Detecting Type 2 Diabetes Mellitus Using Electronic Health Records and Machine Learning

Tayyab Nasir[1] and Moosa Ali[1]

CureMD Research and Development
tayyab.nasir@curemd.com

Abstract. Type 2 diabetes mellitus (T2DM) is one of the most common disease and one of the leading causes of death. The problem of early diagnosis of T2DM is challenging and is necessary to prevent serious complications. In this study,we present the most diverse T2DM prediction dataset consisting of 9,792 patients' data that is collected from 295 different practices around the USA. A novel set of features for the prediction of T2DM have also been presented. Techniques like Principal Component Analysis (PCA),Boruta,and Minimum Redundancy and Maximum Relevance (MRMR) have been used for extracting features from a sparse set of EHR attributes. Seven different machine learning techniques were used to build and evaluate the generated dataset and features for T2DM prediction with an intent to build a system for early prediction of T2DM using routine EHR data. Finally,we present a detailed analysis of the results obtained using different evaluation measures,validating the accuracy,precision,recall,specificity,and F1 score for each technique using 5-fold cross-validation. The fairness of each technique is also evaluated.

Keywords: EHR · Feature Selection · Machine Learning · T2DM · Type 2 Diabetes.

1 Introduction

Diabetes mellitus or simply diabetes is a set of conditions concerning the body's mechanism of managing blood glucose [2, 20]. The human body converts carbohydrates to glucose which is then converted to energy. The pancreas plays a key role in producing insulin which is a hormone responsible for converting glucose from the bloodstream to energy. Thus, insufficient production of insulin or the inability to utilize the produced insulin can cause diabetes [3, 62]. Diabetes causes the blood glucose level to stay high which is a condition known as hyperglycemia [72]. This,in turn,can cause a number of other disease like heart attack,hypertension,urinary organ disease,joint failure,stroke [72, 56, 38, 54, 70, 26, 79] and can affect the nerves,eyes,kidneys,and other vital organs [17, 77, 60, 41]. Diabetes is a chronic disease with no known cure [7, 26, 19, 33, 22, 72, 61]. However,it can be managed and a diabetic person can lead a long life if proper treatment is timely provided [2]. Treatment of diabetes involves medication and

following a healthy lifestyle. Diabetes has been on the rise and its upwards trend is projected to stay the course in the coming years [20].

1.1 Types of Diabetes

There are two major types of diabetes mellitus i.e., Type 1 Diabetes mellitus (T1DM) and Type 2 Diabetes mellitus (T2DM) [2, 20]. T1DM also called insulin-dependent diabetes is a condition that usually begins at an early age. T1DM causes the auto immune system of the body to attack the insulin-producing cells thus, causing a lack of insulin. Patients suffering from T1DM require some sort of insulin injected on regular basis to control their blood glucose levels [20]. T1DM has symptoms that often involve unplanned weight loss,increased thirst (polydipsia),frequent urination (polyuria),and increased hunger [80, 20, 70, 63]. T2DM is non-insulin-dependent diabetes and occurs more often in middle-aged and older people [2, 20, 3]. However, over the past 20 years, T2DM is becoming more common in adolescents and even in children [20]. Almost 90% of the total population having diabetes have T2DM which signifies the importance of diagnosing T2DM [2, 20]. T2DM has the same symptoms as T1DM but the symptoms are milder. Other symptoms of T2DM involve slow wound healing, fatigue, depression, and high blood pressure [30, 20]. Also, previous studies show higher comorbidity of T2DM with various diseases including cardiovascular disease, hypertension, angina, gastric ulcer, and hypothyroidism [81, 29, 43, 27, 16]. Although T2DM itself is less severe than T1DM yet it can cause serious long-term complications like affecting the kidneys, eyes, and nerves [25, 4]. Also, it increases the risk of heart disease and stroke [24]. Similarly, there are several risk factors involved when it comes to T2DM including obesity, ethnicity,older age,high blood pressure,family history,and genetics [19, 13, 65].

1.2 Significance of Diabetes

The number of people that are Type 2 Diabetic is rapidly growing. According to the International Diabetes Federation (IDF) around 536.6 million people were diabetic in the year 2021 [20]. This number is projected to grow to 783.2 million by 2045. As stated almost 90% of the total diabetics are diagnosed with T2DM. Thus, the total number of type 2 diabetic persons in the year 2021 was 483 million [20]. Diabetes has a high mortality rate as well. According to the IDF,6.7 million people died due to some complications caused by diabetes in 2021 alone [20]. Like the world, the USA is also impacted hugely by diabetes. A total of 32.2 million US citizens were diabetic in the year 2021 which is around 11% of the total US population [20, 1]. The US is at number 4 on the list of the countries with most diabetic patients and spends a huge amount every year on fighting diabetes and its underlying complications [20]. In 2021 a total of 966 billion US dollars were spent on fighting diabetes which is 316% higher than the 232 billion US dollars spent in the year 2007 for the same [20].

1.3 Diagnosis of T2DM

For the diagnosis of T2DM glycated hemoglobin (A1C) test is most commonly used [63, 5]. This test uses an average of blood sugar levels over the past two or three months. Other clinical tests including the fasting blood sugar test and oral glucose tolerance test are also used for diagnosing diabetes [58, 63, 5, 20]. Early detection of diabetes is of vital importance for an effective treatment [21]. Early diagnosis of T2DM can help prevent many serious complications including but not limited to loss of vision, kidney failure, amputations of limbs, stroke, and even premature death [45, 35]. Thus, an early diagnosis of T2DM is essential which can not only save millions of dollars every year but also save precious human lives. However,diabetes in general and T2DM,in particular,remain undiagnosed in the vast majority [21].

1.4 The Problem of Undiagnosed T2DM

Early detection of T2DM is challenging [85]. T2DM is often diagnosed very late when the symptoms become severe and complications arises [20, 21]. Diagnosis of diabetes is usually delayed even in the countries that have more developed health care systems. The significance of the problem of undiagnosed diabetes can be estimated from the National Diabetes Statistics Report, which states that around 8.5 million diabetic persons in the US alone are undiagnosed [1]. Hence it is desirable to have a system that can help in the early diagnosis of T2DM, without relying on the aforementioned clinical tests but on some non-invasive measure that can act as an indicator to help identify the risk or presence of T2DM in a patient. Such a system can help the doctors in the early diagnosis of T2DM and thus can help in reducing the yearly health expenses spent tackling complications of T2DM and also save millions of lives every year that are lost due to late diagnosis of T2DM.

1.5 Machine Learning for T2DM Diagnosis

Machine learning has been used for providing solutions to many problems in healthcare. Over the years many machine learning techniques have been used for medical image analysis, disease diagnosis, and prognosis, treatment failure prediction, etc. Predicting T1DM and T2DM has been studied previously using different supervised and unsupervised machine learning algorithms. Data is the building block of any machine learning-based technique, whereas machine learning itself is defined as the branch of artificial intelligence that learns pattern from the given data and use these patterns to make predictions on unseen data. Many datasets have been used for training and evaluating machine learning techniques for T2DM prediction, the details of which are given in the following section 1.6.

