An Early Disease Prediction and Risk Analysis of Diabetic Mellitus using Electronic Medical Records

Rutuja A Gulhane$^1$ and Sunil R Gupta$^2$

$^1$Research Scholar, PRMIT&R, Badnera, Maharashtra, India
$^2$Assistant Professor, PRMIT&R, Badnera, Maharashtra India

E-mail: gulhanerutuja@gmail.com, sunilguptacse@gmail.com

Abstract. In the world today, the fourth leading disease is Diabetes that could lead to other serious complicating diseases. Diabetes is one of the most common chronic disease which can also be the cause of death in many cases. An efficient system for early disease prediction and risk analysis of diabetic mellitus is very much needed as it has the major adverse effects. The large amount of medical data is collected by healthcare industry in the form of Electronic Medical Records. The Electronic Medical Records is communal database for clinical disease and risk prediction that are useful in accurately predicting multiple medical events using machine learning approach. Therefore, this research presents an efficient technique for early prediction and risk analysis of diabetic mellitus disease to improve accuracy and precision using Electronic Medical Records.

1. Introduction

The changing habits of individuals has various adverse effects on human health. Diabetic Mellitus (DM) has become an essential global health concern that has no models for its appearance and at any age of life it can impact humans [1]. According to WHO Diabetes was seventh leading cause of death in 2016, and was estimated to have caused four million deaths globally in 2017 [2]. According to International Diabetes Federation Diabetes Atlas, by 2019, an estimated 463 million people were living with diabetes; and by 2030, this figure is estimated to rise to 578 million and by 2045, to around 700 million [3]. DM is a metabolic disease characterized by high blood glucose (BG) and causes abnormal BG regulation that may lead to death if not properly managed. Diabetes is a kind of chronic disease and often accompanies with other disease that might result in short and long-term health complications [4, 5].

DM is generally divided into two types of diabetes: type-1 (T1DM) & type-2 (T2DM) diabetes. In type 1, DM occurs when insulin can no longer be created in the pancreas. Kids, teens and adults will be diagnosed with type-1 diabetes at a younger age. Type 2 diabetes happens when body does not efficiently use pancreas to prevent production of glucose and promote the use of glucose. Type-2 commonly happens in older adults & affects more people who are overweight [6, 7]. Type 1 diabetes does not have complications at the start, as the pancreas remains moderately useful. The disease is perceived only when 80-90% of pancreatic insulin-producing cells are dead. The uncertainty involved in type 2 diabetes brings lifestyle interruption, psychosocial transition, cost for medical treatment. This causes Hypoglycemia & many health problems [8, 9]. About 10% of people are having type-1 & about 90% of people are having type-2 diabetes [10]. DM can affect blood vessels, which increases the risk that the heart, eyes, kidneys, and nerves will be harmed by serious health issues. Diabetic retinopathy, diabetic neuropathy, diabetic nephropathy, and strokes, etc. are the most prevalent diabetes complications [11]. The effects of diabetes can be suppressed when it is identified in the earlier stages
because it cannot be cured, but it can be prevented, detected, and managed [12, 13]. Therefore, early detection of diabetes is the only solution for prevention of the diabetes mellitus disease [14].

Many types of medical information sources are present, such as, hospital information systems (HIS), electronic health records (EHR), or electronic medical records (EMR). EMR is a collection of medical records of individual patients or many people. EMR contains many kinds of healthcare data such as structured, unstructured and semi-structured data. Prediction of disease becomes quick and easier if data is precise, consistent and free from noise. Chronic diseases can be detected and prevented early with current medical techniques. For early diagnosis, patient's data consisting of various features and diagnostics related to disease should be entered with utmost care to provide quality services [15, 16]. The data mining methods, algorithms and its application are more significant for healthcare services [17]. Data stored in medical databases may contain missing values and redundant data, so it is difficult to extract medical data. As it can impact mining outcomes, before implementing mining algorithms, it is important to have good data planning and data reduction. For disease prediction and health risk analysis, effective machine learning algorithms are therefore used to provide accurate results [18].

