans of treatments, demographics and other information about the patients are available for a detailed study in various research works. Systolic Blood Pressure (SBP), Fasting Blood glucose (FBG), Triglycerides (TG), body mass index (BMI), high density lipoprotein (HDL), Glycated Haemoglobin (HbA(1c) and Gender are the parameters that composed the datasets. Every patient is carefully monitored for a period of 13 years and the standard of defining the parameters is identified to be 8 years. This period is applicable for all patients without any preference to the stage of diseases.

The motto of this research work is to identify the characteristics in the Electronic Health Records, evaluate them according to the progress level and predicting the risk of type 2 diabetes mellitus accurately. This model is specified with the following objectives.

1. Can the machine learning model be derived to predict the diabetes disease?
2. Can the model be applied to detect the prioritized effects / symptoms of type 2 diabetes?
3. Can the proposed method fill in the incomplete information in the datasets and produce an effective dataset to regulate space information?

This model has answered the questions to provide better predictions, to derive specific patients' risks, specific treatments, progress of diseases based on medications and timely treatments.

3.1 Preprocessing the Incomplete Information

The Electronic Medical Records or Electronic Health Records comprises of n instances of patients' information which are completely independent. These instances are denoted by $D = \{S_1, S_2, S_3 \ldots S_n\}$. Such instances are recorded at uniform and standard time intervals, over a period of 8 years for individual patients. The time points are denoted by $T_i = \{t_1, t_2, t_3 \ldots . t_n \mid S_i \mid\}$. The other values are marked as x and collectively they are $X_i = \{x_1, x_2, x_3 \ldots . x_n \mid S_i \mid\}$. Newton's Divided Difference Method is a classical methodology for polynomial equation to interpolate the divided differences. The database comprises of scarce and incomplete information where features and time dimensions are formed into a sparse matrix. The technique is further applied with a pre-processing technique for clearing the incompleteness of such matrix [13].

It is defined that given a set of patients, time and effects of the disease, the sets are formed as follows, (×0, f0), (×1, f1) ……(xn, fn), where×1;×2,…… xn are distinct and not necessarily distributed equally over time. On the other hand, fi is computed mathematically when incomplete information has to be predicted and filled by the methodology. The interpolation function Pn(x) is identified by the following equation.

$$Pn(x)\ D\ f0,\ Pn(x1)\ D\ f1 \ldots .. Pn(xn) = fn$$

The polynomial function is implemented to estimating the missed values in the dataset. This missed information is found to be missed when the records are updated in the Electronic Health records. In addition to the parameters, present in the datasets, the databases are equipped with glucose levels of CGMs and thus facilitates the prediction model. The

approximation technique is applied over the Dimensions of x with respect to time variants. This equation is represented as the Newton form as follows.

$$P_n(x) = f[x_0] + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2] * (x - x_0) * (x - x_1) + f[x_0, x_1, x_2, x_3] * (x - x_0) * (x - x_1) * (x - x_2) + \ldots\ldots + f[x_0, x_1, x_2, \ldots. x_n] * (x - x_0) * (x - x_1) * (x - x_2) \ldots\ldots (x - x_n)$$

$$f[x_0] = f(x_0)$$

$$f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0}$$

$$f[x_0, x_1, x_2] = \frac{\frac{f(x_2) - f(x_1)}{x_1 - x_0} - \frac{f(x_1) - f(x_0)}{x_1 - x_0}}{x_2 - x_0}$$

This method is alternated by Dynamic Linear Method since the time series is considered in the used databases [14]. These models are responsible for a general and non-stationary time series models. DLMs may include terms to model trends, seasonality, covariates and autoregressive components. The goals of this prediction model is to facilitate short term predictions, intervention analysis and monitoring of inputs. Let D be the EMRs information related to patients of datasets namely, P = { p1, p2, p3, …. Pi} which are registered over the time period T. Each patient is marked by the information of their medications, progress with respect to time factors. The series of risk factors that result in diabetes is identified by body mass index, glucose concentrations, insulin secretion, lipoprotein generations over a time interval. In a D dimensional and time series, the real value of an individual patient is denoted by $x_{dk} = y_{dk} =$ { t1,t2,t3 ….. ti [Pi]}. Each patient is already assigned with a unique id Pi, based on the stage of disease. The underlying dimensions, incomplete, sparse data are observed values at different time and based on the stage of diabetes, the number of observations may vary accordingly.