1.6 Datasets for T2DM

Three significant datasets have been used previously for T2DM detection which are the PIMA Indian Diabetes dataset [69], UCI [71], and the MIMIC-III [36].

Out of these, the PIMA Indian diabetes dataset is the most widely used. The Table 1 gives the details of each dataset.

Table 1. Datasets for T2DM.

Dataset	Positive Examples	Negative Examples	Total Examples
PIMA	179	358	537
UCI	51,034	14,806	65,840
MIMIC-III	1,242	38,456	39,698

A large number of studies have used the PIMA dataset to build T2DM prediction models [4]. The very first study that used machine learning for predicting T2DM using the PIMA dataset dates back to 1988 [69]. Techniques like KNN,Logistic Regress,Naïve Bayes,Decision Trees,Support Vector Machines (SVM),Random Forests,AdaBoosting,and Gradient Boosted Trees have all been used to train and evaluate T2DM prediction models using the PIMA dataset [74, 23, 40, 7, 54, 26, 39, 41, 53, 76, 67, 60, 8, 37]. Apart from the conventional machine learning techniques,deep neural networks have also been widely used for T2DM prediction. Previous studies have used multi-layer perceptron networks to predict T2DM, where PIMA was used for training and evaluation [75, 80, 6, 56, 55]. Many other neural network architectures including CNN, RNN (both LSTM and GRU), and a combination of CNN-LSTM-based deep neural networks were evaluated using the PIMA dataset [82, 62, 22, 4]. Additionally, many studies have used voting and stacking-based ensembling to combine the aforementioned techniques to improve the overall performance of T2DM prediction models [74, 42, 40, 60].

Previous studies have also applied machine and deep learning techniques to datasets other than PIMA. All the aforementioned machine learning techniques have been used to build T2DM prediction models using such datasets as well [49, 25, 19, 81, 76]. For examples NHANES dataset [57] was used to train a SVM based model for predicting diabetes [83]. Similarly, another study combined NHANES dataset [57] with lab results of different patients for training Logistic Regression, SVM, Random Forest, and Gradient Boosted Trees for predicting T2DM [42]. Similarly, such datasets have also been used with deep neural networks [70, 17, 16]. Many studies have also collected data consisting of the same features as the PIMA dataset from other sources to build machine learning-based T2DM predictive models [25, 54]. There are many problems in most of the datasets discussed before and the associated studies that use such datasets for training machine and deep learning techniques. The PIMA dataset has a very limited number of examples, which raises the concern of a lack of versatility in the dataset that is required for better generalization of the models. Also, the dataset contains examples belonging to only the female PIMA Indians of Arizona further limiting the variation in the dataset. Also, the aforementioned datasets are collected in a more controlled environment and may lack the inherent features and characteristics of real-world data.

1.7 Use of EHR Data for T2DM Prediction

Over the past decade, a large number of hospitals and practices have moved to digital health records [36, 12]. The widespread adaption of EHR systems has produced a large amount of digital data [31, 59, 64]. Such data can help significantly in clinical research and has been used for a large variety of health analytics [59, 12]. Data coming from EHR systems is more close to the real world and can help build more accurate T2DM prediction systems. Previous work has used data from different EHRs to build machine learning models [65, 52, 84, 44] for T2DM prediction. For example, studies have used Logistic Regression, SVM, Random Forest, Naïve Bayes, and Gradient Boosted Trees on EHR data collected from different hospitals in different parts of the world [14, 18, 77]. Also, deep neural networks have been used for T2DM on EHR data [45, 4]. Most of such studies that utilize some EHR data rely on clinical components like lab results, historical diagnosis, family medical history, or a combination of these. Such studies also have limitations of their own. Although data collected from EHRs might be significantly larger yet in most of the studies data is collected from a single practice or region and might not be an effective sample of the data required for better generalization of the machine and deep learning tehcniques [47, 15, 10]. In addition to the techniques that were mentioned before that rely solely on tabular data features, some of the studies also rely on images and signal data for the prediction of T2DM. A couple of studies have utilized Heart Rate Variation data collected from ECG and built machine learning and deep neural-based models to predict diabetes [73, 72]. Similarly, images of the retina have been used by a couple of studies to predict T2DM using deep CNN and LSTM-based models [32, 66].

1.8 Contributions

This work aims to overcome the shortcomings of the existing research including the lack of volume and diversity of datasets, data being collected in a controlled environment, and the nature of features like tricep skin thickness and blood insulin that are not part of routine EHR data. The key contributions of this work as given as follow:

- Propose a dataset that is derived from multiple EHR systems running in different hospitals and practices in different parts of the US. The intent here is to create a dataset that is diverse enough for better generalization of machine learning techniques and is collected in a real-world environment.
- Identify novel features using state-of-the-art feature selection techniques, that rely only on historical diagnosis, patient vitals, and demographics. The idea is to propose features that rely on routine EHR data and do not require some invasive lab procedure or measurements.
- Use the novel features to build and evaluate multiple Machine Learning techniques that have been previously used for T2DM prediction and analyze the performance of each using different evaluation measures.

- Present a detailed evaluation of the fairness of all our presented techniques under different settings.
- Present an external evaluation of all the techniques using an out-of-sample dataset. The out-of-sample dataset is to be created using data from an EHR that is not used during training, validation, and testing. Also, present the largest test dataset for the external evaluation of all the techniques presented.

The rest of the paper is divided as follows: Section **??** gives an overview of the corpus collection and its specifications, along with the required pre-processing. The same encloses the associated details of the feature selection process and the experiments performed on the said dataset. The results obtained and the analysis performed on the results are presented in Section **??**. Section 4 provides the conclusions and directions for future work.

2 Methodology

2.1 Dataset

For this study, we used data from CureMD which is one of the leading EHR providers in the US. We were provided with EHR data of 738 different practices. Out of these 295 unique practices were selected based on the amount of Type 2 diabetic patients they have. Hence, we can state that we collected data from 295 different EHR environments operating in different parts of the US which makes our data the most diverse of its kind for T2DM prediction. As discussed in the introduction section the motivation behind selecting data from multiple EHRs is to be able to build a system that is trained using real-world data and which relies only on features from routine EHR, providing a non-invasive method of T2DM detection that can be plugged into any EHR. The raw data from CureMD EHRs contained 5,092,987 unique encounters from 2011 to 2021. The data consisted of 1,145,465 unique patients. Many data attributes that contained personal information related to the patients were already replaced with dummy values to comply with the user privacy policy of the EHR. For example data fields like the first name, last name, middle name, social security number (SNN), address, email, etc. were hidden and replaced with dummy values. For this study we used three clinical components details of which are given as follows:

Diagnosis The Diagnosis clinical component represents a set of ICD-10-CM codes, each representing a diagnosis associated with a particular visit. The raw data had a count of 29,714 unique ICD-10-CM codes. These codes will be used as individual features to identify diseases that can either lead to T2DM or is comorbid to T2DM.

