Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-19766-9

Machine and deep learning techniques for the prediction of diabetics: a review

Sandip Kumar Singh Modak[1] · Vijay Kumar Jha[2]

Received: 25 October 2023 / Revised: 27 February 2024 / Accepted: 20 June 2024

Abstract

Diabetes has become one of the significant reasons for public sickness and death in worldwide. By 2019, diabetes had affected more than 463 million people worldwide. According to the International Diabetes Federation report, this figure is expected to rise to more than 700 million in 2040, so early screening and diagnosis of diabetes patients have great significance in detecting and treating diabetes on time. Diabetes is a multi factorial metabolic disease, its diagnostic criteria are difficult to cover all the ethology, damage degree, pathogenesis and other factors, so there is a situation for uncertainty and imprecision under various aspects of the medical diagnosis process. With the development of Data mining, researchers find that machine learning and deep learning, playing an important role in diabetes prediction research. This paper is an in-depth study on the application of machine learning and deep learning techniques in the prediction of diabetics. In addition, this paper also discusses the different methodology used in machine and deep learning for prediction of diabetics since last two decades and examines the methods used, to explore their successes and failure. This review would help researchers and practitioners understand the current state-of-the-art methods and identify gaps in the literature.

Keywords Diabetics · Data Mining · Machine learning · Deep learning

1 Introduction

According to the International Diabetes Federation (IDF) [1] statistics, there were 415 million people suffering from diabetes around the world in 2015. By 2040 this number is expected to rise to over 642 million, as a consequence, diabetes has become the main cause of national disease and death in most countries. Diabetes is a group of metabolic diseases

✉ Sandip Kumar Singh Modak
sandip.modak@sbu.ac.in

Vijay Kumar Jha
vkjha@bitmesra.ac.in

[1] Department of Computer Science & Engineering, Sarala Birla University, Ranchi, Jharkhand, India

[2] Department of Computer Science & Engineering, Birla Institute of Technology, Mesra, Ranchi, India

Published online: 16 July 2024

in which a person has high blood glucose, either because the body does not produce enough insulin, or because the cells do not respond to the insulin that is produced [2, 3]. Most diabetes can be categorized into 3 subgroups: type 1 diabetes (T1D), type 2 diabetes (T2D), and gestational diabetes (GDM). Over the long term, T2D patients become resistant to the normal effects of insulin and gradually lose their capacity to produce enough of this hormone. A wide range of therapeutic options is available for patients with T2D. At the early stages of disease, they commonly receive medications that improve insulin secretion or insulin absorption, but eventually they must receive external doses of insulin. On the other hand, T1D patients have severe impairments in insulin production, and must use external insulin exclusively to manage their blood glucose (BG). Treatment of T1D requires consistent doses of insulin through multiple daily injections (MDIs) or continuous subcutaneous insulin infusion (CSII) using a pump. GDM is treated similarly to T2D, but only occurs during pregnancy due to the interaction between insulin and hormones released by the placenta. Figure 1 represents the statistical data of diabetics' patients from the year 2000 onwards.

In 2000, the global estimate of adults living with diabetes was 151 million. By 2009 it had grown by 88% to 285 million. Today, 9.3% of adults aged 20–79 years – a staggering 463 million people – are living with diabetes. A further 1.1 million children and adolescents under the age of 20, live with type 1 diabetes. A decade ago, in 2010, the global projection for diabetes in 2025 was 438 million. With over five years still to go, that prediction has already been surpassed by 25 million. IDF (International Diabetes Federation) estimates that there will be 578 million adults with diabetes by 2030, and 700 million by 2045. Diabetes is one of the deadliest diseases that claim millions of lives each year. According to the WHO (World Health Organization), it was estimated that 3.4 million deaths are caused due to high blood sugar. It has been found that the over diagnosis of diabetes may lead to comorbidity like cognitive impairment, stroke, cancer, kidney problem etc. Therefore, it should be diagnosed at the earliest. In year 2000, India topped the world with 31.7 million people suffered from diabetes followed

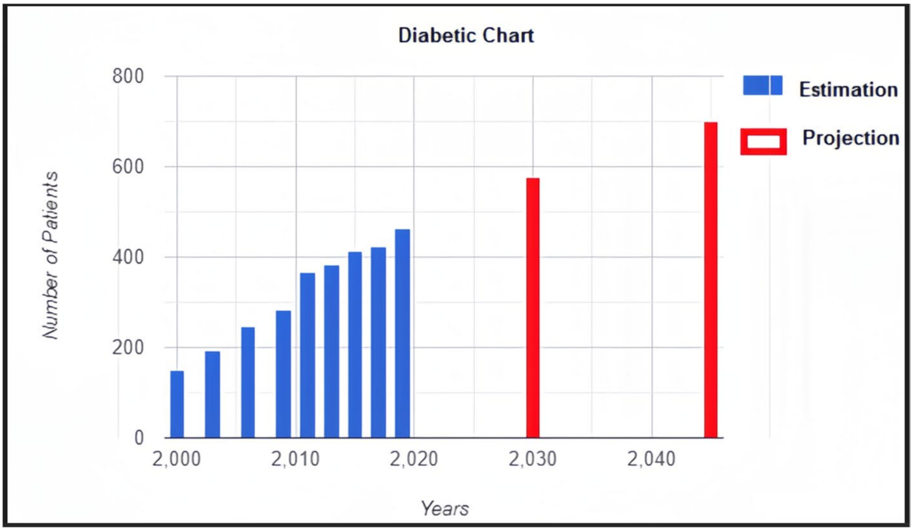


Fig. 1 Population of diabetic patients

by China with second place and United States with third place [4]. It is predicted that by the year 2030 diabetes mellitus may affect up to 79.40 million people in India [5]. In last 40 years, a fourfold rise has been witnessed for this contagious disease [6]. According to International Diabetes Federation, in 2017, there are around 425 million populations suffering from diabetes across the world. It is also estimated that by 2045 the rise in the diabetic population will be increased by 32% [7]. Currently, China, India, USA, Brazil, and Russia are the top five countries with the highest rate of diabetic population. Figure 2 [8] shows the percentage of people affected by diabetes.

Data Mining and Artificial Intelligence (AI) plays an important role in the prediction of diabetes. With the continuous development of artificial intelligence and data mining technology, researchers begin to consider using machine learning and deep learning techniques to search for the characteristics of diabetes. Machine learning techniques can find implied pathogenic factors in virtue of analyzing and using diabetic data, with a high stability and accuracy in diabetes diagnosis. Therefore, machine learning techniques which can find out the reasonable threshold risk factors and physiological parameters provide new ideas for screening and diagnosis of diabetes [9]. Diabetes is a very serious disease that, if not treated properly and on time, can lead to very serious complications, including death. This makes diabetes, one of the main priorities in medical science research, which in turn generates huge amounts of data. Constantly increasing volumes of data are very well suited to be processed using data mining that can readily handle them. Using data-mining methods in diabetes research is one of the best ways to utilize large volumes of available diabetes-related data for extracting knowledge. Both descriptive (association and clustering) and predictive (classification) data-mining methods are used in the process. These data-mining methods are different from traditional statistic approaches in many ways [10].

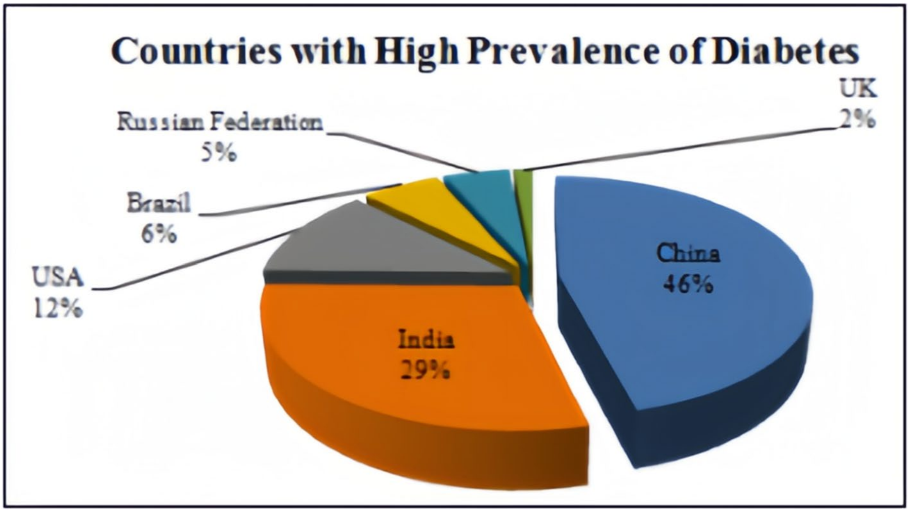


Fig. 2 Worldwide population of diabetic patients

1.1 Machine learning/deep learning and its application for diabetic prediction

Machine learning and deep learning techniques hold great promise in improving the early detection and management of diabetes, potentially leading to better patient outcomes and reduced healthcare costs.

- Data Collection and Preprocessing: The first step involves collecting relevant data from patients, which may include demographic information (age, gender), medical history (family history of diabetes, past diagnoses), lifestyle factors (diet, exercise), and clinical measurements (blood glucose levels, blood pressure, cholesterol levels). This data needs to be preprocessed to handle missing values, normalize features, and remove noise.
- Feature Selection and Engineering: Feature selection involves identifying the most relevant variables that contribute to the prediction of diabetes. Feature engineering may involve creating new features from existing ones or transforming the data to improve model performance.
- Model Selection: Various machine learning algorithms can be employed for diabetic prediction, including logistic regression, decision trees, random forests, support vector machines (SVM), and neural networks. Deep learning techniques, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), can also be used to capture complex patterns in the data.
- Model Training and Evaluation: The selected model is trained using labeled data, where the outcome (diabetes status) is known. The model's performance is evaluated using metrics such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC) on a separate validation dataset.
- Deployment and Integration: Once the model is trained and evaluated, it can be deployed in clinical settings to assist healthcare providers in identifying individuals at risk of diabetes. Integration with electronic health record (EHR) systems can facilitate real-time prediction and decision-making.
- Continuous Monitoring and Updating: As new data becomes available, the model should be periodically retrained and updated to ensure its accuracy and relevance over time. Continuous monitoring of model performance and outcomes can help improve its effectiveness in predicting diabetes and related complications.

Some challenges and considerations in diabetic prediction using machine learning and deep learning include the need for large and diverse datasets, addressing class imbalance (since the number of diabetic patients may be much smaller than non-diabetic patients), interpretability of models, and ensuring privacy and security of patient data. Healthcare data, including patient information and medical records, are sensitive and subject to strict privacy regulations such as HIPAA (Health Insurance Portability and Accountability Act) in the United States. Ensuring the privacy and security of patient data is essential when developing and deploying diabetic prediction models. Integrating ML and DL models into clinical practice requires rigorous validation to demonstrate their effectiveness and safety. The recent proliferation of data mining techniques has given rise to disease prediction systems. Specifically, with the vast amount of medical data generated every day [440].

The remainder of this paper is structured as follows. Section 2 highlights the details of different data mining technique utilized in prediction of diabetes; Section 3 mainly focuses on a detail review of diabetes prediction based on machine learning; Section 4 mainly focuses on a detail review of diabetes prediction based on deep learning; Section 5 mainly for discussion and comparission; and finally concludes the papers in section 6.

2 Data mining techniques

Both Data Mining and Machine learning are areas which have been inspired by each other, though they have many things in common, yet they have different ends. Figure 3 represents the relationship between the machine learning and data mining. Data mining, machine learning, artificial intelligence, and statistics are closely related fields that share common goals and methodologies for analyzing and extracting insights from data. They complement each other and are often used together in various applications, such as predictive modeling, pattern recognition, data visualization, and decision support.