Every irregular and sparse information for every patient Pi will be classified according to the given pair of numbers and denoted by (×1, y1), (×2, y2) … (xi, yi). These number pairs are categorized into upper and lower classes based on the time intervals as depicted in the above Fig. 1. X is divided into (xk + ×1) = 2 in every interval and ((xk + ×1) = 2) + 1 for every other interval. This process is continued until the missing value is filled in the datasets. to _nd any missing value i.e*x*n or *f* (*x*n) in interval two, the available values, *x*n + 1 to *xk*and *x*n -1 are utilized. Likewise, CGMS sensors are found to possess noise to some

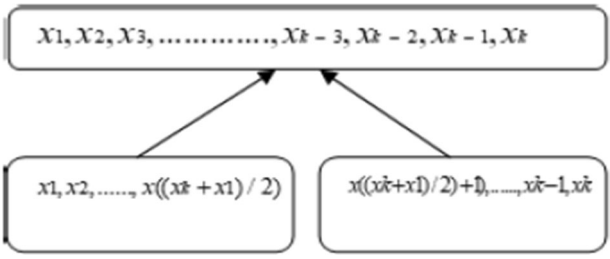

Fig. 1 Classes of parameters

extent and thus need to be cleaned before the information is forwarded for processing in a prediction model. A training dataset needs to be cleaned and thus have to be filled in order to promise accuracy in the prediction system. From the literature survey, it is evident that that these systems tend to clean the missing information and results in over smoothening. This over smoothening process is explicit in a time series environment which results in difficulties in predictions of hypoglycaemic or hyperglycaemic values in the patients. Without smoothening, the entire process.

Algorithm: For predicting missing values

Step 1: X count = x LB where X LB – stands for lower bound of time intervals

Step 2: Initialize Count = 0;

Step 3: Continue the checking process until Xcount is less than or equal to XUB

Step 4: if (Fcount0) is not equal to Null

Step 5: Apply the equation

$$Y_t = F'_t\theta_t + \in_t$$

$$\theta_t = G_t\theta_{t-1} + \omega_t$$

Step 6: Yt is the observation at time t. We assume this is to be a scalar but could also be a vector.

θ is the vector of parameters at time t and of dimension $p \times 1$. Ft is the row vector (dimension 1xp) of covariates at time t and Gt is a matrix of dimension pxp known as evolution or transition matrix. The algorithms helps in step by step process of prediction of missing values. Initially the lower bound of time intervals is set. Further the count value is set to be 0 and iterated until the value reaches upper bound. During the iteration process the observation at specific time is been calculated. The value processed can either be considered to be scalar or vector. Through these values the transition matrix is formed which further identifies the values missing in the dataset.

3.2 Hidden Markov Model

Once the model of filling in the missing values in pre-processing stage is over, the EDLM method is applied to induce a multivariate time series in the D dimension. EHRs are indicated by a sequence of inputs known as reference time points ranging from r_{1n} to r_{in}which are marked to every individual patient Pi. The complete set of approximations is represented mathematically by $P_{yn} = (x_{dn}, y_{dn})$. The same set of inputs are represented in a matrix form with nd in rows and T time slots in the columns. On the other hand, a non-linear autoagressive (NAR) regression model was implemented on the risk factors of CPCSSN datasets and RT_CGM datasets to classify the risk factors with a major and minor impact as depicted in the above Fig. 2. The NAR is an extension of linear autoagressive methods which is completely free from distributions [15]. The technique is preferred in the cases of intrinsic non linearity's that occurs due to intense work out or sudden spike of blood glucose levels. The identified set of major Fig. 3 risk factors were planned to be validated by a medical expert with sufficient years of experience on the same domain. Enhanced Hidden Markov Model is proposed along with the DLM substitution to fill out the missing information in the pre-processing stages.