Vitals The Vitals clinical component represents a set of vitals readings that were taken during a visit. There were 18 unique vital attributes in the raw data including Weight, Height, BMI, BSA, Lean Body Weight, Ideal Body Weight,

Neck Circumference, Waist, Oxygen Saturation, Peak Expiratory Flow, Blood Type, Blood Rh, Finger Stick, Pulse, Respiration, Temperature, Systolic Blood Pressure, Diastolic Blood Pressure.

Demographics The patient's Demographics component consists of attributes including the patient's date of birth (which can be used to calculate the present age of the patient), gender, sexual orientation, race category, employment status, and other personal information attributes that were replaced with dummy values like patient's name, address, social security number, etc.

2.2 Data Preprocessing

The CureMD EHR data in its raw form required a certain set of preprocessing steps to mold it into a form useable for experiments. Data collected from EHR is sparse,noisy,heterogeneous,and unstructured [84, 64, 12]. Also challenges like missing values for certain attributes, and class imbalance is also present [68, 31]. The raw EHRs data that we obtained also have all of these inherent issues. These data challenges and the required preprocessing techniques that were used to tackle these challenges are discussed below:

Handling Records with no History We intend to build a system that can predict T2DM by merely analyzing a patient's diagnosis history, recorded vitals, and demographics. So, we are interested in patients that have some historic visits before being diagnosed with T2DM. However, in our raw EHR data, we encountered a large number of records where a patient was marked with T2DM (indicated with an E11 ICD-10-CM) in the very first encounter. The reason for such cases is obvious, for example, an already T2DM-diagnosed patient may walk into a hospital or practice for the first time without any medical history. Thus, we have to come up with a strategy to tackle such cases. We dropped all patients with T2DM that do not have at least 3 previous encounters before being diagnosed with T2DM. Applying this step our data set was reduced to 4,896 out of the overall 72,626 patients with T2DM. Also, the patient records without T2DM that do not have at least three encounters were dropped.

Handling Missing Data The issue of missing data is a widely encountered problem in the EHRs [68, 31]. Many techniques have been used previously to handle missing data. Techniques like dropping records with missing values, and filling in missing values using mean, median, or from the nearest neighbors have been used in many research studies [23, 54, 81, 53, 15, 8, 86]. The raw EHR data was processed to identify the attributes that have the missing data problem. The problem was only encountered in the attributes of the Vitals component. To deal with missing values a 2-step strategy was applied. In the first step, we performed a missing value ratio (MVR) calculation of each attribute. The missing value ratio is given by the following equation:

$$MVR = 100 * \frac{Number of Records with Missing Values}{Total Number of Records} \quad (1)$$

The Table 2 presents the details of the missing value ratio of different attributes. Based on the results given in the Table we dropped the vitals for which the missing value ratio was higher than 60% as such variables do not possess a sufficient amount of information that can be used to impute the missing values.

Table 2. Details of Missing Values Ratio.

No.	Attribute	MVR %
1	Weight	12.67
2	Height	21.53
3	BMI	28.34
4	BSA	30.24
5	Lean Body Weight	64.71
6	Ideal Body Weight	65.14
7	Neck Circumference	92.22
8	Waist	91.80
9	Oxygen Saturation	89.15
10	Peak Expiratory Flow	99.99
11	Blood Type	99.82
12	Blood Rh	99.94
13	Finger Stick	98.98
14	Pulse	28.91
15	Respiration	53.93
16	Temperature	51.66
17	Systolic Blood Pressure	16.48
18	Diastolic Blood Pressure	16.44

The second step consists of imputing the missing values. For the remaining columns we performed the following steps to fill in the missing values:

- Firstly,all the records were grouped based on the patient they belong to.
- For each patient group,the Exponentially Weighted Moving Average (EWMA) of every column was calculated and stored.
- Next for all the missing values of an attribute,its calculated value (which is calculated using EWMA) was used to fill in the missing values.

EHR data is temporal thus we used EWMA which is used to describe a time series to fill in the missing values. EWMA can assign larger weights to the recent observations and smaller to the ones that are older. This actually will be quite effective in capturing the temporal behavior associated with a patient's EHR record where more weight is given to recent visits and less weight to older visits. For the remaining two components i.e., patient demographics and diagnosis the problem of missing value did not exist in the raw data and no measure was taken.

Merging Multiple Encounters into a Single Record For all the encounters of a single patient, we want to create a single record that contains details of all historic diagnoses, demographics, and vitals. Firstly, we merged all the three components to have single records representing a single encounter enclosing all demographics, vitals, and diagnosis details. Next for each encounter the age of the patient at the time of encounter was calculated using the Date of Birth. Next, we separated all the encounters of type 2 diabetic patients and non-type 2 diabetic patients. For every T2DM patient, all the records before the one in which T2DM was first diagnosed were merged into a single record. This single record has all the diagnoses marked as 1 which were encountered in any visit before the diagnosis of T2DM. However, the vitals and demographics of only the last record before the diagnosis of T2DM were used. A label of 1 was added to the resultant record to indicate a T2DM positive record. The details of this step are given in Figure 1.

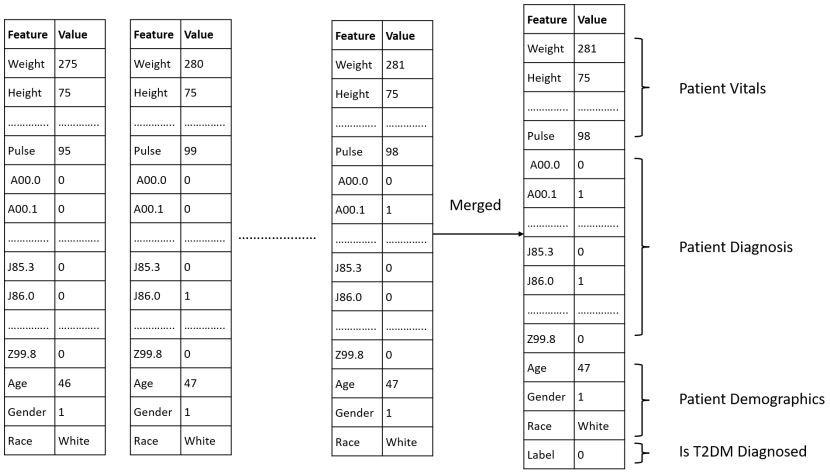

Fig. 1. Merging Encounters of a Patient.

The resultant single vector as shown in the figure is of size 29,727 (29,714 diagnosis + 9 vitals + 3 demographics + 1 for class label) per patient. The same process was carried out for creating feature vectors for non-diabetic patients with the only difference of using all encounters per patient and setting the label to 0.