To perform DM prediction, it needs history of various diabetic patients. Such records are collected from different medical organizations and named as diabetic data set. For detection of diabetes mellitus fuzzy logic, Artificial Neural Network (ANN), Support Vector Machine (SVM), Random Forest (RF), Convolutional Neural Network (CNN), CBR classifiers are used [19]. It is important to select the correct extraction technique as this helps to extract vital information from the signals, thereby enhancing the efficacy of the classifier. [20]. Many techniques are introduced to improve performance of diabetic detection system. But the improvement is still needed for diabetic detection based on accuracy metrics, so this paper presents an efficient early detection and risk level analysis technique.

The research paper of an early disease prediction and risk analysis of diabetic mellitus using electronic medical records is organized as follows: Section I introduces the need of predicting of diabetes mellitus. Section II provides detailed literature review on machine learning based risk prediction model. Section III presents the methodology of early prediction and risk analysis of diabetes mellitus through electronic medical records using machine learning technique. Section IV presents summary and discussion. Finally, Section V concludes the paper.

2. Literature Review

Machine learning has significant contribution in building predictive models in the diagnosis and forecasting of the diseases using electronic medical records (EMR) is being made by recent advances [21]. The emergence of machine learning and electronic medical records for prediction of disease viewed as analogous to the development of imaging, genetic, or some other new cause of highly informative clinical data [22]. Machine learning can be used in numerous ways to achieve rising health data challenges in predicting diabetic mellitus disease and make the system more efficient in practice. The analytic processes such as data mining can solve different problems through data analysis which are based on the category of data that needs to be handled [23]. Recent progress of machine learning methods in designing and applying it on EMR has shown promising outcomes as discussed below:

Cui S et al. [24] presented improved support vector machine (SVM) based method to predict diabetic readmission. There were 3 stages: preprocessing, feature selection and classification. In order to fix imbalanced data and to improve prediction, hybrid feature selection mechanisms were developed to select important features. Synthetic Minority Over-Sampling Technique (SMOTE) data preprocessing were implemented. Finally, using improved SVM-based learning method, classification was carried out. Experimental result has shown that SVM method achieves 81.02% accuracy. Authors claimed that other common methods for readmission predictions were outperformed by decision tree and SVM process.

Zhu C et al. [25] designed data mining model for early prediction of diabetes using principal component analysis, k-means and logistic regression methods. PCA used for dimensionality reduction, k-means technique used for clustering the data and logistic regression used for classification technique. Logistic regression model has been improved by integrating k-means and PCA for prediction of diabetes at an early stage. The authors claimed that the approach can also model a new dataset effectively. The
results showed that PCA improved the logistic regression classifier accuracy of 1.98% higher and the k-means clustering output of 25 more correctly classified data as compared to results of other studies. Also, the approach achieved accuracy of 97.4% for diabetes prediction.

Devi R D H et al. [26] presented hybrid method of farthest first clustering algorithm and support vector machine (SVM) classification algorithm for diagnosing Diabetes Mellitus. Clustering algorithm farthest first was used for gathering data into number of clusters and the time require for computation was reduced significantly as original data diminished. The output of clustering algorithm was given as input to classification algorithm for classification of patients into positive & negative. Results showed that classification accuracy of 99.4% has achieved for diagnosing patient with diabetic and non-diabetic.

Beloufa F and Chikh M A [27] proposed reliable and improved Artificial Bee Colony (ABC) algorithm for diabetes diagnosis by designing optimal fuzzy classifier. Decision-making process of doctor supported by the fuzzy classifier made with ABC approach and provided more information by using optimal fuzzy rules with membership function. The improved ABC algorithm used Pima Indian diabetes dataset for demonstration. Results showed that classification method achieves highest accuracy of 84.21% by using fuzzy rules with improved ABC algorithm in diagnosis of diabetes disease.

Singh N and Singh P [28] have developed ‘‘NSGA-II-Stacking’’ a stacking-based multi-objective evolutionary ensemble system for predicting of Type-2 diabetes mellitus. Initially, data pre-processing step is carried out where the missing values and outliers were identified and the values assigned with the median. The multi-objective optimization algorithm was used to increase classification accuracy and decrease ensemble difficulty for base learner selection. Also, for combine predictions of base learner selection, k-nearest neighbor (KNN) was used as meta-classifier. Experiment results showed that the system achieved 83.8% highest accuracy, 96.1% sensitivity, 79.9% specificity, 88.5% f-score & 85.9% area under ROC curve. Authors claimed that datasets having different features and risk factors can be investigated further for early prediction of DM by modifying NSGA-II stacking method.