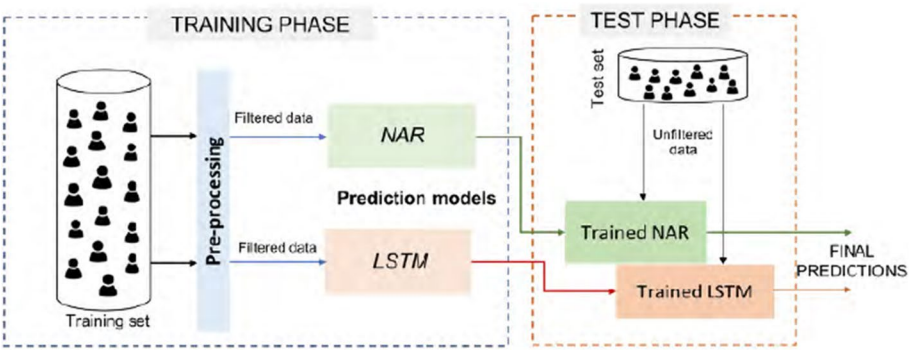

Fig. 2 Process of the prediction system

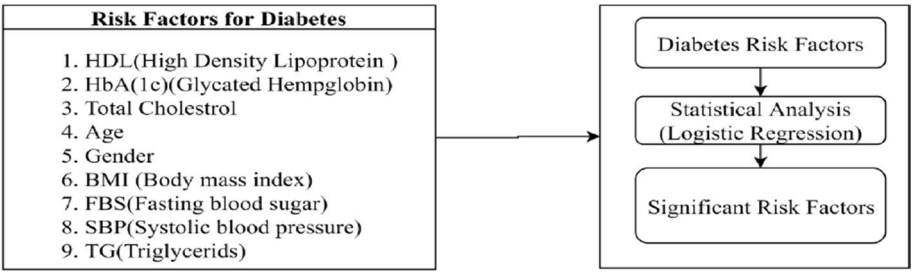

Fig. 3 Analysis of risk factors

The information held in the training and original datasets are continuous and registered for a period of 8 years irrespective of the stage of disease T2DM. An advanced technique that modifies the efficient Hidden Markov Model is applied and from the results, it is evident that the latter is a systematic process and outperforms [16–21] the former. The results demonstrate that the model is a durable, standard and density distribution model when applied on some dynamic entries with respect to time. The hidden nodes of critical importance are sensed with Markov chain assumption when A(pt|pt-1) = A(Pt|Pt-1, Pt-2, Pt-3 ….P0). The proposed enhancement analyses the cardinality between one factor and the other factors to compute a decision tree based on Euclidean distance of relationship between the risk factors. The hidden risk factors are found with the help of a normal Markov Chain application through a stochastic model.

4 Experimental Setup

This investigation includes a number of experiments to analysis and compare the results of the Prognostic Dynamic Linear Methods with NDDM to manage the irregularity and sparsely sampled information in the datasets. The studies are performed over the CPC-SSM and RT_CGM datasets where some values is found to be missing. The methodology intends to fill in the values and identify the risk of developing a diabetic condition. The output of the proposed technique will deliver how the chances of an individual to develop

Type 2 Diabetes Mellitus are. In both of these datasets, time series based Electronic Health Records are related to each risk factor found in a patient. The information in the datasets are taken from individuals above the age group of 18. The next level of investigation is applied to regulate the irregular information obtained in both the datasets. Data values are highlighted with the features and form a matrix to derive the collective information about healthcare industries. The next level of processing completes the filling up of scarce and missing information in the datasets and enable the model of predicting the risk factors of type 2 diabetes mellitus.

The fourth experiment which utilizes the irregular and incomplete datasets is refurbished for implementing the same in the proposed EDLM method. The parameters depicted in the CPCSSN dataset is listed in Table 1.