Artificial Intelligence can enable the computer to think. The computer is made much more intelligent by AI. Machine learning is the subfield of AI study. Various researchers think that without learning, intelligence cannot be developed. There are many types of Machine Learning Techniques that are shown in Fig. 4. Supervised, Unsupervised, Semi Supervised, Reinforcement, Evolutionary Learning and Deep Learning are the types of machine learning techniques. These techniques are used to classify the data set [11]. Both supervised and unsupervised learning techniques are used depending on the nature of the data and the specific problem being addressed. Additionally, semi-supervised learning techniques combine elements of both paradigms by leveraging a small amount of labeled data along with a larger pool of unlabeled data. Reinforcement learning is a powerful

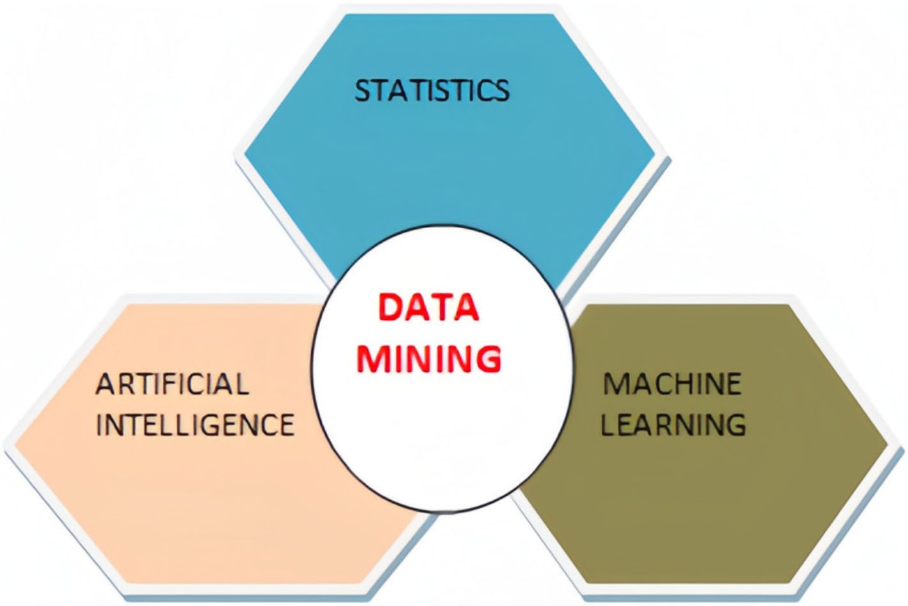


Fig. 3 Relationship of Data Mining with Machine learning

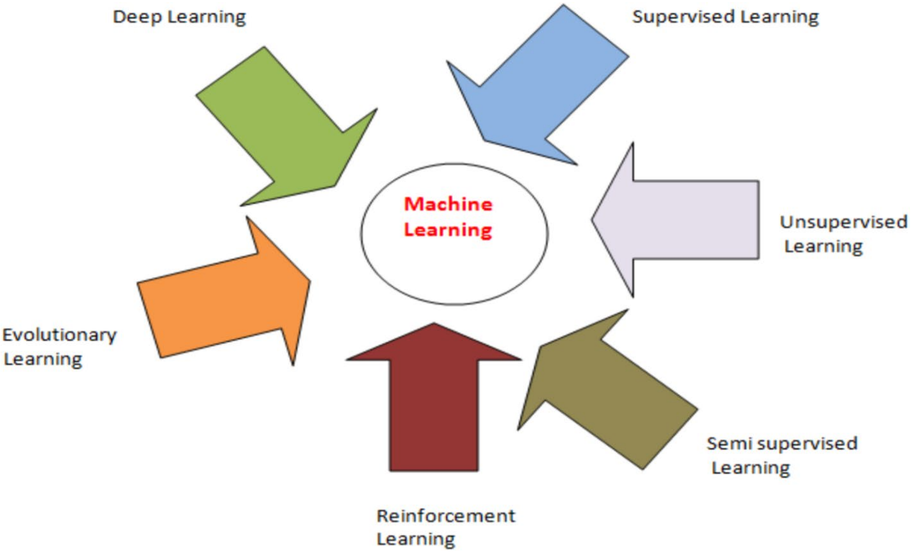


Fig. 4 Different techniques of Machine learning

paradigm in machine learning that enables agents to learn optimal behavior through trial and error, interaction with the environment, and feedback in the form of rewards.

2.1 Supervised learning

A supervised learning technique is used when the historical data is available for a certain problem. The system is trained with the inputs and respective responses and then used for the prediction of the response of new data [12]. Classification and regression are the types of Supervised Learning.

2.1.1 Classification

It gives the prediction of Yes or No, for example, “Is this tumour cancerous?”, and “Does this cookie meet our quality standards?” Common classification approaches include artificial neural network, back propagation, decision tree, support vector machines, Naive Bayes classifier, K-Nearest Neighbors (K-NN), Random forest [12]. Classification is used to classify data into predefined categorical class labels. “Class” in classification, is the attribute or feature in a data set, in which users are most interested. It is defined as the dependent variable in statistics. To classify data (or records), a classification algorithm creates a classification model consisting of classification rules. For example, banks have constructed classification models to categorize the bank loan and mortgage applications into risky or safe. In the medical field, classification can be used to help define medical diagnosis and prognosis based on symptoms and health conditions.

Support Vector Machine (SVM) Support vector machine (SVM) is used in both classification and regression. An SVM classifier, a concept by Vladimir Vapnik, finds the optimal separating hyperplane between positive and negative classes of data. The optimal hyperplane is the one that gives maximum margin between the training examples that lie closest to the hyperplane and the data points on the two sides belong to different classes. In linear SVM the given data set is considered as p-dimensional vector that can be separated by maximum of p-1 planes called hyper-planes. These planes separate the data space or set the boundaries among the data groups for classification or regression problems. The best hyper-plane can be selected among the number of hyper-planes on the basis of distance between the two classes it separates. The plane that has the maximum margin between the two classes is called the maximum-margin hyper-plane. It can handle nonlinear classification tasks efficiently by mapping the samples into a higher dimensional feature space by using a nonlinear kernel function. Since the SVM approach is data-driven and model free, it has important discriminating power for classification. SVM algorithms use a set of mathematical functions that are defined as the kernel. The function of the kernel is to take data as input and transform it into the required form. Different SVM algorithms use different types of kernel functions. These functions can be different types. For example linear, nonlinear, polynomial, radial basis function (RBF), and sigmoid. The most used type of kernel function is RBF [442]. Because it has localized and finite response along the entire x-axis. The kernel functions return the inner product between two points in a suitable feature space. Both SVM and RF are widely used for classification tasks, including the classification of diabetic patients based on various features such as medical history, clinical measurements, and demographic information. Both algorithms can provide insights into feature importance. SVM determines the support vectors, which are the data points closest to the decision boundary, while RF calculates feature importance based on how much each feature contributes to the model's predictive performance. In diabetes research, this can help identify the most relevant features for predicting diabetes or assessing disease progression.

Decision Tree (DT) Decision tree (DT) is a supervised learning that can be used as a regression tree while the response variable is continuous and as a classifcation tree while the response variable is categorical. Whereas the input variables are any types, as like graph, text, discrete, continuous, and so on in the case of both regression and classification. The finding of a solution with the help of decision trees starts by preparing a set of solved cases. The whole set is then divided into 1) a training set, which is used for the induction of a decision tree, and 2) a testing set, which is used to check the accuracy of an obtained solution [443]. A decision tree is a tree structure based model which describes the classification process based on input features.

The steps of DT as follows:

- Construct a tree with its nodes as input features.
- Select the feature to predict the output from the input features whose gives the highest information gain.
- Repeat the above steps to form sub trees based on features which was not used in the above nodes.

The decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. Decision trees classify instances by sorting them down the tree from the root to any leaf node, which provides the classification of the instance.

Naive Bayes Classifier (NB) Naive Bayes classifier is a well-known type of classifiers, i.e., of programs that assign a class from a predefined set to an object or case under consideration based on the values of descriptive attributes. A well-known Bayesian network classifier is the Naïve Bayes' classifier is a probabilistic classifier based on the Bayes' theorem, considering Naïve (Strong) independence assumption [444]. the stage of the calculation Naive Bayes as follows:

- Find the value of prior probability for each class calculate the average of each class.
- Find the value of the likelihood that is a process of calculating the probability of each attribute against the class, the possibility of the emergence of a class when an attribute is selected.
- Find the value of the posterior that is result of calculation likelihood in the form of the probability of the attribute class, calculated to divert the possibility of the attribute of the input with the class, in the process of this can be the probability of the end.

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'. The naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

K-Nearest Neighbors (KNN) KNN algorithms are supervised non-parametric learning algorithms that learn the relationship between input and output observations. It is simply based on the idea that "objects that are 'near' each other will also have similar characteristics. Thus if you know the characteristic features of one of the objects, you can also predict it for its nearest neighbour." k-NN is an improvisation over the nearest neighbour technique. It is based on the idea that any new instance can be classified by the majority vote of its 'k' neighbours, - where k is a positive integer, usually a small number [445].

The criterion is defined by Euclidean distance; and if the two locations contain $O_1=\{x_{11},x_{12},x_{13},\ldots\ldots\ldots .,x_{1n}\}$ and $O_2=\{x_{21},x_{22},x_{23},\ldots\ldots\ldots .,x_{2n}\}$ then the Euclidean distance between them is defined according to Eq. (1).

$$d(O_1,O_2) = \sqrt{\sum_{j=1}^{k}\left(x_{1,j}-x_{2,j}\right)^2} \tag{1}$$

The algorithm is based on the distance between two instances, which represents their similarity. KNN identifies k instances in the training set and then classifies a new instance based on how similar (i.e., near) it is to its neighbors. Generally speaking, a new instance is classified by a majority vote of its neighbors. Thus, when the algorithm is used for classification purposes, the output is the class membership of the new instance.

The hyperparameter k is a user-defined positive odd integer, typically small. If k = 1, the algorithm considers the neighbor nearest to the unclassified instance. If k = 3, then KNN compares the distances of the three neighbors nearest to the unclassified instance.

Random Forest (RF) A random forest classifier is the assembly of tree-structured classifiers. This algorithm supplements the objects from array of input to every tree of the forest. The elements of the unit vector are individually voted for classification by every single tree. The forest filters the most voted classifications out of the forest. The simplest random forest with random features is formed by selecting at random, at each node, a small group of input variables to split on. Grow the tree using CART methodology to maximum size and do not prune. Random forest is a machine learning methodology for classification, which is commonly used in computational biology fields. Independently trained decision trees are merged in a random forest, which is done by subsets randomly sampled with replacement from the training data. Every branch of decision tree discovers a best feature in the training time. The best feature randomly chosen from a subset of feature space. Because trees are trained in subset of feature space and training data, they should not be produced with post-pruning [446] .The prediction of RF is the average or a majority vote of all tree predictions that have been trained. Random forest algorithm has three parameters that should be set in training time. These parameters are the number of growing trees, the minimum node size to split and the number of features to select randomly for each split. RF reduces the degree of over-fitting by combining multiple overfit evaluators (ie, decision trees) to form an ensemble learning algorithm. Each decision tree can get the corresponding classifcation decision result. By using the voting results of each decision tree in the forest, the category of the sample to be tested is determined according to the principle of minority obeying the majority, and the category with higher votes in all decision trees was determined to be the final result.

One of the biggest advantages of random forest is its versatility. It can be used for both regression and classification tasks, and it's also easy to view the relative importance it assigns to the input features. Random forest is also a very handy algorithm because the default hyperparameter it uses often produce a good prediction result. Understanding the hyperparameter is pretty straightforward, and there's also not that many of them. One of the biggest problems in machine learning is overfitting, but most of the time this won't happen thanks to the random forest classifier. If there are enough trees in the forest, the classifier won't overfit the model. The main limitation of random forest is that a large number of trees can make the algorithm too slow and ineffective for real-time predictions [447]. In general, these algorithms are fast to train, but quite slow to create predictions once they are trained. A more accurate prediction requires more trees, which results in a slower model. In most real-world applications, the random forest algorithm is fast enough but there can certainly be situations where run-time performance is important and other approaches would be preferred. And, of course, random forest is a predictive modeling tool and not a descriptive tool, meaning if you're looking for a description of the relationships in your data, other approaches would be better.

2.1.2 Regression

Regression analysis consists of a set of machine learning methods that allow us to predict a continuous outcome variable (y) based on the value of one or multiple predictor variables (x). Briefly, the goal of regression models is to build a mathematical equation that defines y as a function of the x variables. Next, this equation can be used to predict the outcome (y) on the basis of new values of the predictor variables (x). Linear regression is the most simple and popular technique for predicting a continuous variable. It assumes a linear relationship between the outcome and the predictor variables.