El-sappagh S et al. [29] proposed framework for diabetes prediction based on a heterogeneous ensemble classifier. The approach used an ensemble framework by considering classifiers based on random subspace and bagging techniques. The framework uses different data mining techniques, such as naive Bayes, k-nearest neighbors, fuzzy decision tree, decision tree, support vector machine, logistic regression, and artificial neural network, each with separate set of suitable features. The framework accurately selects for each sub-dataset, accurate classifier and appropriate feature set for diabetes prediction. The result showed that framework achieved 90% of accuracy, 94.9% of precision, and 90.2% of recall, which shows that ensemble framework significantly outperforms all other classifiers.

Alama T M et al. [30] introduced early prediction model for diabetes disease. The approach used machine learning & data mining techniques. Important attributes were selected & association of attributes were categorized to predict diabetes. This was achieved by principal component analysis (PCA) technique. It is observed that glucose level & BMI are main features for diabetes prediction which was excavated by apriori method in association rule mining. Random forest (RF), artificial neural network (ANN), and K-means clustering techniques were applied to predict diabetes at an early stage. Among all, ANN technique has valuable treatment results by achieving the best accuracy of 75.7%.

3. Methodology
Early prediction of preventable diseases and classification of risk level is very significant for more efficient health-care resource allocation and better disease management. The aim is to make use of significant features, and design an efficient system to predict DM at early stage and analyze risk level using electronic medical records. The methodology is split into three phases, such as,

PHASE I: Diabetes Prediction
PHASE II: Risk Analysis
PHASE III: Early Prediction.
Figure 1: Work flow of Methodology

PHASE I: Diabetes Prediction
Data preprocessing is an essential step in disease prediction, because of the quality of data, to a large extent, affects the result of prediction. In this step, the preprocessing includes two steps for simplifying the dataset. They are the removal of repeated data and replaced the missing attributes.

i. Removal of repeated data
In this stage, repeated data are removed from dataset using Hadoop Distributed File System. It provides software framework for dispersed storage and processing of big data by using the Map Reduce programming model that consists of map() & reduce() functions to remove duplicate data.

- Map () stage: The input file is given to the mapper function line by line. The mapper processes the data and makes several minor chunks of data.
  \[ m_i = \text{map}(D_i) \]  
  Where, \( \text{map}() \) denotes the map () function and \( D_i \) denotes the input dataset.

- Reduce () stage: The Reducer’s job is to process the data that comes from the mapper. It removes the repeated data and produces a new set of output, which will be stored in the HDFS.
  \[ R_i = \text{red}(m_i) \]  
  Where, \( \text{red}() \) denotes the reduce () function, \( R_i \) is the reduced collection of data.

ii. Replace the missing attributes
If any records encompass unrecorded values, then these values will be filled by replacing the missing value for a particular attribute by the average value for that attribute.

\[ M_i^\theta = \frac{\sum D_i}{n} \]  
Where, \( M_i^\theta \) denotes the missing attribute, \( n \) represents the number of data.

iii. Feature Extraction
In this phase, the important features are extracted from preprocessed data. In DM prediction system, the important features, like Plasma glucose concentration, Diastolic blood pressure (mm Hg), Two-hour serum insulin, Body mass index, Diabetes pedigree function, etc. Extracted features are represented as, 
\[ \tilde{F}(e)_i = \{\tilde{F}(e)_{i1}, \tilde{F}(e)_{i2}, \ldots, \tilde{F}(e)_{in}\} \]  
Here, \( \tilde{F}(e)_i \) signifies extracted feature set and \( F(e)_n \) denotes \( n \) number of extracted features.

iv. Diabetes Disease Prediction utilizing Clustering Algorithm
The clustering algorithm is utilized for categorization of patient results as positive and negative classes. Positive class means, the output of the patient data is a diabetic patient. Here, the value is signified as, 1 and 0, 1 means positive result and 0 implies a negative result. The negative result means the output of the particular data is a non-diabetic patient. The less variation within clusters, the more homogeneous (similar) the data points are within the same cluster. The algorithmic steps of clustering are given below:

\textbf{Step 1:} Specify number of clusters \( K = 2(K_1, K_2) \) then initialize the \( K \) number of class values as \( \tilde{c}_i = \{\tilde{c}_{i1}, \tilde{c}_{i2}, \ldots, \tilde{c}_{in}\} \), here \( \tilde{c}_i \) is the class value dataset \( D_i \) and set of cluster centroid as \( c_i = \{c_{i1}, c_{i2}, \ldots, c_{in}\} \).