HMM based applications and models are helpful in classifying the problem into two stages namely the training phase and a decoding phase. In the phase of training, a given collection of parameters is comparatively analysed for finding the probability of relationships between each other in a controlled sequence of events. The decoding phase is accountable for determining the efficient pathways that leads the sequence of events in the right track. This model is executed to strategize the input sequence and validation information. This research work deployed a hold out method for every investigation in the development and implementation. The given datasets are divided into two categories namely the training dataset and testing dataset. In the above experiments, the training and testing datasets are prone to information about the risk factors tabulated in the above Table 1. The Table 2 illustrates how the cardinality between the risk factors are derived to demonstrate the prediction Fig. 4 of T2DM using the proposed auto nonlinear auto aggressive model (NAR).

Table 1 Parameters from CPCSSN dataset

Prediction	Findings	
	Regularly sampled data	Irregularly sampled data
Demographic (Gender, Age)		
Sample size without duplicate	911	1918
Female, sample size (%)	556, (61.03)	775 (61.03)
Male age mean ± SD, Years	58.97 ± 11.83	63.19 ± 11.74
Female age mean ± SD, Years	53.03 ± 11.02	57.53 ± 11.92
Vital Signs/Clinical Measures		
sBP, mean ± SD,mm HG	127.611 ± 15.86	128.611 ± 15.86
Diabetes Mellitus frequency (%)	214 (23.49)	584 (23.49)
Lab Values		
Fasting blood glucose, mean ± SD, mmol/L	5.573 ± 1.93	6.029 ± 1.51
Triglycerides, mean ± SD, mmol/L	1.705 ± 1.027	1.72 ± 1.02
High-Density Lipo protein, mean ± SD, mmol/L	1.313 ± 0.366	1.356 ± 0.39
Light Density Lipoprotein, mean ± SD, mmol/L	2.47 ± 0.97	2.442 ± 0.851
HbA(lc), mean ± SD, mmol/L	6.316 ± 0.824	6.286 ± 0.95
Cholesterol mean ± SD, mmol/L	4.938 ± 1.178	5.409 ± 0.59
Body Mass Index, mean ± SD, kg/m^2	28.76 ± 5.818	29.81 ± 6.362
SD-Standard Deviation; sBP – systolic Blood Pressure; HbA(lc) – Glycated Hemoglobin		

Table 2 Proposed method approximation over other methods

	SE	Asymptotic significance	LB	UB
Irregular datasets	0.082	0.343	0.428	0.739
Regular datasets	0.053	0.00	0.754	0.963
Conventional approximation	0.071	0.004	0.602	0.879
Proposed approximation	0.064	0.00	0.678	0.973

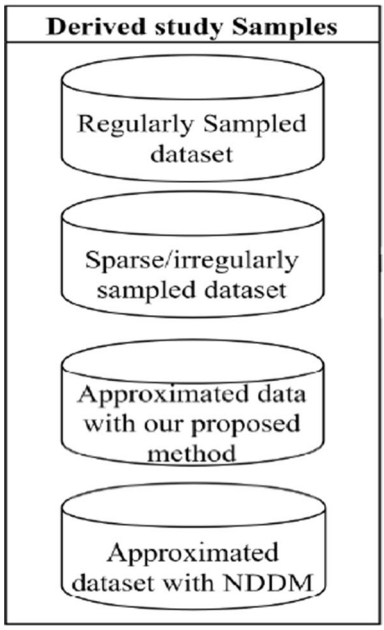

Fig. 4 Proposed Approach of prediction

5 Results and Discussions

The statistical investigations are carried out using the Python libraries and IBM statistics in a Python platform. The research work was carried out to perform an approximation of missing values in an inevitable dataset for diabetes predictions. An Enhanced DLM was designed and implemented to remove the irregular data and fill in scarce and missing information of the datasets. The model intended to predict the risk factors and symptoms for a continuous period of 8 years for every patient who is suspected to be affected by varying glucose levels. The experiments demonstrated compares the proposed EDLM method with the conventional NDDM methodology in the following figure and it is unmistakable. The predicted values are nearly similar to the predicted values delivered by the model.