The linear regression equation can be written as y = b0 + b*x + e, where b0 is the intercept, b is the regression weight or coefficient associated with the predictor variable x and e is the residual error. When it has multiple predictor variables, say x1 and x2, the regression equation can be written as y = b0 + b1*x1 + b2*x2 +e. In some situations, there might be an interaction effect between some predictors that is for example, increasing the value of a predictor variable x1 may increase the effectiveness of the predictor x2 in explaining the variation in the outcome variable [448].

In some cases, the relationship between the outcome and the predictor variables is not linear. In these situations, it needs to build a non-linear regression, such as polynomial and spline regression. Regression is a supervised learning technique which helps in finding the correlation between variables and enables us to predict the continuous output variable based on the one or more predictor variables. It is mainly used for prediction, forecasting, time series modeling, and determining the causal-effect relationship between variables.

In Regression, plot a graph between the variables which best fits the given data points, using this plot, the machine learning model can make predictions about the data. In simple words, "Regression shows a line or curve that passes through all the data points on target-predictor graph in such a way that the vertical distance between the data points and the regression line is minimized." The distance between data points and the line tells whether a model has captured a strong relationship or not. The most common regression techniques are: Linear regression (LR), Logistic regression, Polynomial regression, support vector regression, Decision tree regression and Random forest regression.

2.2 Unsupervised learning

Unsupervised Learning is a machine learning technique in which the users do not need to supervise the model. Instead, it allows the model to work on its own to discover patterns and information that was previously undetected. It mainly deals with the unlabeled data. It allows users to perform more complex processing tasks compared to supervised learning, whereas unsupervised learning can be more unpredictable compared with other natural learning methods. Unsupervised learning algorithms include clustering, association rule, neural networks, etc.

2.2.1 Clustering

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them. Clustering is very much important as it determines the intrinsic grouping among the unlabeled data present. There are no criteria for a good clustering. It depends on the user, what are the criteria they may use which satisfy their needs [449] . There are different forms of clustering, which is explained below.

- Density-Based Methods: These methods are considered the clusters as the dense region having some similarity and different from the lower dense region of the space. These methods have good accuracy and the ability to merge two clusters.

Example DBSCAN (Density-Based Spatial Clustering of Applications with Noise), OPTICS (Ordering Points to Identify Clustering Structure) etc.

- Hierarchical Based Methods: The clusters formed in this method form a tree-type structure based on the hierarchy. New clusters are formed using the previously formed one. It is divided into two categories Agglomerative (bottom up approach) and Divisive (top down approach).

2.2.2 Association rule

Association rule mining finds interesting associations and relationships among large sets of data items. This rule shows how frequently an item set occurs in a transaction. A typical example is Market Based Analysis. Association rule learning is a type of unsupervised learning technique that checks for the dependency of one data item on another data item and maps accordingly so that it can be more profitable. It tries to find some interesting relations or associations among the variables of the dataset. It is based on different rules to discover the interesting relations between variables in the database [450]. To measure the associations between thousands of data items, there are several metrics. These metrics are given below:

- Support: Support is the frequency of A or how frequently an item appears in the dataset. It is defined as the fraction of the transaction T that contains the item set X. If there are X datasets, then for transactions T, it can be written as$Supp\ (x) = \frac{freq(x)}{T}$.
- Confidence: Confidence indicates how often the rule has been found to be true. Or how often the items X and Y occur together in the dataset when the occurrence of X is already given. It is the ratio of the transaction that contains X and Y to the number of records that contain X.
- Lift: It is the strength of any rule. It is the ratio of the observed support measure and expected support if X and Y are independent of each other. It has three possible values: If Lift= 1: The probability of occurrence of antecedent and consequent is independent of each other. If lift>1: It determines the degree to which the two item sets are dependent on each other. If lift<1: It tells us that one item is a substitute for other items, which means one item has a negative effect on another. Apriori Algorithm is the most common association technique used in machine learning application.

2.3 Semi-supervised learning

Semi-supervised learning is the type of machine learning that uses a combination of a small amount of labeled data and a large amount of unlabelled data to train models. This approach to machine learning is a combination of supervised machine learning, which uses labeled training data, and unsupervised learning, which uses unlabelled training data. Semi-supervised machine learning is a combination of supervised and unsupervised learning. The basic procedure involved is that first, the developer will cluster similar data using an unsupervised learning algorithm and then use the existing labeled data to label the rest of the unlabelled data [451].

2.4 Reinforcement learning

Reinforcement learning differs from the supervised learning in a way that in supervised learning the training data has the answer key with it so the model is trained with the correct answer itself whereas in reinforcement learning, there is no answer but the reinforcement agent decides what to do to perform the given task. In the absence of a training dataset, it is bound to learn from its experience [452].

2.5 Evolutionary learning

Evolutionary algorithms are a heuristic-based approach to solving problems that cannot be easily solved in polynomial time, such as classically NP-Hard problems, and anything else that would take far too long to exhaustively process. When used on their own, they are typically applied to combinatorial problems; however, genetic algorithms are often used in tandem with other methods, acting as a quick way to find a somewhat optimal starting place for another algorithm to work off of [453].

2.6 Deep learning

Deep learning is an artificial intelligence (AI) function that imitates the workings of the human brain in processing data and creating patterns for use in decision making. Deep learning is a subset of machine learning in artificial intelligence that has networks capable of learning unsupervised from data that is unstructured or unlabelled also known as deep neural learning or deep neural network [454]. It has the following characteristics.

- Deep learning is an AI function that mimics the workings of the human brain in processing data for use in detecting objects, recognizing speech, translating languages, and making decisions.
- Deep learning AI is able to learn without human supervision, drawing from data that is both unstructured and unlabelled.
- Deep learning, a form of machine learning, can be used to help detect fraud or money laundering, among other functions.

The different types of neural networks in deep learning are convolutional neural networks (CNN), recurrent neural networks (RNN), artificial neural networks (ANN), etc. The neural network has been widely used to train predictive models for applications such as image processing, disease prediction, and face recognition [439].

2.6.1 ANN

Artificial Neural Network, or ANN, is a group of multiple perceptrons/ neurons at each layer. ANN is also known as a Feed-Forward Neural network because inputs are processed only in the forward direction. ANN consists of 3 layers – Input, Hidden and Output. The input layer accepts the inputs, the hidden layer processes the inputs, and the output layers produce the result. Essentially, each layer tries to learn certain weights. In a neural network, one neuron to the other neuron connection exists with some strength known as weight or synaptic weight. The neural network consists of feedback, and information can flow from the input layer to the output layer via one or more hidden layers, and vice versa known as a feedback neural network [455].

2.6.2 CNN

A Convolutional Neural Network (CNN) is a Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other. The pre-processing required in a CNN is much lower as compared to other classification algorithms. While in primitive methods, filters are hand-engineered, with enough training, CNN has the ability to learn these filters/characteristics. Convolutional neural network (CNN) is one of the most popular and used of DL networks . Because of CNN, DL is very popular nowadays. The main advantage of CNN compared to its predecessors is that it automatically detects the significant features without any human supervision which made it the most used [456].

2.6.3 RNN

In a feed-forward neural network, the information only moves in one direction — from the input layer, through the hidden layers, to the output layer. The information moves straight through the network and never touches a node twice.

Feed-forward neural networks have no memory of the input they receive and are bad at predicting what's coming next. Because a feed-forward network only considers the current input, it has no notion of order in time. It simply can't remember anything about what happened in the past except its training. In a RNN the information cycles through a loop. When it makes a decision, it considers the current input and also what it has learned from the inputs it received previously [457].

A usual RNN has a short-term memory. In combination with an LSTM they also have a long-term memory. A recurrent neural network, however, is able to remember those characters because of its internal memory. It produces output, copies that output and loops it back into the network. Therefore, a RNN has two inputs: the present and the recent past. This is important because the sequence of data contains crucial information about what is coming next, which is why a RNN can do things other algorithms can't. A feed-forward neural network assigns, like all other deep learning algorithms, a weight matrix to its inputs and then produces the output. RNNs apply weights to the current and also to the previous input. Furthermore, a recurrent neural network will also tweak the weights for both through gradient descent and back propagation through time (BPTT).

3 Review based on machine learning technique in diabetes prediction

3.1 Supervised learning (Classification)

3.1.1 Support vector machine (SVM)

Diabetic retinopathy has become a common eye disease in most developed countries. It occurs in 80% of all diabetic cases and is the leading cause of blindness [13]. Regular screening is the most efficient way of reducing the preventable eye damages. There are two kinds of symptoms in the diabetic retinopathy. One is dark lesion that includes hemorrhages, and microaneurysms. The other is bright lesion such as exudates and cotton wool spots. Microaneurysms are commonly detected in the retinal fluorescein angiography [14, 15].

In 2005, Zhang and Chutatape [16] introduced an SVM approach for detection of hemorrhages in background diabetic retinopathy. This paper focuses on the detection of hemorrhages which have "dot" and "blot" configurations in the background diabetic retinopathy with their color similar to the blood vessels. In this paper, a top-down strategy is applied to detect hemorrhages. The SVM classifier uses features extracted by combined 2DPCA (Two-Dimensional Principal Component Analysis) instead of explicit image features as the input vector. After locating the hemorrhages in the ROI (Region of Interest), the boundaries of the hemorrhages can be accurately segmented by the post-processing stage. The paper demonstrates a new implementation of various techniques on the problem and shows the improvement it offers over the others. Combined 2DPCA is proposed and virtual SVM is applied to achieve higher accuracy of classification. The test result demonstrates that the TP (True Positive) rate of SVM is 89.1%, while that of ANN is 84.6% at FP rate of two FP per image. The Gaussian kernel is used in SVM. The SVM based on SRM (Structural Risk Minimization) appears to be superior to ANN that employs ERM (Empirical Risk Minimization). It also compared the performance of SVM with VSVM (Virtual SVM) and found that classification accuracy of VSVM that uses the rotation invariance and illuminance invariance is better than SVM. When number of FP remains 2 per image, the TP rate of VSVM is 94% while the TP of SVM is 93.2%.

In 2006, Stoean et al. [17] proposed an ESVM (Evolutionary support vector machine) technique for diagnosis of diabetes mellitus. The main aim of this paper is to validate the new paradigm of evolutionary support vector machines (ESVMs) for binary classification also through an application to a real world problem, i.e. the diagnosis of diabetes mellitus. Different algorithms like (CPLEX (COmmercial Solvers for Integer Programming and Mathematical Programming by Linear Programming Extensions), SVM light, Active SVM, and Critical SVM) have been utilized for experimental evaluation and compare the performance with ESVM. The test result depicts that proposed technique offers a good enough accuracy in comparison to the state-of-the-art classical approaches and to the standard SVM formulation. Possibly, application of parameter tuning methods like SPO on ESVMs with a polynomial kernel would lead to better values for the evolutionary parameters that would improve the proportion of self-determined training errors. The proposed method achieves training accuracy of 77.95%, whereas test accuracy is 80.22%.

In 2008, Balakrishnan et al. [18] introduced a feature selection approach for finding an optimum feature subset that enhances the classification accuracy of the Naive Bayes classifier. Experiments were conducted on the Pima Indian Diabetes Dataset to assess the effectiveness of our approach. The results confirm that SVM Ranking with Backward Search approach leads to promising improvement in feature selection and enhances classification accuracy. Polat et al. [19] proposed a new cascade learning system based on Generalized Discriminant Analysis and Least Square Support Vector Machine. The proposed system consists of two stages. The first stage, used Generalized Discriminant Analysis to discriminant feature variables between healthy and patient (diabetes) data as a pre-processing process. The second stage used LS-SVM in order to classification of diabetes dataset. While LS-SVM obtained 78.21% classification accuracy, using 10-fold cross validation, the proposed system called GDA–LS-SVM obtained 82.05% classification accuracy using 10-fold cross validation.

In 2009, WU et al. [20] developed a semi-supervised based learning method (LapSVM) for diabetes disease diagnosis. Firstly, LapSVM was trained as a fully-supervised learning classifier to predict diabetes dataset and 79.17% accuracy was obtained. Then, it was trained as a semi-supervised learning classifier and got the prediction accuracy 82.29%. The obtained accuracy 82.29% is higher than other previous reports. The experiments led to the finding that LapSVM offers a very promising application, i.e., LapSVM can be used to solve a fully-supervised learning problem by solving a semi-supervised learning problem. The result suggests that LapSVM can be of great help to physicians in the process of diagnosing diabetes disease and it could be a very promising method in the situations where a lot of data are not class-labelled.