Handling Class Imbalance Class imbalance is a problem that arises when the data is distributed unevenly among the classification categories [78, 11, 34]. In case of class imbalance, the performance of traditional classification techniques is affected where the performance of the techniques becomes biased towards the majority class and is reduced for the minority class [78, 48, 9]. Many

techniques have been presented over the years to deal with class imbalance including oversampling, under sampling, SMOTE, MUTE, ADASYN [9, 11, 50, 46, 28]. Class imbalance is another inherent property of health care and medicine data [79, 15, 45]. Techniques like SMOTE,ADASYNC,data over sampling,data under-sampling,etc. have been used for balancing data for T2DM prediction as well [45, 22, 15]. The EHRs data that is used as part of this study has a huge class imbalance for T2DM patients. As mentioned before out of the total of 1,145,465 unique patients only 72,626 patients were identified with T2DM. This number was further reduced to 4,896 patients after filtering out patients that have at least 3 previous encounters. To handle this class imbalance, we applied the under-sampling technique to the majority class. However, rather than just selecting a random sample naively we applied a more systematic approach. We sampled data from each practice's EHR such that the number of samples for T2DM negative patients was equal to the number of T2DM positive patients for that particular practice. Thus, maintaining a more diverse sampled T2DM negative distribution.

After applying the aforementioned preprocessing steps, we achieved a data set with 4,896 positive and 4,896 negative examples. We call this our CureMD T2DM Prediction Data (CTPD) which is used for the set of experiments and analysis discussed in the later sections. Apart from the preprocessing steps mentioned before some additional preprocessing steps including standardization (of the normally distributed variables) and normalization (of non-uniformly distributed variables) of the numeric features were performed in particular to certain machine learning techniques. Figure 2 gives an overview of the data preprocessing pipeline.

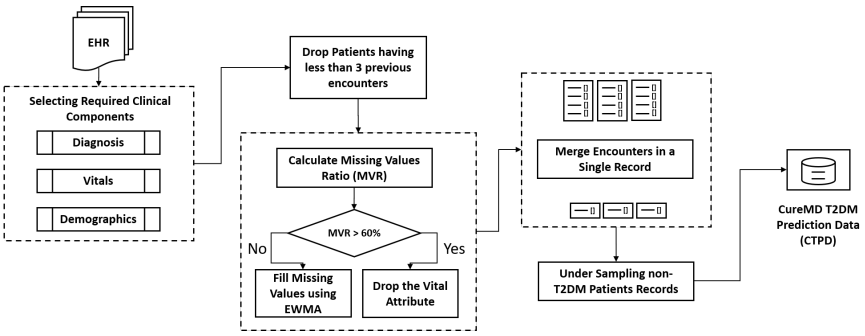

Fig. 2. Data Processing Pipeline.

2.3 Feature Selection

It is important to note that EHR data is high dimensional, sparse, and noisy [12, 84]. The dataset that is discussed in the previous section has the problem of high

dimensionality and sparsity where a total of 29,727 features are part of the final data. Out of the total features, we have 29,714 dichotomous features that represent the presence or absence of a past disease diagnosis. Such features are quite sparse and to handle such sparse raw data some feature selection process is required which can help identify relevant features that can be used for training different machine learning techniques. Many studies for T2DM prediction have used different feature selection techniques to reduce the number of features to be used for training their machine learning models [80, 62, 55]. Techniques ranging from the conventional ones including PCA,information gain,chi-square,fisher score,correlation coefficient,variance thresholding,ANOVA,etc. [80, 53, 30],to the more advanced like Boruta,MRMR,and Recursive Feature Elimination [62, 23, 8] have been used for selecting appropriate features. All such techniques have been applied to EHR and non-EHR-based datasets [19, 86, 52]. For example, Mutual Information, PCA, ANOVA, and Fischer Discriminatory Ratio were applied for feature selection to the PIMA dataset [53]. Similarly, ANOVA, Chi-Square, and Recursive Feature Elimination were applied to EHR data collected from a Korean hospital [15]. As discussed in the Data Section we collected data related to 3 major clinical components including Vitals, Diagnosis, and Demographics. Each of the components has challenges of its own and belongs to a separate data type. We applied a separate set of features selection process to each of the components to select appropriate features, details of each are given as follows:

Feature Selection from Historic Diagnosis For every patient we collected all the past diagnoses that a patient was associated with during every visit before being diagnosed with T2DM. All diagnoses are recorded as ICD-10-CM codes. The presence of a diagnosis in history is represented as 1 and the absence of a diagnosis in history as 0. The major challenge that we face here is the dimensionality and sparsity of the features we have. It is desirable to reduce the dimensionality of the data for which we implied three separate techniques i.e.,PCA,Boruta,and MRMR. Each of the three techniques was run using different hyperparameter settings and then an intersection of the features from each was taken. All the codes that were marked as important features were manually validated as well to eliminate any such codes that may be redundant or falsely identified. For example,we encountered a code Z00 marked as an important feature by all three techniques. The code itself represents 'Encounter for general examination' which has no significance in determining T2DM but is generally associated with almost all visits. Hence such a feature was dropped. A total of 347 diagnosis codes were marked important by the aforementioned process which are used in combination with the selected vitals and demographic for experiments.

Feature Selection from Vitals Next, we discuss the feature selection technique that we used to select the most important vitals. All of the information that we have for various vitals is numeric. We have the following associated vitals variables including Weight, Height, BMI, BSA, Pulse, Respiration, Temperature, Systolic Blood Pressure, and Diastolic Blood Pressure. We calculated

the variance and information gain of each feature which is given in the table. We used variance thresholding to drop the features with the least variance including BSA, Temperature, and Respiration.

Table 3. Variance of Vitals Attributes.

No.	Vital Attribute	Variance %
1	Weight	2233.41
2	Height	18.63
3	BMI	45.83
4	BSA	0.08
5	Pulse	122.15
6	Respiration	2.91
7	Temperature	0.368
8	Systolic Blood Pressure	268.56
9	Diastolic Blood Pressure	95.74

Feature Selection from Demographics We have three features that were associated with demographics including Race Category, Age, and Gender. As per many studies, all these features have been used and marked important so we selected all these in all our experiments [42, 51, 54].

2.4 Experiments

For the prediction of T2DM, a variety of machine learning techniques have been used [18, 86, 60, 10]. We picked up the 6 most used supervised machine learning techniques that have been previously used for T2DM prediction [52, 65, 37, 67] training them using the features selected:

- Naïve Bayes.
- AdaBoost.
- Gradient boosting Trees.
- Random Forests.
- Logistic Regress.
- SVM.