\textbf{Step 2:} Select the number of \( K_i \) cluster centers randomly to cluster the dataset.

\textbf{Step 3:} Calculate the distance between each class values and cluster centers.

\[ E_{d1} = |\tilde{c}_i - c_1| \]  
\[ E_{d2} = |\tilde{c}_i - c_2| \]  

\textbf{Step 4:} Assign datapoint to cluster center whose distance from cluster center is minimum of all centers.

\textbf{Step 5:} Repeat steps 2, 3, and 4 until the same points are assigned to each cluster in consecutive rounds.

The final clustering result contains two classes \('0', '1'\). Here, '0' indicates DM negative and '1' indicates DM positive. In the case of DM positive, Phase II is performed to analyze the future risks in diabetes mellitus. Otherwise, Phase III is carried out for the early prediction of DM.

PHASE II: Risk Analysis
Diabetes can lead to problems that affect many parts of the body, including brain, heart, eyes, kidneys, and nerves. Thus, proper assessment of patient's prognosis plays a central role in the management of diabetes. This phase is utilized to ascertain Risk Analysis of diabetic disease after 3 years.

The risk analysis phase consists of four steps, that are, merge input dataset, pre-processing, feature extraction, and classification. First, data from diabetic disease is merged by utilizing common attributes, like sex, age, etc. to form a new dataset. Then, the data is arranged into a standard format for obtaining a better result, which is done in the pre-processing step. After that, the important features, such as BP, Anemia, numbness, Blood urea, and Blood sugar, etc. are extracted from dataset for efficient analysis. The extracted feature set is denoted as, \( F_D^I \). In this step, extracted features \( F_D^I \) are given as input to machine learning algorithm to ascertain risk value for diabetic neuropathy, retinopathy, and nephropathy after 2 or 3 years. The machine learning algorithms classify the risk level of diabetes patients effectively.

PHASE III: EARLY PREDICTION OF DM
In this phase, early detection of diabetes mellitus is carried out. Early diagnosis of diabetes and prediabetes is important so that patients can start to manage the disease early and possibly prevent or delay the serious disease complications. The steps involved in early prediction of DM is given as follows:

\textbf{i. Non-Diabetic Patient Dataset}
From the hospital laboratory, the method gathers patient's previous health information about blood glucose tests, age, fasting blood glucose (FBG), Oral glucose Tolerance (OTG), HbA1c examination.

\textbf{ii. Feature extraction}
In this phase, important features are extracted from dataset. Technique extracts following features:

\begin{table}[h]
\centering
\caption{Extracted Feature set}
\end{table}
Above mentioned features are chosen for DM early prediction. Extracted feature set is expressed as, 
\[ F_{i}^{nd} = \{F_{1}^{nd}, F_{2}^{nd}, \ldots, F_{n}^{nd}\} \quad n = 5 \] (7)

Here, \( F_{i}^{nd} \) signifies the extracted feature set and \( F_{n}^{nd} \) denotes the \( n \) number of extracted features.

### Table 2. Diagnosis level of Diabetes Ranges

|                           | Age  | BMI             | HbA1c | FPG (mg/dL)  | OGT (mg/dL) |
|---------------------------|------|-----------------|-------|--------------|-------------|
| Diabetes                  |      | - 6.5 or above  | 126 or above | 200 or above |
| Pre-diabetes              | Above 45 | 25 to 29.9 | 5.7 to 6.4 | 100 to 125 | 140 to 199 |
| Normal                    | -    | 18.5 to 24.9    | Above 5 | 99 or below  | 139 or below |