In 2010, Yu et al. [21] develop and validate SVM models for two classification schemes: Classification Scheme I (diagnosed or undiagnosed diabetes vs. pre-diabetes or no diabetes) and Classification Scheme II (undiagnosed diabetes or prediabetes vs. no diabetes). The SVM models were used to select sets of variables that would yield the best classification of individuals into these diabetes categories. The overall discriminative ability of classification Schemes I and II are represented by their AUC values (83.47% and 73.18%, respectively). Barakat et al. [22] proposed a support vector machine (SVM) for the diagnosis of diabetes. In particular, use an additional explanation module, which turns the "black box" model of an SVM into an intelligible representation of the SVM's diagnostic (classification) decision. Result in a real-life diabetes dataset shows that intelligible SVMs provide a promising tool for the prediction of diabetes, where a comprehensible ruleset have been generated, with prediction accuracy of 94%, sensitivity of 93%, and specificity of 94%.

In 2011, Calisir et al. [23] proposed an automatic diagnosis system for diabetes on Linear Discriminant Analysis (LDA) and Morlet Wavelet Support Vector Machine Classifier: LDA–MWSVM is introduced. The Linear Discriminant Analysis (LDA) is used to separate features variables between healthy and patient (diabetes) data in the first stage. The healthy and patient (diabetes) features obtained in the first stage are given to inputs of the MWSVM classifier in the second stage. Finally, in the third stage, the correct diagnosis performance of this automatic system based on LDA–MWSVM for the diagnosis of diabetes is calculated by using sensitivity and specificity analysis, classification accuracy, and confusion matrix, respectively. The classification accuracy of this system was obtained at about 89.74%.Gupta et al. in [24] present a study aimed to do the performance analysis of several data mining classification techniques using three different machine learning tools over the healthcare datasets. In this study, different data mining classification techniques have been tested on four different healthcare datasets. The standards used are percentage of accuracy and error rate of every applied classification technique. The experiments are done using the 10 fold cross validation method. A suitable technique for a particular dataset is chosen based on highest classification accuracy and least error rate. The test result based on PIMA Indian Diabetes dataset show that an accuracy rate of 96.74% and 3.18% of the error rate is achieved using SVM technique which is superior when contrasted with another technique.

In 2020, Xue et al. [88] proposed an automatic diagnosis system for diabetes using supervised machine-learning algorithms like Support Vector Machine (SVM), Naive Bayes classifier and LightGBM to train on the actual data of 520 diabetic patients and potential diabetic patients aged 16 to 90. Although the naive Bayes classifier is the most popular classification algorithm, the final accuracy rate on the given data set is only 93.27%. SVM has the highest accuracy rate, with an accuracy rate of 96.54%. The accuracy of LightGBM is only 88.46%. It is found that SVM has the highest accuracy through the confusion matrix evaluation test.

In 2021, Chaves et al. [91] introduce a comparative study of data mining techniques for early diagnosis of diabetes. We use a publicly accessible data set containing 520 instances, each with 17 attributes. Naive Bayes, Neural Network, AdaBoost, k-Nearest Neighbors, Random Forest and Support Vector Machine methods have been tested. The results suggest that Neural Networks should be used for diabetes prediction. The proposed model presents an AUC of 98.3% and 98.1% accuracy, an F1-Score, Precision and Sensitivity of 98.4% and a Specificity of 97.5%. In the first experiment, author applied the Naive Bayes classifier, which correctly predicted 452 instances out of 520, a success rate of 86.92%. In the second experiment, author applied the Neural Network classifier, which correctly predicted 510 instances out of 520, a success rate of 98.08%. In the third experiment, author applied the AdaBoost classifier, which correctly predicted 506 instances out of 520, a success rate of 97.31%. In the fourth experiment, author applied the kNN classifier, which correctly predicted 506 instances out of 520, a success rate of 97.31%.In the fifth experiment, author applied the Random Forest classifier, which correctly predicted 504 instances out of 520, a success rate of 96.92%.In the last experiment, author applied the SVM classifier, which correctly predicted 505 instances out of 520, a success rate of 97.12%.

In 2022, Li et al. [408] proposed an effective biomarkers for an efficient diagnosis of type 2 diabetes. The sensitivity and specificity of the SVM model for identifying patients with type 2 diabetes were 100%, with an area under the curve of 1 in the training as well as the validation dataset. In 2023, Lei et al. [441] propose a publicly verifiable and secure SVM classification scheme (PVSSVM) for cloud-based health monitoring services. It utilize homomorphic encryption and secret sharing to protect the model and data confidentiality in the cloud server, respectively. Based on a multi-server verifiable computation framework, PVSSVM achieves public verification of predicted results. The proposed scheme achieves a reduction of approximately 83.71% in computation overhead through batch verification, as compared to one-by-one verification. Table 1 represent the related work on the diagnosis of diabetes based on SVM algorithm.

3.1.2 Decision Tree (DT)

In 2002, Breault et al. [102] introduce a classification tree approach in Classification and Regression Trees (CART) with a binary target variable of HgbA1c >9.5 and 10 predictors: age, sex, emergency department visits, office visits, comorbidity index, dyslipidemia, hypertension, cardiovascular disease, retinopathy, end-stage renal disease. The first level of the tree shows that just dividing people using an age cut-point of 65.581 years of age, 19.4% of younger people (n ¼ 3987) have a bad HgbA1c. This is 2.8 times the rate of bad HgbA1c values in those who are older (7.0%, n ¼ 3966). The dataset contains the data from 442 bed tertiary care hospital, a 500 physician multi-specialty clinic in 25 locations.

In 2004, Haung [105] investigated the potential for data mining in order to spot trends in the data and attempt to predict outcome. Feature selection has been utilized to enhance the efficiency of the data mining algorithm. Decision Tree (C4.5) is utilized in this work and results show that before feature selection, discretized C4.5 had the best performance of classification. And after feature selection C4.5 obtained the best result. Diabetic data has been collected from the Ulster community and trust hospital for the year 2000 to 2004. The dataset contained 2017 type-2 diabetic patients' clinical information having 1124 males and 893 females.

In 2008, Liou et al. [109] proposed to detect fraudulent or abusing the reporting by health care providers using their invoice for diabetic outpatient services. The proposed

Table 1 (SVM based) Diabetes diagnosis summary

Author, Year, References	Algorithms	Experimental Results	Datasets
Zhang, 2005 [16]	SVM and VSVM	TP (True Positive) rate of SVM is 89.1% while that of ANN is 84.6%, TP rate of VSVM is 94% while the TP of SVM is 93.2%.	SNEC (Singapore National Eye Centre)
Stoean, 2006 [17]	ESVM (Evolutionary support vector machine)	The proposed method achieves training accuracy of 77.95% whereas test accuracy is 80.22%.	Pima Indian diabetes
Balakrishnan, 2008 [18]	SVM Ranking with Backward Search approach.	Leads to promising improvement on feature selection and enhances classification accuracy.	Pima Indian diabetes
Polat, 2008 [19]	Generalized Discriminant Analysis and Least Square Support Vector Machine.	LS-SVM obtained 78.21% classification accuracy using 10-fold cross validation, the proposed system called GDA–LS-SVM obtained 82.05% classification accuracy using 10-fold cross validation.	UCI Repository of Machine Learning Databases
WU, 2009 [20]	LapSVM	79.17% accuracy was obtained. Then, it was trained as a semi-supervised learning classifier and got the prediction accuracy 82.29%.	Pima Indians diabetes dataset
Yu, 2010 [21]	SVM	The overall discriminative ability of classification Schemes I and II are represented by their AUC values (83.47% and 73.18%, respectively).	National Health and Nutrition Examination Survey (NHANES)
Barakat, 2010 [22]	SVM	Comprehensible ruleset have been generated, with prediction accuracy of 94%, sensitivity of 93%, and specificity of 94%.	Data from 4682 subjects of age 20 years and above was collected.
Calisir, 2011 [23]	(LDA) and Morlet Wavelet Support Vector Machine Classifier	The classification accuracy of this system was obtained at about 89.74%	Pima Indian diabetes
Gupta, 2011 [24]	SVM	Accuracy rate of 96.74% and 3.18% of error rate is achieved	Pima Indian diabetes

Table 1 (continued)

Author, Year, References	Algorithms	Experimental Results	Datasets
Marling, 2011 [25]	LibSVM	The CBR research paradigm has provided a strong framework for medical research as well as for artificial intelligence (AI) research.	CBR Research
Zolfaghari, 2012 [26]	SVM and Back Propagation Neural Network Classifiers	Results show 88.04% accuracy	Pima Indian diabetes
Giveki, 2012 [27]	Feature Weighted Support Vector Machines (FW-SVMs) and Modified Cuckoo Search (MCS)	The proposed MI-MCS-FWSVM method obtains 93.58% accuracy	UCI Repository of Machine Learning Databases
Hashim, 2012 [28]	SVM	Classification accuracy obtained using support vector machine (SVM) with 82.74% and 83.87% for clinical and textural approach respectively.	Messidor benchmarked database
Karatsiolis, 2012 [29]	Region based Support Vector Machine	The suggested algorithm raised average classification success rate to 82.2%	Complete data set consists of 500 normal cases and 268 abnormal cases.
Kumari, 2013 [30]	SVM	The suggested algorithm achieved accuracy rate of 78%, Sensitivity=80% and specificity=76.5%.	Pima Indian diabetes
Farran, 2013 [31]	SVM and KNN	The best random classifier for the data set achieved a classification accuracy of 51.1%. Fivefold CV results for k-NN (at k=6) and SVM (RBF kernel, σ=0.1, C=10) achieve accuracies of 75.6±2.7% and 87.4%±1.1%, respectively.	10 632 (2853 diabetic and 7779 non-diabetic) participants.
Mansour, 2013 [32]	Discrete Cosine Transform (DCT) analysis and SVM	Achieve a diagnostic accuracy with 97.0% sensitivity and 98.7% specificity.	1200 retinal images with variable color, brightness, and quality.

Table 1 (continued)

Tapak, 2013 [33]	Four machine-learning classifiers (neural networks, support vector machines, fuzzy c-mean, and random forests)	Support vector machines showed the highest total accuracy (0.986) as well as area under the ROC (0.979). Also, this method showed high Specificity (1.000) and sensitivity (0.820).	6,500 subjects from the Iranian national non-communicable diseases
Anthimopoulos, 2014 [34]	Linear support vector machine	The system achieved classification accuracy of the order of 78%.	5000 food images was created and organized into 11 classes.
C hoi, 2014 [35]	ANN and SVM	The SVM model showed the areas under the curve of 0.731 in the external datasets, which is higher than those of the ANN model (0.729) and the screening score model (0.712), respectively.	Korean National Health and Nutrition Examination Survey (KNHANES)
Roychowdhury, 2014 [36]	GMM,K-NN,SVM,Adaboost	Achieves 100% sensitivity, 53.16% specificity and 0.904 AUC, compared to the best reported 96% sensitivity,51% specificity and 0.875 AUC, for classifying images as with or without DR.	1200 images from the publicly available MESSIDOR data set.
Baitharu, 2015 [101]	SVM trained using linear, polynomial, puk and Radial Basic Function (RBF) kernels	A preliminary study has been made between SVM using the best choice of kernel. Results had revealed that SVM trained using Linear Kernel is the best choice for dealing with Diabetes dataset.	Diabetes dataset available UCI Machine Learning Repository
Cai, 2015 [37]	logistic regression (LR), linear discriminant analysis (LDA), Naïve Bayes (NB) and support vector machine (SVM)	As regards the choice of classifier used to estimating selected features, it is found that the SVM classifier give the best and stable accuracy on both datasets.	17,473 genes were processed for the following analysis.
Jaya, 2015 [38]	Fuzzy support vector machine (FSVM)	The area under the receiver operating characteristic curve reached 0.9606, which corresponds to a sensitivity of 94.1 % with a specificity of 90.0 %.	200 retinal images collected from diabetic retinopathy

Table 1 (continued)