K-Fold Cross Validation K-Fold cross-validation has been used previously for T2DM prediction and has been able to yield improved results validating the effectiveness of the techniques used [15, 42, 26, 86]. To validate our techniques, we first used a 5-fold cross-validation trying different combination of train and test splits over the dataset. The results obtained are discussed in the following the results section.

3 Results

3.1 Evaluation Measures

Many evaluation measures have been used to mark the performance of T2DM predictive systems which include Accuracy, Precision, Recall, and F1-Score, [74, 7, 70, 26, 39, 25, 30, 53, 76, 8, 86, 52]. We used all these measures to find the performance of each of the models discussed above.

3.2 Performance of Individual Models

As stated in the experiments section 2.4 we used seven different machine learning techniques to evaluate T2DM prediction using our proposed EHR dataset. As stated, certain feature selection techniques were applied to obtain an optimal set of features which were used for training the aforementioned machine learning techniques. The performance of each of the 6 techniques was evaluated using 4 different evaluation measures. Firstly, in Table 4 we have compared the accuracy of each technique on the training and testing data. Secondly, Table 5 represents the details of performance of each technique on test set alone against the 4 evaluation measures including accuracy,precision,recall,and F1-score.

Table 4. Training and Testing Accuracy of Machine Learning Techniques for T2DM Prediction.

Machine Learning Technique	Testing Accuracy %	Training Accuracy %
Naïve Bayes	64.1	65.8
Ada Boost	84.9	86.1
Gradient Boosted Trees	84.3	86.6
Random Forest	82.1	95.6
Logistic Regression	85.5	88.7
SVM	86.7	91.2

Table 5. Performance of Machine Learning Techniques for T2DM Prediction.

Machine Learning Technique	Accuracy %	Precision %	Recall %	F-Score %
Naïve Bayes	64.1	60.2	98.5	76.1
Ada Boost	84.9	85.2	85.1	85.1
Gradient Boosted Trees	84.3	82.6	87.9	85.2
Random Forest	82.1	81.7	86.1	83.8
Logistic Regression	85.5	85.6	87.5	86.5
SVM	86.7	84.9	90.7	87.7

From Tables 4 and 5 it can be observed that in terms of accuracy SVM performs best. SVM also has the highest recall and F-score. It is also important

to note that the Logistic Regression performs quite as well as SVM and has a smaller difference between the training and testing accuracies i.e., less overfitted to the training data.

3.3 Results of K-Fold Cross Validation

To further validate the performance of each technique we applied 5-fold cross-validation as discussed in the experiments section. Firstly, in Tables 6 and 7 we have presented the training and testing accuracy achieved using the 5-fold cross-validation along with the mean accuracies.

Table 6. Training Accuracy for T2DM Prediction using 5-Fold CV.

Machine Learning Technique	Accuracy Per Fold %	Mean %
Naïve Bayes	64.80 68.70 62.90 64.08 62.60	64.62
Ada Boost	86.90 86.90 86.90 86.90 86.50	86.82
Gradient Boosted Trees	86.10 86.40 86.60 86.50 85.90	86.30
Random Forest	95.60 95.70 95.70 95.70 95.20	95.58
Logistic Regression	88.60 88.80 88.70 88.40 88.10	88.52
SVM	91.50 91.30 91.30 91.10 90.90	91.22

Table 7. Testing Accuracy for T2DM Prediction using 5-Fold CV.

Machine Learning Technique	Accuracy Per Fold %	Mean %
Naïve Bayes	62.80 67.40 63.04 65.20 61.70	64.03
Ada Boost	85.30 85.10 85.70 84.90 86.90	85.58
Gradient Boosted Trees	84.30 84.30 83.80 84.50 85.40	84.46
Random Forest	83.10 82.70 83.30 84.70 82.90	83.34
Logistic Regression	85.60 85.50 86.30 86.03 87.40	86.17
SVM	86.03 86.20 86.20 86.03 87.30	86.35

Next Tables 8, 9 and 10 present the precision, recall and F1-score obtained for each fold on the test set along with the mean value for each.

Tables 6, 7, 8, 9,and 10 validate that SVM performs best for T2DM prediction along with Logistic regression that has almost identical results to SVM. Do note that a higher recall can be observed for Naïve Bayes however, it performs consistently poor for other evaluation measures.

3.4 Evaluating Fairness of the Techniques

It is desirable to have a machine learning model for T2DM prediction that is able to correctly predict for a diverse set of unseen examples. The motivation

Table 8. Precision for T2DM Prediction using 5-Fold CV.

Machine Learning Technique	Precision Per Fold %	Mean %
Naïve Bayes	56.90 60.60 57.90 59.70 56.70	58.36
Ada Boost	85.50 84.20 84.80 85.60 86.30	85.28
Gradient Boosted Trees	81.50 81.80 82.50 83.10 83.30	82.44
Random Forest	80.99 80.70 81.50 83.30 81.20	81.54
Logistic Regression	85.02 84.20 85.10 86.60 86.10	85.40
SVM	83.90 83.60 84.06 85.40 84.70	84.33

Table 9. Recall for T2DM Prediction using 5-Fold CV.

Machine Learning Technique	Recall Per Fold %	Mean %
Naïve Bayes	98.60 98.10 98.40 97.80 96.80	97.94
Ada Boost	84.10 85.90 87.60 84.60 87.60	85.96
Gradient Boosted Trees	87.70 87.80 86.40 87.30 88.50	87.54
Random Forest	85.60 85.50 86.80 87.30 85.40	86.12
Logistic Regression	85.70 87.10 88.60 85.90 88.80	87.22
SVM	88.40 89.90 89.70 87.60 91.00	89.32

Table 10. F1-Score for T2DM Prediction using 5-Fold CV.

Machine Learning Technique	F1-Score Per Fold %	Mean %
Naïve Bayes	72.20 74.90 72.90 74.20 71.60	73.16
Ada Boost	84.80 85.10 86.20 85.10 86.90	85.62
Gradient Boosted Trees	84.50 84.70 84.40 85.20 85.80	84.92
Random Forest	83.20 83.10 84.10 85.30 83.20	83.78
Logistic Regression	85.40 85.70 86.80 86.20 87.50	86.32
SVM	86.10 86.60 86.80 86.50 87.70	86.74

behind collecting data from different EHRs, running in different parts of the US was to serve the same purpose of having a diverse dataset that can help our machine learning techniques to generalize better. In order to validate that our models are able to learn without much bias towards any certain group we perform the following analysis to evaluate the fairness of each of the techniques used. Firstly we divide the entire dataset based on certain patient groups. These groups involved patient being divided based on their age, gender and race. For each of this division we evaluated all the techniques to find if there exists any bias towards any certain group. The details for fairness of each technique evaluated using the accuracy for different groups is given in the Table 11.

Table 11. Evaluation of Fairness of T2DM Prediction Techniques.