In the secondary test conducted over the proposed work, a non-linear autoagressive model is applied to assess the risk factors Fig. 5 of T2DM and provides a prediction value to all the disease parameters. The subject of the proposed method is to design and implement a standard classification strategy by carefully determining the potential features based on the relevance, probabilities and cardinalities. According to the NAR model implemented in this research work, every feature other than high cholesterol was linked to diabetes. The results also highlighted HbA(1C) to be the major contributor of diabetes diseases primarily

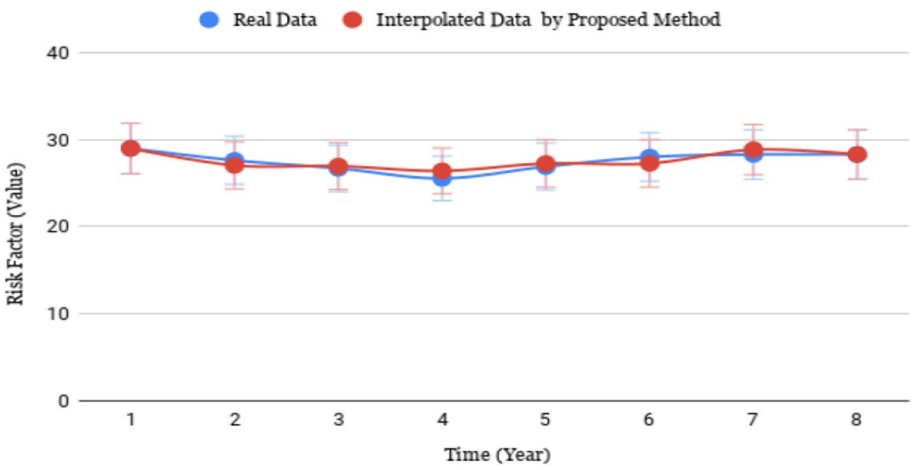

Fig. 5 Risk factor prediction

among the other factors in the given databases. The HbA(1C) has shown the most potential risk factor of Type 2 Diabetes Mellitus with a ratio of $p < 0.0005$, OR = 12.556[94.5% CI, 10.912–14.481]). This can also justify that this factor is the highest relevant risk factor of T2DM in the history of all medical records considered in the databases. The next factor ranking second is FBG with ($p < 0.0005$, OR = 5.917 [95% CI, 1.281–5.967]); proving the results and predictions of a clinical analysis.

The other contributing factors are Body Mass Index, triglycerides and age factor are ranked subsequently according to the model. They have shown the progressions as follows. T2DM ($p < 0.0005$, OR = 1.036 [95% CI, 1.030–1 0.052; $p < 0.0005$, OR = 1.183 [95% CI, 1.093–1.281; $p < 0.0005$, OR = 1.002 [95% CI, 0.999–1.006 respectively). The same result set also depicts that not all obese patients are prone to diabetes but the level of fat tissues determines the level of insulin secreted or to be injected. When the fat tissue is more, then the cells avoid the insulin. As discussed in the previous sections, high cholesterol does not add up to the risk factors and have shown no significant relationship with diabetes type 2. This factor is eliminated from further analysis for complexity reasons.

The methodology is evaluated with the division of four subsets of data from the input databases namely the two original datasets without any approximations, the other two with approximations with the help of prediction models. Of these datasets, the irregular datasets and conventional approximation datasets are unaltered by any external techniques and they are considered to be the ordinary information without any modifications or minimal modifications. The other two datasets are passed through the careful and standard methods to fill in the missing information, the other is passed through the proposed methodology. The output of the proposed approximations held the unique records of 1919 patients with over 15,000 clinical entries in an 8-year window. The Table 2 portrays the outcome of proposed method over the conventional strategies.