Arjun, 2015 [39]	SVM	In the proposed system, the SVM algorithm is shown the superior performance in terms of high precision, recall and accuracy values.	768 tuples and 8 attributes
Kang, 2015 [40]	Efficient and effective ensemble of SVMs	The accuracy of the proposed method was approaching 77.952%, Overall, SVM ensembles outperformed sampling-based single SVMs, and the RBF kernel was better than the polynomial kernel in most cases.	Seoul National University Hospital in the Republic of Korea
Ramanathan, 2015 [41]	(SVM) and fuzzy modelling (SVM-Fuzzy)	The proposed system design showed classification accuracy of 96%. Out of 50 input data, 48 showed correct classification.	Pima Indian diabetes dataset
Santhanam, 2015 [42]	SVM+GA+K-mean	The proposed model has attained an average accuracy of 98.79 %	Pima Indian diabetes dataset
Sowjanya, 2015 [43]	J48, Naïve Bayes, SVM with Polykernel and RBF kernel and Multilayer Perceptron	J48 algorithm proven to give better results compared to other algorithms with sensitivity (0.890), specificity (0.928) and ROC areas of 0.928.	145 instances with 105 of males and 40 of female candidates.
Tafa, 2015 [44]	SVM+Naive-Bayes	Both SVM and naïve Bayes algorithm have individually shown high overall classifier performances of 95, 52% and 94, 52%, respectively.	402 instances taken from three different locations in Kosovo.
Abdillah, 2016 [45]	SVM-RBF	The highest performance of SVM-RBF using 10-fold cross validation was obtained from 500 training data with optimal parameter , which yields accuracy, sensitivity, specificity, and AUROC of 80.22%, 82.56%, 9.12%, and 0.8084 respectively.	Pima Indian diabetes dataset

Table 1 (continued)

Bano, 2016 [46]	SVM+K-NN	By experimental study, it achieved 95.80% accuracy using k-NN classifier and 80.92% accuracy of SVM.	Pima Indian diabetes dataset
GILL, 2016 [47]	SVM+Neural network	The proposed hybrid model obtained the accuracy of 96.09%.	Pima Indian diabetes dataset
Huang, 2016 [48]	SVM+Decision Tree	The proposed technique achieved 99.88% accuracy on training data and 64.93% on testing data.	Diabetic Retinopathy Debrecen Dataset
Kose, 2016 [49]	(SVM) and Cognitive Development Optimization Algorithm (CoDOA)	The proposed technique achieved 87% accuracy on training data and 50% on testing data.	Pima Indian diabetes dataset
Malik, 2016 [50]	SVM+RBF+LR+ANN	SVM using RBF kernel showed the best performance for classifying high FBGLs with approximately 85 % accuracy, 84 % precision, 85 % sensitivity and 85 % F_1 score.	Pima Indian diabetes dataset
Negi, 2016 [51]	SVM	The proposed technique achieved 73% accuracy on training data and 72% on testing data.	Pima Indian diabetes dataset+ Diabetes 130-US dataset
Osman, 2017 [52]	SVM and K-mean Clustering	The diagnosis accuracy results after improvement by using the hybrid method between K-means and SVM algorithm are 99.74, 99.78, and 99.81 for training data and 99.82, 99.85, and 99.90 for testing data respectively.	Pima Indian diabetes dataset
Carrera, 2017 [53]	SVM	Proposed technique achieved maximum sensitivity of 94.6% and a predictive capacity value of 93.8%.	400 retinal images labeled according to a 4-grade scale diabetic retinopathy.
Khalil, 2017 [54]	SVM, Fuzzy-C-Mean,K-mean, PNN	Proposed technique achieved accuracy (SVM= 96.87%, Fuzzy-C-Mean=95.4%, K-mean=87.8% and PNN=93.7%)	Black Lion General Specialized Hospital, Addis Ababa, Ethiopia.

Table 1 (continued)

Rathore, 2017 [55]	SVM, Decision Tree	Diabetes Detection is done successfully with SVM Classifier with 82% accuracy.	Pima Indian diabetes dataset
Wang, 2017 [56]	SVM	The AUC values achieve 0.96, 0.67, 0.66, 0.50 in the four categories respectively. The first group of differently coexpressed gene pairs obtains the best classification performance AUC of 0.96.	NCBI GEO database
Zhang,2017 [57]	SVM	The accuracy rate of diabetes predication was increased from 77.83% to 78.77%. Th accuracy rate and area under curve (AUC) were not reduced after reducing the dimensions of tongue features with principal component analysis (PCA), while substantially saving the training time. During the training for selecting SVM parameters by genetic algorithm (GA), the accuracy rate of cross-validation was grown from 72% or so to 83.06%.	Tongue images of 296 diabetic subjects and 531 subjects were collected by the TDA-1 digital tongue instrument.
Cui, 2018 [58]	Improved SVM with GA	Experimental results indicate that the proposed SVM-based method achieves an accuracy of 81.02%, a sensitivity of 82.89%, a specificity of 79.23%.	Health Facts database (Cerner Corporation, Kansas City, MO).
Dagliati, 2018 [59]	SVM,NB,RF,LR	AUC values are higher for SVMs and RF when the data sets are balanced. However, SVMs and RF models are harder to interpret, especially considering that our final goal is the model application into clinical practice.	Data of nearly 1,000 T2DM patients.
Joshi, 2018 [60]	SVM,LR	Experiments revealed that SVM showed better performance in accuracy as the best result is around 0.79.	Not reported

Table 1 (continued)

Rao, 2018 [61]	SVM	The results demonstrate that the proposed CEHO exhibit better to competitive results across all datasets.	PIMA, WBC
Mule, 2018 [62]	SVM	Results are compared using parameters of sensitivity, specificity, and accuracy, and found high accuracy of 96.32% which can help to detect and prevent diabetic retinopathy.	STARE dataset
Abdullah, 2018 [63]	SVM	The experimental results shows that data classification using SVM upon varied kernel function has improvement over the accuracy with 86.65%, precision 76.21% and recall with 81.11% respectively. Among the kernels experimented, polynomial kernel performed better than the other kernels with increased correlation results.	Dataset is collected from UCI repository. It consists of 12 attributes including 400 patient records.
Sisodia, 2018 [64]	SVM,DT,Naive Bayes	Results obtained show Naive Bayes outperforms with the highest accuracy of 76.30% comparatively other algorithms.	Pima Indian diabetes dataset
Brisimi, 2018 [100]	Linear and kernelized Supp ort Vec tor Machines (SVM), random forests, and logistic regression.	It explored a diverse set of methods, namely kernelized, linear and 1-regularized linear Support Vector Machines, 1-regularized logistic regression and random forests. We proposed a likelihood ratio test-based method, K-LRT, that is able to identify the K most significant features for each patient that lead to hospitalization.	Boston Medical Center (BMC)

Table 1 (continued)

Tsao, 2018 [65]	SVM,ANN,DT,LR	Experimental results demonstrated that prediction performance by support vector machines performed better than the other machine learning algorithms and achieved 79.5% and 0.839 in accuracy and area under the receiver operating characteristic curve.	“DM shared care” database in a private hospital in northern Taiwan.
Alirezaei, 2019 [66]	SVM	It is concluded that the multi-objective firefly (MOFA) and multi-objective imperialist competitive algorithm (MOICA) with a 100% classification accuracy outperform the non-dominated sorting genetic algorithm (NSGA-II) and multi-objective particle swarm optimization (MOPSO) with the accuracies of 98.2% and 94.6%, respectively.	PIMA Indian Type-2 diabetes dataset
Bernardini, 2019 [67]	Sparse Balanced Support Vector Machine	Results evidence that the SB-SVM overcomes the other state-of-the-art competitors providing the best compromise between predictive performance and computation time.	FIMMG dataset
Raj, 2019 [68]	SVM, Naive Bayes	The prediction efficiency of Naïve Bayes algorithm for sample datasets is 62.5% and prediction efficiency of Support vector machine for sample datasets is 82%.	800 data sets is taken for the study.
He, 2019 [69]	Structured Output Support Vector Machine (SOSVM)	AUCs for MMED, SOSVM, SVMkernel, SVM-linear and Naive Bayes at that time point are: 0.78 +0.17, 0.77+ 0.16, 0.74+ 0.17, 0.58+ 0.23, and 0.69 +0.19.	100 subjects in total, and each subject has different number of time points ranging from 12 to 14.
Karkuzhali, 2019 [70]	SVM-RBF	SE= 95, SPE= 98.75, ACC=98 and Computation time : 28	Began Singh Eye Hospital

Table 1 (continued)

Abbas, 2019 [71]	SVM	Shows an average accuracy of 96.80% and a sensitivity of 80.09% obtained on a holdout set.	Oral glucose tolerance test (OGTT)
Lokuarachchi, 2019 [72]	SVM,KNN,DT and Ensemble	The sensitivity and specificity of (kNN 92.31% 91.89% , SVM 88.46% 86.48% , Decision Tree 82.69% 75.67% ,Ensemble 78.86% 83.78%).	Diaretdb1 dataset which contains 89 retinal images with an array size of 1500 ×1152 pixels.
Aminah, 2019 [73]	Naive Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), Ensemble Learning (EL) and k Nearest Neighbor (kNN).	The results show that the best accuracy is 85.6%, with specificity is 0.90, and the sensitivity is 0.80.	26 subjects, of which 15 were non diabetic subjects (5 male and ten female
Qomariah, 2019 [74]	SVM,CNN	From the results of the experiments, the highest accuracy values are 95.83% and 95.24% for base 12 and base 13 respectively.	Tested using 77 and 70 retinal images from Messidor database of base 12 and base 13 respectively.
Selvathi, 2019 [75]	SVM	Experimental results obtained for various feature combinations gives maximum accuracy of 86. 22%, sensitivity of 94. 07% and specificity of 79. 17% using SVM classifier with five-fold validation.	283 thermal images of an eye is obtained from IGCAR Kalpakkam, TamilNadu.
Sneha, 2019 [76]	SVM, Random Forest, Naive Bayes, DT and KNN.	Accuracy (SVM =77.73, Random forest =75.39, NB =73.48, Decision tree =73.18 and KNN= 63.04).	The dataset is collected from UCI machine repository archive.ics.uci.edu-Diabetes.
Hao, 2019 [77]	SVM, LDA, Random Forest.	For detecting diabetes, the method with the highest out-of-sample prediction accuracy is SVM with polynomial kernel. The algorithm can detect diabetes with 96.35% accuracy. However, all the algorithms have a low accuracy when predicting diabetic patients with hypertension and hyperlipidemia (below 70%).	Collected data from 50 healthy people, 139 diabetic patients without hypertension and hyperlipidemia,

Table 1 (continued)

Azad, 2020 [78]	SVM eHealth Cloud system	It is worth to mention that the system giving remarkable accuracy 77.50% by coarse Gaussian SVM in 10-fold validation and fine Gaussian SVM gives 98.8% accuracy in No validation set.	PIMA Indian Diabetes Dataset
Harimoorthy, 2020 [79]	Improved SVM-Radial bias kernel method.	From the experiment results, improved SVM-Radial bias kernel technique produces accuracy as 98.3%, 98.7% and 89.9% in Chronic Kidney Disease, Diabetes and Heart Disease dataset respectively.	Chronic Kidney Disease (CKD),Diabetes and Heart diseases from UCI dataset.
Jayabalan, 2020 [80]	SVM	SVM achieves 98% precision, 98% recall on this dataset, which is greater than other learning techniques.	Not Reported
Kazerouni, 2020 [81]	SVM,KNN,LR,ANN	SVM algorithm, the mean AUC obtained 95% after stratified 10-fold cross-validation, and the SD obtained 0.05. The mean sensitivity and specificity were 95 and 86%, respectively.	200 unrelated Iranian subjects, 100 T2DM patients , and 100 healthy individual..
Nnamoko, 2020 [82]	(SVM-RBF),C4.5 decision tree, Naive Bayes and Repeated Incremental Pruning to Produce Error Reduction (RIPPER)	Naïve Bayes trained with SMOTEd produced slightly better Precision than IQRd +SMOTEd by 0.001%. Similarly, the SVM-RBF model is 0.60% more precise when trained with SMOTEd dataset than with IQRd SMOTEd.	PIMA Indian Diabetes Dataset
Shuja, 2020 [83]	Bagging, SVM (Support Vector Machine), MLP (Multi-Layer Perceptron), Simple Logistic and Decision Tree.	Accuracy without SMOTE (Bagging=94.14, SVM= 90.73, MLP= 93.46, SLR=92.23 and DT= 92.50), Accuracy with SMOTE (Bagging =94.21, SVM= 89.01, MLP =93.83, SLR =90.26 and DT= 94.70).	The dataset was collected from one of the leading diagnostic labs in the Kashmir valley.