Group	Group Value	Naïve Bayes	Ada Boost	Gradient Boosted Trees	Random Forest	Logistic Regression	SVM
Age	Over80	54.1	75	78.1	76.2	77.4	78.6
	Middle Aged	58	82.9	82.5	82.7	86.7	84.8
	Aged	61.1	76.9	77.4	77.4	76.9	79.3
	Adult	75.3	88.7	86.6	86.3	89.9	91
	Adolescent	N/A	N/A	N/A	N/A	N/A	N/A
	Child	N/A	N/A	N/A	N/A	N/A	N/A
Race	Asian	60	80	80.9	81.8	83.6	79.1
	Black	62.8	83.9	82.6	88.5	85.3	89.9
	White	71.8	84.1	85.1	84.6	86.4	86.9
	Native American	72.5	70	67.5	72.5	85	85
	Bi Racial	54.5	72.2	73.7	72.7	81.8	70.7
	Latino	N/A	N/A	N/A	N/A	N/A	N/A
Gender	Males	60.2	81.6	81.3	83	82.1	83.4
	Females	75.2	83	83.7	83.7	83.8	85.3

Note for age we created age groups by grouping ages based on the ranges specified in the Table 12. Also, note in Table 11 N/A represent the group for which we have significantly fewer examples points for which the fairness was not evaluated.

Table 12. Age Group Ranges.

No.	Age Group	Age Range
Child	13 or lower	
Adolescent	13 to 19	
Adult	19 to 45	
Middle Aged	45 to 65	
Aged	65 to 80	
Over 80	80 and above.	

4 Conclusion and Future Work

In this study,we proposed a novel T2DM prediction dataset collected from different EHRs which is the most diverse dataset for T2DM. The dataset consists of 4,896, type 2 diabetic and 4,896 non-type 2 diabetic patients and comes from 295 different EHR systems running in different hospitals and practices around the US. We also extracted novel features from the proposed EHR dataset for the prediction of T2DM that rely only on routine EHR data and do not require some complicated lab procedure or result. The features based on a patient's history of diagnosis, vitals, and demographics were then used to train and evaluate 6 different machine learning techniques achieving a best accuracy of 86.7% for early detection of T2DM. We further validated our techniques using 4 different evaluation measures and a 5-fold cross-validation. We analyzed the fairness of the techniques presented for different user groups based on age, race, and gender. We further aim to extend this study by proposing novel ensemble machine learning and deep learning technique to improve the overall performance of T2DM prediction. We also aim to apply the same methodology for the early diagnosis of other diseases including hypertension, cardiovascular disease, etc.

References

1. Centers for disease control and prevention. national diabetes statistics report website. https://www.cdc.gov/diabetes/data/statistics-report/index.html
2. Types of diabetes mellitus. https://www.webmd.com/diabetes/types-of-diabetes-mellitus
3. Abhari, S., Kalhori, S.R.N., Ebrahimi, M., Hasannejadasl, H., Garavand, A.: Artificial intelligence applications in type 2 diabetes mellitus care: focus on machine learning methods. Healthcare informatics research **25**(4), 248–261 (2019)
4. Alhassan, Z., McGough, A.S., Alshammari, R., Daghstani, T., Budgen, D., Al Moubayed, N.: Type-2 diabetes mellitus diagnosis from time series clinical data using deep learning models. In: International Conference on Artificial Neural Networks. pp. 468–478. Springer (2018)
5. Association, A.D., et al.: Standards of medical care in diabetes—2015 abridged for primary care providers. Clinical diabetes: a publication of the American Diabetes Association **33**(2), 97 (2015)
6. Ayon, S.I., Islam, M.M.: Diabetes prediction: a deep learning approach. International Journal of Information Engineering and Electronic Business **12**(2), 21 (2019)
7. Battineni, G., Sagaro, G.G., Nalini, C., Amenta, F., Tayebati, S.K.: Comparative machine-learning approach: a follow-up study on type 2 diabetes predictions by cross-validation methods. Machines **7**(4), 74 (2019)
8. Birjais, R., Mourya, A.K., Chauhan, R., Kaur, H.: Prediction and diagnosis of future diabetes risk: a machine learning approach. SN Applied Sciences **1**(9), 1–8 (2019)
9. Bunkhumpornpat, C., Sinapiromsaran, K., Lursinsap, C.: Mute: Majority undersampling technique. In: 2011 8th International Conference on Information, Communications & Signal Processing. pp. 1–4. IEEE (2011)

10. Cahn, A., Shoshan, A., Sagiv, T., Yesharim, R., Goshen, R., Shalev, V., Raz, I.: Prediction of progression from pre-diabetes to diabetes: Development and validation of a machine learning model. Diabetes/metabolism research and reviews **36**(2), e3252 (2020)
11. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)
12. Cheng, Y., Wang, F., Zhang, P., Hu, J.: Risk prediction with electronic health records: A deep learning approach. In: Proceedings of the 2016 SIAM international conference on data mining. pp. 432–440. SIAM (2016)
13. Collins, G.S., Mallett, S., Omar, O., Yu, L.M.: Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. BMC medicine **9**(1), 1–14 (2011)
14. Dagliati, A., Marini, S., Sacchi, L., Cogni, G., Teliti, M., Tibollo, V., De Cata, P., Chiovato, L., Bellazzi, R.: Machine learning methods to predict diabetes complications. Journal of diabetes science and technology **12**(2), 295–302 (2018)
15. Deberneh, H.M., Kim, I.: Prediction of type 2 diabetes based on machine learning algorithm. International journal of environmental research and public health **18**(6), 3317 (2021)
16. Ducat, L., Philipson, L.H., Anderson, B.J.: The mental health comorbidities of diabetes. Jama **312**(7), 691–692 (2014)
17. El_Jerjawi, N.S., Abu-Naser, S.S.: Diabetes prediction using artificial neural network. International Journal of Advanced Science and Technology **121** (2018)
18. Farran, B., Channanath, A.M., Behbehani, K., Thanaraj, T.A.: Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: machine-learning algorithms and validation using national health data from kuwait—a cohort study. BMJ open **3**(5), e002457 (2013)
19. Fazakis, N., Kocsis, O., Dritsas, E., Alexiou, S., Fakotakis, N., Moustakas, K.: Machine learning tools for long-term type 2 diabetes risk prediction. IEEE Access **9**, 103737–103757 (2021)
20. Federation, I.: Idf diabetes atlas tenth edition 2021. international diabetes federation. idf diabetes atlas, 10th edn. brussels, belgium: International diabetes federation; 2021 (2021)
21. Franciosi, M., De Berardis, G., Rossi, M.C., Sacco, M., Belfiglio, M., Pellegrini, F., Tognoni, G., Valentini, M., Nicolucci, A., Group, I.S.: Use of the diabetes risk score for opportunistic screening of undiagnosed diabetes and impaired glucose tolerance: the igloo (impaired glucose tolerance and long-term outcomes observational) study. Diabetes care **28**(5), 1187–1194 (2005)
22. García-Ordás, M.T., Benavides, C., Benítez-Andrades, J.A., Alaiz-Moretón, H., García-Rodríguez, I.: Diabetes detection using deep learning techniques with oversampling and feature augmentation. Computer Methods and Programs in Biomedicine **202**, 105968 (2021)
23. Ghosh, P., Azam, S., Karim, A., Hassan, M., Roy, K., Jonkman, M.: A comparative study of different machine learning tools in detecting diabetes. Procedia Computer Science **192**, 467–477 (2021)
24. Group, D.P.P.R., et al.: Long-term effects of lifestyle intervention or metformin on diabetes development and microvascular complications over 15-year follow-up: the diabetes prevention program outcomes study. The lancet Diabetes & endocrinology **3**(11), 866–875 (2015)