The proposed method has explicitly shown improved predictions, accuracy and precision of the outcomes. The quality of such predictions are validated through a comparative analysis with other state of art Auto Regression AROC models. The table also shows how the predictive performance of Enhanced Markov Chain Model is reasonably accurate than the other methods. This is validated by (AROC 80.4%, p-value < 0.0005, SE = 0.064 [95%

CI, (0.679–0.930)]) than the performance over approximated dataset using NDDM (AROC 0.740 p-value < 0.0005, SE D 0.071 [95% CI, (0.602–0.879)]).

6 Conclusion

The results in the previous sections promises a good quality prediction system over the existing methods which prudently analyses the datasets and replace the missing or irregular information with a near true value. The results have shown effectiveness and efficiency of the overall method in predicting diabetes. The purpose of this model is to implement the intelligence of study over forecasting models and equipping them with intelligence for better quality outcomes. The EHRs need to be complete at all times to ensure a proper diagnosis and treatment plan. The proposed model has shown the pathway towards successful predictions of an alarmingly increasing disease in developing countries. Hence this approach explicitly states that prediction, accuracy have increased when compared to the existing approaches. The performance of Enhanced Markov Chain Model is proven to be the best compared to AROC models. The near likely projections have promised better model to be proactive and meaningful for a complete diagnosis. Apart from the efficient and intelligent lives saving measure, the EHRs are properly utilized for virtual doctoring and automated clinical systems. This model has been helpful in designing an efficient tool for advising the patients about a healthy lifestyle, take precautions and build an intelligent system in a digital world. The cost effective model of the same version may still be designed as a future work.

Compliance with ethical standards

Conflicts of interest We authors not having any conflict of interest among ourselves to submit and publish our articles in Wireless Personal Communications journal.

References

1. American Diabetes Association. (2004). Diagnosis and classification of diabetes mellitus. *Diabetes Care, 27*(1), S5.
2. Arulananth, T. S., Balaji, L., Baskar, M., et al. (2020). PCA Based Dimensional Data Reduction and Segmentation for DICOM Images. *Neural Processing Letters*. https://doi.org/10.1007/s11063-020-10391-9.
3. Y. Y. Liu, S. Li, F. Li, L. Song and J. M. Rehg (2015) ``Efficient learning of continuous-time hidden Markov models for disease progression,'' in Proc.Adv. Neural Inf. Process. Syst. 21; 3600-3608.
4. Alberti, K. G. M. M., Zimmet, P. Z., & nition, De. (1998). Diagnosis and classification of diabetes mellitus and its complications" Part 1: Diagnosis and classification of diabetes mellitus Provisional report of a WHO consultation. *Diabetic Med., 15*(7), 539–553.
5. Ekhlaspour, L., Mondesir, D., Lautsch, N., Balliro, C., Hillard, M., Magyar, K., et al. (2017). Comparative accuracy of 17 point-of-care glucose meters. *J. Diabetes Sci. Technol., 11*(3), 558–566.
6. Perveen, S., Shahbaz, M., Keshavjee, K., & Guergachi, A. (2019). Metabolic syndrome and development of diabetes mellitus: Predictive modelling based on machine learning techniques. *IEEE Access., 7,* 1365–1375.
7. Cho, N., Shaw, J., Karuranga, S., Huang, Y., Fernandes, J. D. R., Ohlrogge, A., & Malanda, B. (2018). IDF diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes Res. Clin. Pract., 2*(138), 271–281.