Table 1 (continued)

Mishra, 2020 [84]	Naive Bayes, ANN, SVM, KNN, Random Forest, LSTM, CNN, BLSTM.	This paper proposed an ensemble approach of CNN and LSTM to predict diabetes and provided an Accuracy of 97.14%, Precision of 97.30%, recall of 96.30%, F1-Score of 96.79%, and AUC of 0.97.	PIMA Indian diabetes dataset.
Viloria, 2020 [85]	SVM	Achieved 99.2% with Colombian patients and an accuracy of 65.6% with a data set of patients of a different ethnic background.	500 patients from a public hospital in Colombia.
Wang, 2020 [86]	WRank-SVM	Compared with the other six popular multi-label methods, our WRank-SVM can effectively predict the schemes for hypoglycemic drugs of type 2 diabetes. Meanwhile, receiver operating characteristic (ROC) curve is employed to statistically show the effectiveness of the model.	2443 diabetics provided by the Chinese People's Liberation Army General Hospital.
Srivastava, 2020 [87]	KhmAW-SVM	Proposed KhmAW-SVM method achieves 94.28%, 99%, 89.93% and 92.38% accuracy rates for heart disease, Parkinson's disease, liver disease and diabetes disease respectively.	Pima Indian diabetes datasets.
Xue, 2020 [88]	(SVM), Naive Bayes classifier and LightGBM.	SVM has the highest accuracy rate, with an accuracy rate of 96.54%. The accuracy of LightGBM is only 88.46%.Accuracy of Naive Bayes is 93.27%.	520 diabetic patients and potential diabetic patients aged 16 to 90.
Suresh, 2020 [99]	SVM	PIMA dataset will increase the accuracy of almost all algorithms but the SVM and linear regression leads over others.	PIMA Indian diabetes dataset.
Ahmad, 2021 [89]	LR,SVM,DT,RF,Ensemble	SVM performed best on the HbA1c-labeled dataset while RF performed best on the FPG-labeled dataset, The performance of DT and EVM classifiers improved.	3000 patients data collected over two years from 2016 until 2018 through different departments.

Table 1 (continued)

Alabdulwahhab, 2021 [90]	LDA,SVM,KNN,RF	Random forest outperforms the other methods by accurately classifying 86% of the DR patients on the test data.	327 diabetic patients in Almajmaah, Saudi Arabia.
Chaves, 2021 [91]	Naive Bayes, Neural Network, AdaBoost, k-Nearest Neighbors, Random Forest and Support Vector Machine.	The SVM classifier, which correctly predicted 505 instances out of 520, a success rate of 97.12%.	Data set containing 520 instances, each with 17 attributes.
Dinesh, 2021 [92]	KPCA+GA+SVM	Proposed KPCA-GA-SVM obtains accuracy of 99.53% and also reduced feature size compared to GA-SVM of 98.79% accuracy.	Pima Indians Diabetes Dataset
Khanam, 2021 [93]	DT, KNN, RF, NB, AB, LR, SVM	LR and SVM provided approximately 77%–78% accuracy for both train/test split and K-fold cross-validation method. NN with two hidden layers is considered the most efficient and promising for analyzing diabetes with an accuracy rate of approximately 86% for all varying epochs (200, 400, and 800).	PIMA Indian diabetes dataset.
Reddy, 2021 [94]	Decision tree, random forest, adaptive boosting, bagging and SVM with Gaussian kernel	The proposed algorithm namely SVM with Gaussian kernel achieved better values of evaluation metrics compared to other algorithms. The values obtained for proposed algorithm for accuracy, balanced accuracy, Youden's J index, concordance and Somers' D statistics are 81.3%, 0.8118, 0.6236, 0.6596 and 0.3192 respectively.	The entire dataset contains 24 attributes with 1151 instances.

Table 1 (continued)

Rodríguez-Rodrí guez, 2021 [95]	LR, RF, SVM, GP, MLP, IBK, PCA.	In view of the obtained results, Random Forest (RF), as both a predictive algorithm and an FS strategy, offers the best average performance (Root Median Square Error, RMSE = 18.54 mg/dL) throughout the 12 considered predictive horizons (up to 60 min in steps of 5 min), showing Support Vector Machines (SVM), the best accuracy as forecasting algorithm when considering.	25 subjects underwent continuous monitoring for up to 14 days as normal routines.
Tang, 2021 [96]	PSO-SVM with Radial basis kernel	The test set shows that the AUC area of the PSO-SVM model is 0.989, and the AUC areas of the neural network, naive Bayes and logistic regression are 0.958, 0.926 and 0.955, respectively. It can be seen from the figure that the performance of the PSO-SVM model is significantly higher than that of the comparison algorithm, and it has better predictive performance in diabetes risk prediction.	University of California, Irvine. The data set has a total of 520 samples.
Hossain, 2021 [97]	SVM	The model shows the average accuracy of 84.23% across the 10 folds which is slightly better than the accuracy of the LR model..	1995–2018 administrative dataset
Senthil, 2021 [98]	SVM, Random Forest	The SVM algorithm and Random forest giving the highest specificity of 91.55% and 92.8%, respectively holds best for the analysis of diabetic data.	Data for 769 patients contains both sick and healthy patients.
Sistla, 2022 [407]	SVM	With the available dataset, the accuracy score of training data was 77.5 percent and the accuracy score of test data was 80.5 percent.	PIMA Indian diabetes dataset.

Table 1 (continued)

Li, 2022 [408]	SVM	The sensitivity and specificity of the SVM model for identifying patients with type 2 diabetes were 100%, with an area under the curve of 1 in the training as well as the validation dataset.	GSE164416 dataset
Rastogi, 2023 [409]	SVM	The performance of the proposed mechanism is analysed using the confusion matrix, sensitivity and accuracy performance metrices. In logistic regression, the accuracy is high, i.e., 82.46%, in comparison to other data mining techniques.	PIMA Indian diabetes dataset.
Aslan, 2023 [410]	SVM	The accuracy rates obtained with ResNet18 and Resnet50 are 80.86% and 80.47%, respectively. In the second approach, in the classification made with SVM using 2560 features, the highest accuracy is calculated as 91.67% with the quadratic kernel function.	PIMA Indian diabetes dataset.
Lei, 2023 [441]	SVM	The proposed scheme achieves a reduction of approximately 83.71% in computation overhead through batch verification, as compared to one-by-one verification.	PIMA Indian diabetes dataset.

work is validated in Taiwan's National health insurance system and three kinds of a data mining algorithm like decision tree, logistic regression and neural network have been applied for this proposed work. The experimental result shows that the correct identification rate of decision tree based algorithm (99%) outperforms than the logistic regression model (92%) and neural network model (96%).

In 2010, Patil et al. [112] developed a Hybrid Prediction Model (HPM) model which uses Simple K-means clustering algorithm aimed at validating chosen class label of the given data (incorrectly classified instances are removed, i.e. pattern extracted from original data) and subsequently applying the classification algorithm to the result set. C4.5 algorithm is used to build the final classifier model by using the k-fold cross-validation method. The Pima Indians diabetes data were obtained from the University of California at Irvine (UCI) machine learning repository datasets. The proposed HPM obtained a classification accuracy of 92.38%. In order to evaluate the performance of the proposed method, sensitivity and specificity performance measures that are used commonly in medical classification studies were used.

In 2012, Kelarev et al. [116] introduced detection and monitoring of cardiovascular autonomic neuropathy, CAN, in diabetes patients. Using a small set of features identified previously, this work consists of empirical investigation and comparison of several ensemble methods based on decision trees for a novel application of the processing of sensor data from diabetes patients for pervasive health monitoring of CAN. The experiments relied on an extensive database collected by the Diabetes Complications Screening Research Initiative at Charles Sturt University and concentrated on the particular task of the detection and monitoring of cardiovascular autonomic neuropathy. The best outcomes have been obtained by the novel combined ensemble of AdaBoost (accuracy=94%) and Bagging (accuracy=92.99%) based on J48.

In 2014, Kaur et al. [129] proposed an improved J48 algorithm for the prediction of diabetics. In this proposed work, the modified J48 classifier is used to increase the accuracy rate of the data mining procedure. The data mining tool WEKA has been used as an API of MATLAB for generating the J-48 classifiers. Experimental results showed a significant improvement over the existing J-48 algorithm. Proposed algorithm has large accuracy difference than other algorithms. It has accuracy rate of 99.87% rather than others that show maximum of 77.21%accuracy. The experiment is carried out at Pima Indians diabetes dataset.

In 2018, Zou et al. [155] introduced a decision tree, random forest and neural network to predict diabetes mellitus. The dataset is the hospital physical examination data in Luzhou, China. It contains 14 attributes. In this study, five-fold cross validation was used to examine the models. In order to verity the universal applicability of the methods, chose some methods that have the better performance to conduct independent test experiments. It randomly selected 68994 healthy people and diabetic patients' data, respectively as training set. Due to the data unbalance, it randomly extracted 5 times data. And the result is the average of these five experiments. In this study, it used principal component analysis (PCA) and minimum redundancy maximum relevance (mRMR) to reduce the dimensionality. The results showed that prediction with random forest could reach the highest accuracy (ACC = 0.8084) when all the attributes were used.

In 2020, Pei et al. [167] proposed a J48 decision tree based diabetic prediction system for chinesh people. A total of 10,436 participants who had a health check-up from January 2017 to July 2017 were recruited. With appropriate data mining approaches, 3454 participants remained in the final dataset for further analysis. Seventy percent of these participants (2420 cases) were then randomly allocated to either the training dataset for the construction of the decision tree or the testing dataset (30%, 1034 cases) for evaluation of the performance of the decision tree. The proposed approach achieved an accuracy of classification of 90.3% with a precision of 89.7% and a recall of 90.3%. Table 2 represent the related work on diagnosis of diabetes based on Decision Tree (DT) algorithm.

3.1.3 K nearest neighbour (KNN)

In 2010, Lee et al. [175] proposed a monitoring and advisory system for diabetes patient management using a Rule-Based method and KNN. This paper proposes a system that can provide appropriate management for diabetes patients, according to their blood sugar level. The system is designed to send the information about the blood sugar levels, blood pressure, food consumption, exercise, etc., of diabetes patients, and manage the treatment by recommending and monitoring food consumption, physical activity, insulin dosage, etc., so that the patient can better manage their condition. The system is based on rules and the K Nearest Neighbor (KNN) classifier algorithm, to obtain the optimum treatment recommendation. Also, a monitoring system for diabetes patients is implemented using Web Services and Personal Digital Assistant (PDA) programming.

In 2015, Farahmandian et al. [181] introduced a case study on data mining algorithms which are crucial in the diagnosis and prediction of diabetes. In this work Support Vector Machine (SVM), K Nearest Neighbors (KNN), Naïve Bayes, ID3, C4.5, C5.0, and CART algorithms are used. Evaluation and conclusion of data mining algorithms which contain 768 records of different patients have been carried out on Pima dataset. Results have shown that the degree of Accuracy in SVM algorithm equals to 81.77.

In 2018, Dey et al. [186] implement a Web Application to Predict Diabetes Disease using Machine Learning Algorithm. This work consists of development of an architecture which has the capability to predict where the patient has diabetes or not. The main aim of this exploration is to build a web application based on the higher prediction accuracy of some powerful machine learning algorithm. It used a benchmark dataset namely Pima Indian which is capable of predicting the onset of diabetes based on diagnostics manner. With an accuracy of 82.35% prediction rate Artificial Neural Network (ANN) shows a significant improvement of accuracy. The proposed model achieved an accuracy of 66.5% using KNN.

In 2020, Gupta et al. [195] introduced a Performance enhancement of diabetes prediction by finding optimum K for KNN classifier with feature selection method. The proposed work KNN and machine learning methods are used in the prediction model to classify whether the patient is diabetic or non-diabetic. The PIMA diabetes dataset is used for research purpose in the python implemented model. A research study has been performed to improve the performance of the KNN classifier by using a feature selection method, normalization and considering the different number of neighbors. The performance of classifier is measured based on different metrics such as accuracy, precision, sensitivity, specificity, f1 score and error rate. The best performance of KNN is achieved when no of neighbors (K) is 33, 40 or 45. The accuracy and error rate is same on these K and it is 87.01% and 12.99 % respectively, while a little variation is shown in other metric's values.