25. Haq, A.U., Li, J.P., Khan, J., Memon, M.H., Nazir, S., Ahmad, S., Khan, G.A., Ali, A.: Intelligent machine learning approach for effective recognition of diabetes in e-healthcare using clinical data. Sensors **20**(9), 2649 (2020)
26. Hasan, M.K., Alam, M.A., Das, D., Hossain, E., Hasan, M.: Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access **8**, 76516–76531 (2020)
27. Hassing, L.B., Hofer, S.M., Nilsson, S.E., Berg, S., Pedersen, N.L., McClearn, G., Johansson, B.: Comorbid type 2 diabetes mellitus and hypertension exacerbates cognitive decline: evidence from a longitudinal study. Age and ageing **33**(4), 355–361 (2004)
28. He, H., Bai, Y., Garcia, E.A., Li, S.: Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence). pp. 1322–1328. IEEE (2008)
29. Hossain, M.E., Uddin, S., Khan, A., Moni, M.A.: A framework to understand the progression of cardiovascular disease for type 2 diabetes mellitus patients using a network approach. International Journal of Environmental Research and Public Health **17**(2), 596 (2020)
30. Howlader, K.C., Satu, M., Awal, M., Islam, M., Islam, S.M.S., Quinn, J.M., Moni, M.A., et al.: Machine learning models for classification and identification of significant attributes to detect type 2 diabetes. Health information science and systems **10**(1), 1–13 (2022)
31. Hripcsak, G., Albers, D.J.: Next-generation phenotyping of electronic health records. Journal of the American Medical Informatics Association **20**(1), 117–121 (2013)
32. Islam, M.T., Al-Absi, H.R., Ruagh, E.A., Alam, T.: Dianet: A deep learning based architecture to diagnose diabetes using retinal images only. IEEE Access **9**, 15686–15695 (2021)
33. Ismail, L., Materwala, H., Tayefi, M., Ngo, P., Karduck, A.P.: Type 2 diabetes with artificial intelligence machine learning: methods and evaluation. Archives of Computational Methods in Engineering **29**(1), 313–333 (2022)
34. Japkowicz, N.: The class imbalance problem: Significance and strategies. In: Proc. of the Int'l Conf. on artificial intelligence. vol. 56, pp. 111–117. Citeseer (2000)
35. Jayanthi, N., Babu, B.V., Rao, N.S.: Survey on clinical prediction models for diabetes prediction. Journal of Big Data **4**(1), 1–15 (2017)
36. Johnson, A.E., Pollard, T.J., Shen, L., Lehman, L.w.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., Mark, R.G.: Mimic-iii, a freely accessible critical care database. Scientific data **3**(1), 1–9 (2016)
37. Joshi, R.D., Dhakal, C.K.: Predicting type 2 diabetes using logistic regression and machine learning approaches. International journal of environmental research and public health **18**(14), 7346 (2021)
38. Joshi, T.N., Chawan, P., et al.: Diabetes prediction using machine learning techniques. Ijera **8**(1), 9–13 (2018)
39. Kahramanli, H., Allahverdi, N.: Design of a hybrid system for the diabetes and heart diseases. Expert systems with applications **35**(1-2), 82–89 (2008)
40. Kalagotla, S.K., Gangashetty, S.V., Giridhar, K.: A novel stacking technique for prediction of diabetes. Computers in Biology and Medicine **135**, 104554 (2021)
41. Kaur, G., Chhabra, A.: Improved j48 classification algorithm for the prediction of diabetes. International journal of computer applications **98**(22) (2014)

42. Kayaer, K., Yildirim, T., et al.: Medical diagnosis on pima indian diabetes using general regression neural networks. In: Proceedings of the international conference on artificial neural networks and neural information processing (ICANN/ICONIP). vol. 181, p. 184 (2003)
43. Kim, H.S., Shin, A.M., Kim, M.K., Kim, Y.N.: Comorbidity study on type 2 diabetes mellitus using data mining. The Korean journal of internal medicine **27**(2), 197 (2012)
44. Kopitar, L., Kocbek, P., Cilar, L., Sheikh, A., Stiglic, G.: Early detection of type 2 diabetes mellitus using machine learning-based prediction models. Scientific reports **10**(1), 1–12 (2020)
45. Krasteva, A., Panov, V., Krasteva, A., Kisselova, A., Krastev, Z.: Oral cavity and systemic diseases—diabetes mellitus. Biotechnology & Biotechnological Equipment **25**(1), 2183–2186 (2011)
46. Kubat, M., Matwin, S.: Addressing the curse of imbalanced data sets: One-sided sampling. In: Proceedings of the fourteenth international conference on machine learning. pp. 179–186 (1997)
47. Lai, H., Huang, H., Keshavjee, K., Guergachi, A., Gao, X.: Predictive models for diabetes mellitus using machine learning techniques. BMC endocrine disorders **19**(1), 1–9 (2019)
48. Lakshmi, T.J., Prasad, C.S.R.: A study on classifying imbalanced datasets. In: 2014 First international conference on networks & soft computing (ICNSC2014). pp. 141–145. IEEE (2014)
49. Lee, B.J., Kim, J.Y.: Identification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on machine learning. IEEE journal of biomedical and health informatics **20**(1), 39–46 (2015)
50. Ling, C.X., Li, C.: Data mining for direct marketing: Problems and solutions. In: Kdd. vol. 98, pp. 73–79 (1998)
51. Lu, H., Uddin, S., Hajati, F., Moni, M.A., Khushi, M.: A patient network-based machine learning model for disease prediction: The case of type 2 diabetes mellitus. Applied Intelligence **52**(3), 2411–2422 (2022)
52. Mani, S., Chen, Y., Elasy, T., Clayton, W., Denny, J.: Type 2 diabetes risk forecasting from emr data using machine learning. In: AMIA annual symposium proceedings. vol. 2012, p. 606. American Medical Informatics Association (2012)
53. Maniruzzaman, M., Rahman, M., Al-MehediHasan, M., Suri, H.S., Abedin, M., El-Baz, A., Suri, J.S., et al.: Accurate diabetes risk stratification using machine learning: role of missing value and outliers. Journal of medical systems **42**(5), 1–17 (2018)
54. Mujumdar, A., Vaidehi, V.: Diabetes prediction using machine learning algorithms. Procedia Computer Science **165**, 292–299 (2019)
55. Nadesh, R.K., Arivuselvan, K., et al.: Type 2: diabetes mellitus prediction using deep neural networks classifier. International Journal of Cognitive Computing in Engineering **1**, 55–61 (2020)
56. Naz, H., Ahuja, S.: Deep learning approach for diabetes prediction using pima indian dataset. Journal of Diabetes & Metabolic Disorders **19**(1), 391–403 (2020)
57. NCfH, S.: About the national health and nutrition examination survey (2017)
58. Organization, W.H., et al.: Diabetes Mellitus: Report of a WHO Study Group [meeting held in Geneva from 11 to 16 February 1985]. World Health Organization (1985)
59. Pathak, J., Kho, A.N., Denny, J.C.: Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. Journal of the American Medical Informatics Association **20**(e2), e206–e211 (2013)