8. Li, Y., Swift, S., & Tucker, A. (2013). Modelling and analysing the dynamics of disease progression from cross-sectional studies. *J. Biomedical Information., 46*(2), 266–274.
9. Saraçoglu, R. D. (2012). Hidden Markov model-based classification of heart valve disease with PCA for dimension reduction. *Eng. Appl. Artificial Intelligence., 25*(7), 1523–1528.
10. Ramkumar, J., Baskar, M., Viswak, M., & Ashish, M. D. (2020). Smart Shopping with Integrated Secure System based on IoT. *International Journal of Advanced Science and Technology., 29*(5), 301–312.
11. Suchithra, M., Baskar, M., Ramkumar, J. P., & Kalyanasundaram, B. (2020). Amutha, "Invariant packet feature with network conditions for efficient low rate attack detection in multimedia networks for improved QoS." *J Ambient Intell Human Comput*. https://doi.org/10.1007/s12652-020-02056-1.
12. Baskar, M., Ramkumar, J., Karthikeyan, C., et al. (2021). Low rate DDoS mitigation using real-time multi threshold traffic monitoring system. *J Ambient Intell Human Comput*. https://doi.org/10.1007/s12652-020-02744-y.
13. Mhetre, N. A., Deshpande, A. V., & Mahalle, P. N. (2016). Trust management model based on fuzzy approach for ubiquitous computing. *Int. J. Ambient Computing Intelligence., 7*(2), 33–46.
14. El, M., Nahas, S. Kassim., & Shikoun, N. (2012). Profile hidden Markov model for detection and prediction of hepatitis C virus mutation. *Int. JournalComput. Science. Issues., 9*(5), 251.
15. Dong, X., Chen, S., & Pan, S. J. (2017). "Learning to prune deep neural networks via layer-wise optimal brain surgeon''. Proc. 31st Int." *Conf. Neural Inf. Process. Syst., 43,* 4860–4874.
16. Zhu, T., Li, K., Herrero, P., Chen, J., & Georgiou, P. (2018). A deep learning algorithm for personalized blood glucose prediction. *Proc. Int. Workshop Knowledge Discovery Healthcare Data., 87,* 1–5.
17. Rodbard, D. (2016). "Continuous glucose monitoring: A review of successes, challenges and opportunities''' in Diabetes Technol." *Therapeutics, 18*(S2), S2-3.
18. Abbott Diabetes Care Division. (2018). WELCOME to the Forefront of Diabetes Care. [Online]. Available: http://www.diabetescare.abbott/
19. D. Madigan (2005) ``Bayesian data mining for health surveillance,'' in Spatial and Syndromic Surveillance for Public Health. A. B. Lawson and K. Klienman, Eds. Chichester, U.K.: Wiley. 203_221.
20. D. Chen, Z. Runtong, S. Xiaopu, W. V. Li and H. Zhao (2018) "Predicting the interaction between treatment processes and disease progression by using hidden Markov model,'' Symmetry.
21. Pugazhenthi, S., Qin, L., & Reddy, P. H. (2017). Common neurodegenerative pathways in obesity, diabetes and Alzheimer's disease. *Biochim. Bio-phys. Acta-Mol. Basis Disease., 1863*(5), 1037–1045.

J.Omana is Assistant Professor in Computer Science and Engineering at Prathyusha Engineering College. She has professional teaching experience of over 8 years. Her area of research interest is Data Mining, Data Science and Analytics. She is currently pursuing Ph.D at Anna University and has published several research articles in reputed National and International journals.

M. Moorthi is a Professor& Head of the Department in the Medical Electronics at Saveetha Engineering College, Chennai. He has professional teaching experience of over 20 years. His areas of research interest are Digital image processing and Multimedia Compression, Microprocessor and Microcontroller. He is Anna University Recognized Supervisor for Ph.D/M.S Scholars. He is currently guiding 07Ph.D Students in Anna University, He has guided 3 Ph.D and 30 Engineering undergraduate and PG student's projects over the past 20 years.He has obtained several funding from AICTE,CSIR,BRNS, IEEE, He has delivered seminars and invited talks to students and faculties in many Engineering institutions. He has published several research articles in reputed National and International journals with high citation indices. He has been the Convener for IEEE International Conference. Currently, he is the IEEE Student branch and Society advisor at Saveetha Engineering College. He is currently acting as an Executive member in IEEE Photonics Society, Madras section.

Authors and Affiliations

J. Omana[1] · M. Moorthi[2]

M. Moorthi
moorthidmp@gmail.com

1 Department of Computer Science and Engineering, Prathyusha Engineering College, Anna University, Thiruvallur, India

2 Department of Electronics and Communication Engineering, Saveetha Engineering College, Chennai, India