iii. Euclidean Distance Ranking (EDR) Method for Early Prediction of DM

In this phase, the early prediction of DM is carried out utilizing the Euclidean Distance Ranking method. The EDR method is a basic and simple technique that measures the distinguishing between two feature set with real values. The techniques use two feature set for EDR calculation, one is extracted feature set \( F_{i}^{nd} \) and another one is “diabetic-prediction feature set”, which is denoted as \( F_{i}^{dp} \). The real value of each feature in \( F_{i}^{dp} \) is shown in table 2. The Euclidean Distance between two features is calculated as:

\[ Eucl_{dis}(F_{i}^{dp}, F_{i}^{nd}) = \sqrt{\sum_{i=1}^{n}(F_{i}^{dp} - F_{i}^{nd})^2} \] (8)

\( Eucl_{dis}(F_{i}^{dp}, F_{i}^{nd}) \) measures the numerical difference for each corresponding features of feature set \( F_{i}^{dp} \) and \( F_{i}^{nd} \). The stages are found centered on the \( Eucl_{dis} \), which means that the method fixed some threshold if the \( Eucl_{dis} \) is below the threshold, then the stage is denoted as the ‘risk’. Otherwise, the stage is denoted as 'no risk'. After calculating each feature stage, the final result is a decision on whether a person has potential against DM or not is predicted using following rules.

### Table 3. Rules involved in Early Prediction of DM

| Rule Details | Early Prediction Output |
|--------------|-------------------------|
| Features (F3 or F4 or F5) | Features (F1 or F2) |
| No risk | No risk | No potential |
| Risk | No risk | Potential |
| No Risk | Risk | No potential |
| Risk | Risk | High potential |
Based on the above-mentioned rules, the patient's potential against DM is predicted at an early stage. This early prediction system can help to reduce the risk of serious complications, such as premature heart disease and stroke, blindness, limb amputations, and kidney failure.

4. Summary and Discussion
Availability of huge amount of clinical data in the form of electronic medical records is the origin of the research which has ability to transform healthcare industry. In many of the research, the most important challenge is to build models that accurately diagnose and predict diabetic mellitus disease by developing different algorithms and methods for processing and analysis of biomedical data. On analyzing existing techniques, it is observed that more diabetes mellitus cases are caused due to delayed detection. So, early recognition of diabetes is needed to treat diabetic mellitus and also diabetes patients will have so many risks in the future that may damage brain, heart, eyes, kidneys, and nerves. It is also needed to provide efficient and accurate system to enhance workflow, clinical tasks, streamlining work of early disease prediction and risk level classification. Machine learning models yield better performance in many tasks and require less manual feature engineering. Thus, the system will be useful to improve quality of life and avoid health risks.

5. Conclusion
The research focuses on applying machine learning methods and its applications in prediction of diabetic mellitus disease and health risks for healthcare systems using electronic medical records (EMR). A continuously growing large amount of EMR has to be processed and analyzed using data mining to improve accuracy and precision of disease prediction and future health risks. The objective is to analyze electronic medical records to efficiently classify sets of patient data in different categories (e.g. normal and disease affected) by investigating machine learning based approaches to improve efficiency and accuracy of risk analysis. It helps to improve performance of risk level classification for different future stages of disease affected patients. Further, machine learning algorithms is applied to achieve better results in healthcare systems. The approach is designed in such a way that it will give greater results than previous diabetes prediction model. Various machine learning algorithms are used in state-of-art works but low accuracy problem still exists. So, this paper presents an efficient and robust technique to improve accuracy and precision of early disease prediction and risk level analysis of Diabetic Mellitus.

In future work, finding from this paper in changing medical practices, leading to more cost effective by considering unstructured data like family health history, previous health check, etc. The further research can be carried out on how to improve performance of risk prediction for different types of diseases.