In 2018, Sarkar et al. [197] proposed a K-Nearest Neighbor Learning based Diabetes Mellitus Prediction and Analysis for eHealth Services. The proposed work consists of optimal K Nearest Neighbor (Opt-KNN) learning based prediction model based on the patient's habitual attributes in various dimensions. This approach determines the optimal number of neighbors with low error rate for providing better prediction outcome in the resultant model. The effectiveness of this machine learning eHealth model is examined by conducting experiments on the real-world diabetes mellitus data collected from medical hospitals. The setting of this K-value should be optimized according to the patterns in the dataset. It achieved lowest error rate when K=3. Thus, proposed Opt-KNN based prediction model dynamically selects K=3 as an optimal value to build an effective disease risk prediction model.

In 2021, Mohanty et al. [199] developed a KNN based prediction model for diabetic patients. The proposed work machine learning based algorithm is used to figure out various patterns in our dataset and to calculate the accuracy of this data, with hope that this serves as a stepping stone towards developing tools that can help in medical diagnosis/treatment in future. Creating an efficient diagnostic tool will help improve healthcare to a great extent. The fundamental factors considered in this dataset are age, gender, region of stay and Blood groups. The data should be 98 % accurate for it to be acceptable in real time diagnostic tool development.

In 2021, Patra et al. [200] introduced an Analysis and Prediction of Pima Indian Diabetes Dataset using the SDKNN Classifier Technique. The proposed technique is based on a new distance calculation formula to find nearest neighbors in KNN. It utilized standard deviation of attributes as a powerful tool to measure the distance between train dataset and test dataset. This concept is applied on Pima Indian Diabetes Dataset (PIDD). The analysis is carried out on data set by splitting 90% of training data and 10% of testing data. The proposed approach achieved an accuracy rate of 83.2%, which shows better improvement as compared to the other technique. Table 3 represents the related work on diabetes based on KNN.

3.1.4 Naive Bayes (NB)

In 2010, Sopharak et al. [202] proposed a Machine learning approach for automatic exudate detection in retinal images from diabetic patients. The proposed work consists of a series of experiments on feature selection and exudates classification using naive Bayes and support vector machine (SVM) classifiers. First fit the naive Bayes model to a training set consisting of 15 features extracted from each of 115,867 positive examples of exudate pixels and an equal number of negative examples. Perform feature selection on the naive Bayes model, repeatedly removing features from the classifier, one by one, until classification performance stops improving. To find the best SVM, begin with the best feature set from the naive Bayes classifier, and repeatedly add the previously-removed features to the classifier. The result reveals that the naive Bayes and SVM classifiers perform better than the NN classifier. The overall best sensitivity, specificity, precision, and accuracy are 92.28%, 98.52%, 53.05%, and 98.41%, respectively.

In 2013, Lee [206] introduced a prediction of fasting plasma glucose status using anthropometric measures for diagnosing of Type 2 Diabetes. This study aims to predict the fasting plasma glucose (FPG) status that is used in the diagnosis of type 2 diabetes by a combination of various measures among Korean adults. A total of 4870 subjects (2955 females and 1915 males) participated in this study. Based on 37 anthropometric measures, we compared predictions of FPG status using individual versus combined measures using two machine-learning algorithms. The values of the area under the receiver operating characteristic curve in the predictions by logistic regression and naive Bayes classifier based on the combination of measures were 0.741 and 0.739 in females, respectively, and were 0.687 and 0.686 in males, respectively.

Table 2 (Decision Tree based) Diabetes diagnosis summary

Author, Year, References	Algorithms	Experimental Results	Datasets
Breault, 2002 [102]	CART	The first level of the tree shows that just dividing people using an age cut-point of 65.581 years of age, 19.4% of younger people (n ¼ 3987) have a bad HgbA1c. This is 2.8 times the rate of bad HgbA1c values in those who are older (7.0%, n ¼ 3966).	442 bed tertiary care hospital, a 500 physician multi-specialty clinic in 25 locations.
Miyaki, 2002 [103]	CART	The results indicate that the risk factors which type 2 diabetic patients must focus on have an order of priority, and that they differ with patient category.	165 type 2 diabetic patients from the Keio University Hospital.
Duhamel, 2003 [104]	C4.5	The imputation using decision trees performed better than imputation by mode for 6 of the 18 variables analyzed. For other variables, whether no model could be found or, if such a model existed, it provided results similar or poorer than imputation by mode. When the missing values rate was high (>10%) or when the number of categories was high (>_4), imputation by decision trees failed.	23601 records corresponding to the non-insulin-dependent type II diabetic patients.
Haung, 2004 [105]	C4.5	Before feature selection, discretized C4.5 had the best performance of classification. And after feature selection C4.5 obtained the best result.	Ulster community and Hospital Trust
Harper, 2005 [106]	CART	Overall the results were promising and each tool made a statistically significant contribution in each study (values of r and the percentage correctly classified were all significant at the 95% level).	diabetic dataset collected for over 30 years from a leading unit in the UK.

Table 2 (continued)

Author, Year, References	Algorithms	Experimental Results	Datasets
Sigurdardottir, 2007 [107]	C4.5	Of 464 titles extracted, 21 articles reporting 18 studies met the inclusion criteria. Data mining showed that for initial glycosylated hemoglobin (HbA1c) level 7.9% the diabetes education intervention achieved a small change in HbA1c level, or from +0.1 to 0.7%. For initial HbA1c 8.0%, a significant drop in HbA1c level of 0.8–2.5% was found.	Extracted from Medline and Scopus using educational intervention
Huang, 2007 [108]	C4.5	Using the reduced features, a best predictive accuracy of 95% and sensitivity of 98% was achieved.	Ulster Community and Hospitals Trust (UCHT)
Liou, 2008 [109]	CART	The classification model performs the best with an overall identification rate of 99%.	NHI Diabetic dataset.
Toussi, 2009 [110]	C5.0 decision-tree	C5.0 models are robust in the presence of problems such as missing data and large numbers of input fields. They tend to be easier to understand than some other model types, because the interpretation of the rules derived from the model is very straightforward. C5.0 models also allow the use of a powerful boosting method for increasing the accuracy of classification.	Avicenna University Hospital of Bobigny, France.
Hische, 2010 [111]	Decision Tree	A clinical decision tree included age and systolic blood pressure (sensitivity 89.3%, specificity 37.4%, and positive predictive value (PPV) 48.0%), while a tree based on clinical and laboratory data included fasting glucose and systolic blood pressure (sensitivity 89.7%, specificity 54.6%, and PPV 56.2%).	1737 individuals of the cross-sectional Metabolic Syndrome Berlin Potsdam

Table 2 (continued)

Patil, 2010 [112]	C4.5	A classification accuracy of 92.38% is achieved. In order to evaluate the performance of the proposed method, sensitivity and specificity performance measures that are used commonly in medical classification studies were used.	Pima Indians diabetes dataset.
Ahmad, 2011 [113]	C4.5 (J48)	Results showed that a pruned J48 tree performed with higher accuracy, which is 89.3% as compared to 81.9% by the multi-layer perceptrons.	Pima Indians diabetes dataset.
AlJarullah, 2011 [114]	C4.5 (J48)	566 instances were correctly classified representing 78.1768 % of total records and 158 instances were incorrectly classified representing 21.8232 % of total records.	Pima Indians diabetes dataset.
Karegowda, 2011 [115]	Decision Tree	The results prove that, GA-optimized BPN approach has outperformed the BPN approach without GA optimization. In addition the hybrid GA-BPN with relevant inputs lead to further improvised categorization accuracy compared to results produced by GA-BPN alone with some redundant inputs.	Pima Indians diabetes dataset.
Kelarev, 2012 [116]	ADTree, J48, NBTree, RandomTree, REPTree and SimpleCart.	The best outcomes have been obtained by the novel combined ensemble of AdaBoost (accuracy=94%) and Bagging (accuracy=92.99%) based on J48.	Diabetes Complications Screening Research Initiative
Hemant, 2012 [117]	C4.5(J48),Adaboost	The best accuracy for the given dataset is achieved in bagging algorithm compared to other classifirs. The proposed approach helps doctors in their diagnosis de decisions and also in their treatment planning procedures for different categories.	Dataset consists of 768 different entries.

Table 2 (continued)

Hussein, 2012 [118]	J48, Decision Stump, REP Tree and RF	RF has demonstrated 99.7% of correctly classified cases.	935 records has been collected from hospitals in Oman.
Rajesh, 2012 [119]	C4.5	A classification rate of 91% was obtained for C4.5 algorithm.	Pima Indians diabetes dataset.
Chang-ping, 2012 [120]	LR,DT,MLP	The AUC of DT, MLP and LR were 0.8863, 0.8536 and 0.8802, respectively. As the larger the AUC of a specific prediction model is, the higher diagnostic ability presents, MLP performed optimally, and then followed by LR and DT in terms of 10–100 times 2–10 fold stratified cross-validation.	274 patients from the Metabolic Disease Hospital in Tianjin
Chen, 2012 [121]	Fisher linear discriminate analysis (FLDA), SVM and DT (CART).	Both SVM and CART are superior to those from LDA and no significant difference between SVM and CART can be seen.	Type-2 diabetes group from 58 patients. Shenyang area, northeast of China.
Karegowda, 2012 [122]	C4.5	The proposed cascaded model with categorical data obtained the classification accuracy of 93.33 % when compared to accuracy of 73.62 % using C4.5 alone for PIMA Indian diabetic dataset.	Pima Indian diabetic database
Karthikeyani, 2012 [123]	C4.5, SVM and K-NN.	Achieved Accuracy level of 86%, 74.8% and 78% for C4.5, SVM and KNN respectively.	ICMR-INDIAB
Ameri, 2013 [124]	Decision Tree (C5.0)	The accuracy of the C5.0 model on the data was shown to be 89.74% and on the Artificial Neural Network was 51.28%.	856 patient records related to the 2009’s cases in the Diabetes Center
Meng, 2013 [125]	C5.0	The ANN model reached a classification accuracy of 73.23% with a sensitivity of 82.18% and a specificity of 64.49%; and the decision tree (C5.0) achieved a classification accuracy of 77.87% with a sensitivity of 80.68% and specificity of 75.13%. The decision tree model (C5.0) had the best classification accuracy, followed by the logistic regression model, and the ANN gave the lowest accuracy.	Dataset from Guangzhou, China; 735 patients confirmed to have diabetes.

Table 2 (continued)

Karthikeyani, 2013 [126]	C4.5	Achieved C4.5 with 72% accuracy, SVM with 70.66%.	Pima Indians diabetes dataset.
Rahman, 2013 [127]	C4.5	The best algorithm in WEKA is J48graft classifier with an accuracy of 81.33% that takes 0.135 seconds for training.	Pima Indians diabetes dataset.
Varma, 2014 [128]	fuzzy decision tree GG-FSDT	The average testing accuracy of the proposed method is 75.8%.	Pima Indians diabetes dataset.
Kaur, 2014 [129]	Modified J48	Proposed algorithm has large accuracy difference than other algorithms. It has accuracy rate of 99.87% rather than others that show maximum of 77.21%accuracy.	Pima Indians diabetes dataset.
Seera, 2014 [130]	CART,RF	The experimental outcomes positively demonstrate that the hybrid intelligent system is effective in undertaking medical data classification tasks. More importantly, the hybrid intelligent system not only is able to produce good results but also to elucidate its knowledge base with a decision tree.	Pima Indians diabetes dataset.
Uppin, 2014 [131]	C4.5	83.63% accuracy and total time to build the model is 0.03 in diagnosis of diabetes.	Pima Indians diabetes dataset.
Ramezankhani, 2014 [132]	Decision Tree	The overall classification accuracy was 90.5%, with 31.1% sensitivity, 97.9% specificity; and for the subjects without diabetes, precision and f-measure were 92% and 0.95, respectively.	Tehran Lipid and Glucose Study (TLGS) database.
Bashir, 2014 [133]	Decision Tree with Ensemble Technique.	Experiments are conducted on Biostat and Pima Indian diabetes datasets. Evaluation of results indicates that Bagging ensemble outperforms other techniques for both the diabetes datasets.	Pima Indians diabetes dataset.
Habibi, 2015 [134]	C4.5 (J48)	The precision and accuracy of the model was 71.7 and 97.6 percent, respectively.	Diabetes control system in Tabriz, Iran

Table 2 (continued)