60. Prema, N., Varshith, V., Yogeswar, J.: Prediction of diabetes using ensemble techniques. Int. J. Recent Technol. Eng **7**(6), 203–205 (2019)
61. Punthakee, Z., Goldenberg, R., Katz, P.: Definition, classification and diagnosis of diabetes, prediabetes and metabolic syndrome. Canadian journal of diabetes **42**, S10–S15 (2018)
62. Rahman, M., Islam, D., Mukti, R.J., Saha, I.: A deep learning approach based on convolutional lstm for detecting diabetes. Computational biology and chemistry **88**, 107329 (2020)
63. Ramachandran, A.: Know the signs and symptoms of diabetes. The Indian journal of medical research **140**(5), 579 (2014)
64. Rashidian, S., Abell-Hart, K., Hajagos, J., Moffitt, R., Lingam, V., Garcia, V., Tsai, C.W., Wang, F., Dong, X., Sun, S., et al.: Detecting miscoded diabetes diagnosis codes in electronic health records for quality improvement: temporal deep learning approach. JMIR medical informatics **8**(12), e22649 (2020)
65. Robertson, G., Lehmann, E.D., Sandham, W., Hamilton, D.: Blood glucose prediction using artificial neural networks trained with the aida diabetes simulator: a proof-of-concept pilot study. Journal of Electrical and Computer Engineering **2011** (2011)
66. Samant, P., Agarwal, R.: Machine learning techniques for medical diagnosis of diabetes using iris images. Computer methods and programs in biomedicine **157**, 121–128 (2018)
67. Sarwar, M.A., Kamal, N., Hamid, W., Shah, M.A.: Prediction of diabetes using machine learning algorithms in healthcare. In: 2018 24th international conference on automation and computing (ICAC). pp. 1–6. IEEE (2018)
68. Shivade, C., Raghavan, P., Fosler-Lussier, E., Embi, P.J., Elhadad, N., Johnson, S.B., Lai, A.M.: A review of approaches to identifying patient phenotype cohorts using electronic health records. Journal of the American Medical Informatics Association **21**(2), 221–230 (2014)
69. Smith, J.W., Everhart, J.E., Dickson, W., Knowler, W.C., Johannes, R.S.: Using the adap learning algorithm to forecast the onset of diabetes mellitus. In: Proceedings of the annual symposium on computer application in medical care. p. 261. American Medical Informatics Association (1988)
70. Sonar, P., JayaMalini, K.: Diabetes prediction using different machine learning approaches. In: 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC). pp. 367–371. IEEE (2019)
71. Strack, B., DeShazo, J.P., Gennings, C., Olmo, J.L., Ventura, S., Cios, K.J., Clore, J.N.: Impact of hba1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. BioMed research international **2014** (2014)
72. Swapna, G., Vinayakumar, R., Soman, K.: Diabetes detection using deep learning algorithms. ICT express **4**(4), 243–246 (2018)
73. Swapna, G., Kp, S., Vinayakumar, R.: Automated detection of diabetes using cnn and cnn-lstm network and heart rate signals. Procedia computer science **132**, 1253–1262 (2018)
74. Taz, N.H., Islam, A., Mahmud, I.: A comparative analysis of ensemble based machine learning techniques for diabetes identification. In: 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST). pp. 1–6. IEEE (2021)
75. Temurtas, H., Yumusak, N., Temurtas, F.: A comparative study on diabetes disease diagnosis using neural networks. Expert Systems with applications **36**(4), 8610–8615 (2009)

76. Tigga, N.P., Garg, S.: Prediction of type 2 diabetes using machine learning classification methods. Procedia Computer Science **167**, 706–716 (2020)
77. Viloria, A., Herazo-Beltran, Y., Cabrera, D., Pineda, O.B.: Diabetes diagnostic prediction using vector support machines. Procedia Computer Science **170**, 376–381 (2020)
78. Vluymans, S.: Learning from imbalanced data. In: Dealing with Imbalanced and Weakly Labelled Data in Machine Learning using Fuzzy and Rough Set Methods, pp. 81–110. Springer (2019)
79. Wang, Q., Cao, W., Guo, J., Ren, J., Cheng, Y., Davis, D.N.: Dmp_mi: an effective diabetes mellitus classification algorithm on imbalanced data with missing values. IEEE Access **7**, 102232–102238 (2019)
80. Wei, S., Zhao, X., Miao, C.: A comprehensive exploration to the machine learning techniques for diabetes identification. In: 2018 IEEE 4th World Forum on Internet of Things (WF-IoT). pp. 291–295. IEEE (2018)
81. Xiong, X.l., Zhang, R.x., Bi, Y., Zhou, W.h., Yu, Y., Zhu, D.l.: Machine learning models in type 2 diabetes risk prediction: results from a cross-sectional retrospective study in chinese adults. Current medical science **39**(4), 582–588 (2019)
82. Yahyaoui, A., Jamil, A., Rasheed, J., Yesiltepe, M.: A decision support system for diabetes prediction using machine learning and deep learning techniques. In: 2019 1st International Informatics and Software Engineering Conference (UBMYK). pp. 1–4. IEEE (2019)
83. Yu, W., Liu, T., Valdez, R., Gwinn, M., Khoury, M.J.: Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC medical informatics and decision making **10**(1), 1–7 (2010)
84. Zheng, T., Xie, W., Xu, L., He, X., Zhang, Y., You, M., Yang, G., Chen, Y.: A machine learning-based framework to identify type 2 diabetes through electronic health records. International journal of medical informatics **97**, 120–127 (2017)
85. Zhu, T., Li, K., Herrero, P., Georgiou, P.: Deep learning for diabetes: a systematic review. IEEE Journal of Biomedical and Health Informatics **25**(7), 2744–2757 (2020)
86. Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., Tang, H.: Predicting diabetes mellitus with machine learning techniques. Frontiers in genetics **9**, 515 (2018)