References
[1] Baiju B V and Aravindhar D J 2019 1st international conference on innovations in information and communication technology (ICIICT) IEEE pp 1-6
[2] Health Topics, “Diabetes”, World Health Organization, [Online] Available https://www.who.int/news-room/fact-sheets/detail/diabetes
[3] Saeedi P, Petersohn I, Salpea P, Malanda B, Karuranga S, Unwin N, Colagiuri S, Guariguata L, Motala A A, Ogurtsova K, Shaw J E, Bright D and Williams R 2019 Diabetes Research and Clinical Practice vol 157
[4] Woldaregay A Z, Arsand E, Walderhaug S, Albers D, Mamykina L, Botiss T and Hartvigsen 2019 Artificial Intelligence in Medicine pp 109-134
[5] Zheng H, Park H W, Li D, Park K H and Ryu K H 2018 5th NAFOSTED Conference on Information and Computer Science (NICS), IEEE pp 110-113
[6] Qalhati N S J M Al, Hussain S A and Singh A V 2018 7th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), IEEE pp 840-846
[7] Fabris C, Facchinetti A, Sparacino G and Cobelli C 2014 36th Annual International Conference of IEEE Engineering in Medicine & Biology Society pp 6643-46
[8] Yahyaoui A, Jamil A, Rasheed J and Yesiltepe M 2019 *1st International Informatics and Software Engineering Conference (UBMYK)* IEEE pp 1-4

[9] Sumangali K, Geetika B S R and Ambarkar H 2016 *International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)*, IEEE, (Kumaracoil, India) pp 389-392

[10] Undre P, Kaur H and Patil P 2015 *International conference on pervasive computing (ICPC)* IEEE: Advance Communication Technology and Application for Society pp 1-4

[11] Fiarni C, Sipayung E M and Maemunah S 2019 *Procedia Computer Science* vol 161 pp 449-457

[12] Kannadasan K, Edla D R and Kuppili V 2019 *Clinical Epidemiology and Global Health* vol 7 pp 530-535

[13] El-Sappagh S, Ali F, El-Masri S, Kim K, Ali A and Kwak K S 2019 *IEEE* vol 7 pp 21917-47

[14] Rosyadi I, Ramadhon E, Wahono A T U and Hijrati Y N 2018 *1st International Conference on Biometrics, Biotechnology and Biomedical Engineering*, BioMIC vol 1 pp 1-4

[15] Lan G, Lee C, Lee Y, Tseng V, Chin C, Day M, Wang S, Chang C and Cheng S 2012 *Conference on Technologies and Applications of Artificial Intelligence* pp 27-32

[16] Jain D and Singh V 2018 *Egyptian Informatics J* vol 19 pp 179-189

[17] Jothi N, Rashid N A and Husain W 2015 *Procedia Computer Science Elsevier in Information Systems International Conference (ISICO)* vol 72 pp 306-313

[18] Shinde S A and Rajeswari P R 2019 *International Journal of Engineering and Technology* vol 7 pp 1019-23

[19] Han L, Luo S, Yu J, Pan L and Chen S 2015 *IEEE Journal of biomedical and health informatics* vol 19 pp 728-734

[20] Lekha S and Suchetha M S 2017 *IEEE Journal of biomedical and health informatics* vol 22 pp 1630-1636

[21] Battineni G, Sagaro G G, Chinalalapudi N and Amenta F 2020 *Journal of Personalized Medicine* vol 10

[22] Adkins D E 2017 *American Journal of Psychiatry* vol 174 pp 93-94

[23] Islam M, Hasan M, Wang X, Germack H and Noor-E-Alam M 2018 *Publication of Healthcare (Basel)* vol 6

[24] Cui S, Wang D, Wang Y, Yu P and Jin Y 2018 *Computer methods and programs in biomedicine* vol 166 pp 123-135

[25] Zhu C, Idemudia C U and Feng W 2019 *Informatics in Medicine* vol 17

[26] Devi R D H, Bai A and Nagarajan N 2020 *Obesity Medicine* vol 17

[27] Beloufa F and Chikh M A 2013 *Computer Methods and Programs in Biomedicine* vol 112 pp 92-103

[28] Singh N and Singh P 2020 *Biocybernetics and Biomedical Engineering* vol 40 pp 1-22

[29] El-Sappagh S, Elmogy M, Ali F, Abuhmed T, Islam S M R and Kwak K S 2019 *Electronics Article Special Issue Internet of Things (IoT) Based Wireless Health: Enabling Technologies and Applications* vol 8 pp 1-29

[30] Alama T M, Iqbal M A, Alia Y, Ijazb A W S, Baigh T I, Hussainc A, Malikb M A, Razab M M, Ibrarb S and Abbas Z 2019 *Informatics in Medicine, Elsevier* vol 16