Kandhasamy, 2015 [135]	J48 Decision Tree, K-Nearest Neighbors, and Random Forest, Support Vector Machines.	Decision tree J48 classifier achieves higher accuracy of 73.82 % than other three classifiers. In other case that is after pre-processing the dataset we have more accurate result when compared to the first study. In this case, both KNN (k=1) and Random Forest performance much better than the other three classifiers and they provide 100% accuracy.	http://mldata.org/repository/data/viewslug/datasets-uci-diabetes
Iyer, 2015 [136]	Decision Tree (J48) and Naïve Bayes	From the results obtained, both the methods have a comparatively small difference in error rate, though the percentage split of 70:30 for Naïve Bayes technique gives the least error rate as compared to other two J48 implementations. Both the models are efficient in the diagnosis of diabetes using the percentage split of 70:30 of the data set.	Pima Indians diabetes dataset.
Vijayan, 2015 [137]	AdaBoost algorithm with Decision Stump.	The accuracy obtained for AdaBoost algorithm with decision stump as base classifier is 80.72% which is greater compared to that of Support Vector Machine, Naive Bayes and Decision Tree	University of California, Irvin repository of machine learning
Thirumal, 2015 [138]	Naïve Bayes, Decision tree (c4.5), k-Means, SVM and kNN.	C4.5 algorithm outperforms the other algorithms with the accuracy of 78.25%. Naïve Bayes ranks second with accuracy of 77.8 % and k-NN scores 77.73%, finally SVM acquired 77.4% as accuracy.	Pima Indians diabetes dataset.
Nongyao, 2015 [139]	Decision Tree, ANN, LR, Naive Bayes and RF	The accuracy rate of the proposed technique is (DT=85%, ANN=84.53%, LR=82.3%, Naive Bayes=81% and RF=85.55), Findings suggest that the best performance of disease risk classification is Random Forest algorithm.	26 Primary Care Units (PCU) in Sawanpracharak Regional Hospital.

Table 2 (continued)

Heydari, 2016 [140]	SVM, ANN, DT, nearest neighbors, and Bayesian network	Artificial neural network with an accuracy rate of 97.44 % has the best performance on the chosen dataset. Accuracy rates for support vector machine, decision tree, 5-nearest neighbor, and Bayesian network are 81.19, 95.03, 90.85, and 91.60 %, respectively.	2536 cases screened for type 2 diabetes, in the city of Tabriz, Iran.
Ahmed, 2016 [141]	J48	According to the results, the accuracy of the model was 70.8% and all others measures of validation were accepted.	JABER ABN ABU ALIZ clinic center for diabetes in Sudan.
Ahmed, 2016 [142]	Naive Bayse, logistic and J48	Based on the results of experiment, Logistic algorithm has been selected as best one with high accuracy rate of 74.8%.	Cerner Corporation, Kansas City, MO
Daghistani, 2016 [143]	C4.5 and Random Forest	C4.5 and Random Forest achieved Recall and Precision over 90% on the training data set while SOM was able to achieve Recall and Precision over 79% on the training data set.	Ministry of National Guard Health Affairs (MNGHA), Saudi Arabia
Orabi, 2016 [144]	Decision Tree	The result of the second experiment on numerical age data was promising as the best rotation predicted with average accuracy 84 %.	Egyptian National Research Center
Perveen, 2016 [145]	J48 (c4.5)	Experimental result shows that, overall performance of Adaboost ensemble method is better than bagging as well as standalone J48 decision tree.	Canadian Primary Care Sentinel Surveillance network
Pradeep, 2016 [146]	J48	The proposed approach achieved the best results in terms of accuracy and precision.	UCI repository of machine learning database.
Shetty, 2016 [147]	ID3	The correctly classified instance is 94% and incorrectly classified is 6% for ID3 algorithm.	150 records and 14 attributes with one class attribute.

Table 2 (continued)

Songthung, 2016 [148]	CHAID (Chi-squared Automatic Interaction Detector) Decision Tree	Risk scoring has a coverage that is lower than classifiers such as Naive Bayes and Decision Tree whose coverage is almost 100%. Naive Bayes has good coverage and good high-risk percentages compared to both risk scoring and Decision Tree.	Ministry of Public Health (MoPH),Thailand
Srikanth, 2016 [149]	Decision Tree	True positive rate of (J48=0.48, ADTree=0.79 and BFTree =0.77)	Pima Indians diabetes dataset.
Teimouri, 2016 [150]	Decision Tree, Support Vector Machine, Naive Bayes, Neural Network	Support Vector Machine with an accuracy of 95.32% shows better performance than the other methods.	Data collected from 1412 prescriptions with 414 kinds of drugs
Chen, 2017 [151]	J48	The proposed approach of K-mean clustering with J48 classifier achieved the accuracy level of 90.04%.	Pima Indian Diabetes Dataset.
Kasbekar, 2017 [152]	C5.0	The results obtained show an accuracy of 96.4% in the primary group and an accuracy of 94% in the test group.	Hospital records of 301 diabetic foot patients
Sayadi, 2017 [153]	J48	The results showed that T2DM could be predicted via decision tree model without laboratory tests. This model can be used in pre-clinical and public health screening programs.	11302 cases from the database of Healthy Heart House of Shiraz, Iran.
Yuvaraj, 2019 [154]	SVM, DT, RF, Naive Bayes.	From the result statistics obtained it is very clear that hadoop cluster based random forest algorithm performs much better in terms of all the different performance measures when compared to the other two machine learning algorithms.	National Institute of Diabetes from 75,664 patients.

Table 2 (continued)

Zou, 2018 [155]	J48,NN,RF	There is not much difference among random forest, decision tree and neural network, but random forests are obviously better than other classifiers in some methods. The best result for Luzhou dataset is 0.8084, and the best performance for Pima Indians is 0.7721.	Luzhou and Pima Indians Dataset.
Kadhm, 2018 [156]	Decision Tree	By experiments, the proposed system achieved high classification result which is 98.7% comparing to the existing system using Pima Indians diabetes (PID) dataset.	Pima Indian Diabetes Dataset.
Esmaily, 2018 [157]	DT,RF	The decision tree model has 64.9% accuracy, 64.5% sensitivity, 66.8% specificity, and area under the ROC curve measuring 68.6%, while the random forest model has 71.1% accuracy, 71.3% sensitivity, 69.9% specificity, and area under the ROC curve measuring 77.3% respectively.	9528 subjects recruited from MASHAD database.
Barhate, 2018 [158]	K-NN, LR, DT, RF, Gradient Boosting, SVM and Neural Network.	Results demonstrate that Random Forests performed well on the data set giving an accuracy of 79.7%.	Pima Indian Diabetes Dataset.
Mahmud, 2018 [159]	ANN,SVM,LR,DT,RF,NB	NB and SVM outperformed the other classification techniques in terms of accuracy by obtaining the highest accuracy as 74% and 73%, respectively. However, the artificial neural network exhibits lowest performance than the other classification algorithms.	Pima Indian Diabetes Dataset.

Table 2 (continued)

Fiarni, 2019 [160]	Naive Bayes, J48	Given the accuracy of the proposed model is 68%, with the highest accuracy on Retinopathy prediction model. This mean the model could be implemented on automatic system, even though more effort needs to increase these current accuracy levels.	Sri Pamela Hospital and Kumpulan Pane Hospital, Indonesia.
Hebbar, 2019 [161]	DT,RF	The proposed algorithm is evaluated on real-life data set. The results show the accuracy of 72% for Decision Tree and 76.5% Random Forest, respectively.	Data set consisting of 768 observations capturing nine attributes for each.
Pei, 2019 [162]	J48	The accuracy, precision, recall, and area under the receiver operating characteristic curve (AUC) value for identifying potential diabetes were 94.2%, 94.0%, 94.2%, and 94.8%, respectively.	China Medical University(10,436 records)
Chaudhary, 2019 [163]	Svm, k-nearest neighbors, decision tree, naïve Bayes approach, and logistic regression.	The performance analysis is analyzed in terms of accuracy rate among all the classification methods such as decision tree, logistic regression, k-nearest neighbors, naïve Bayes, and SVM. It is found that logistic regression gives the most accurate results to classify the diabetic and non diabetic samples.	Pima Indian Diabetes Dataset.
Sun, 2019 [164]	Logistic Regression (LR), Support Vector Machines (SVM), Random Forests (RF), Decision Trees (DT), Naïve Bayes (NB).	The experimental results show that random forest (RF) in the machine learning model can get 92% accuracy with good performance.	Medical Big Data Center of the 301 Hospital, 4 million records.

Table 2 (continued)

Choubey, 2020 [165]	Logistic Regression, K-Nearest Neighbor, ID3 DT, C4.5 DT, and Naive Bayes.	Performance on the basis of Pima Indian Diabetes Dataset, the descending order are PCA_Logistic Regression, PSO_Logistic Regression, Logistic Regression, PCA_ Naive Bayes, PSO_Naive Bayes, Naive Bayes, C4.5 DT, PCA_ID3 DT, PSO_ID3 DT, ID3 DT, PCA_C4.5 DT, PSO_C4.5 DT, PCA_KNN, PSO_KNN, KNN.	Pima Indian Diabetes Dataset.
Al- Zebari, 2019 [166]	(DT), (LR), Discriminant Analysis (DA), (SVM), (k-NN) and ensemble learners.	The best accuracy score 77.9% is produced by the LR method and the worst one 65.5% is produced by the Coarse Gaussian SVM technique.	Pima Indian Diabetes Dataset.
Pei, 2020 [167]	Decision Tree (J48)	Performance of all classifirs, and shows that J48 exhibits better results than others (accuracy=0.903, precision=0.897, recall=0.903, F-measure=0.899, and AUC=0.872).	10,436 participants who had a health check-up from January 2017 to July 2017.
Maniruzzaman, 2020 [168]	Naïve Bayes (NB), decision tree (DT), Adaboost (AB), and random forest (RF).	The overall ACC of ML-based system is 90.62%. The combination of LR-based feature selection and RF-based classifier gives 94.25% ACC and 0.95 AUC for K10 protocol.	National Health and Nutrition Examination Survey
Pranto, 2020 [169]	Decision tree, K-nearest neighbor, random forest, and Naive Bayes.	The results show that both random forest and Naïve Bayes classifier performed well on both datasets.	Pima Indian Diabetes Dataset.
Tigga, 2020 [170]	Logistic Regression; kNN; SVM; Naïve Bayes; DT; Random Forest.	The experimental result shows that the accuracy of Random Forest of our dataset is 94.10% which is the highest among the rest. Random forest is also giving highest accuracy for PIMA dataset.	Pima Indian Diabetes Dataset.

Table 2 (continued)

Haq, 2020 [171]	ID3	The proposed method DT (ID3) +DT achieved 99% test accuracy, 99.8% accuracy with k-floods and 99.9% accuracy with LOSO validation.	online: https://www.kaggle.com/johndasilva/diabetes.
Taser, 2021 [172]	C4.5, random tree, REPTree, decision stump, Hoeffding tree, and NBTree.	The results indicate that the bagging and boosting approaches outperform the individual DTB classifiers, and real Adaptive Boosting (AdaBoost) and bagging using Naive Bayes Tree (NBTree) present the best accuracy score of 98.65%.	Sylhet Diabetes Hospital of Sylhet, Bangladesh.
Chen, 2021 [173]	CART	Classification accuracy of the proposed approach is 75.36%.	Pima Indian Diabetes Dataset.
Emon, 2021 [174]	J48	The overall performance of logistic regression shows the best result.	UCI machine learning repository.
Ahamed, 2022 [411]	Decision Tree	LGBM classifier has the highest accuracy of 95.20% in comparison with the other algorithms.	PIMA
Özge, 2023 [412]	Decision Tree	The highest accuracy values were obtained with the Extra Trees algorithm with 99.2%.	520 patients.

In 2016, Songthung et al. [148] proposed a novel approach to enhance the Type 2 Diabetes Mellitus Risk Prediction using different machine learning technique. The proposed work consists of an extensive dataset gathered from 12 hospitals in Thailand during 2011-2012 with 22,094 records of screened population who are females age 15 years or older. This study used RapidMiner Studio 7.0 with Naive Bayes and CHAID (Chi-squared Automatic Interaction Detector) Decision Tree classifiers to predict high risk individuals and compared the results with existing hand-computed diabetes risk scoring mechanisms. The result shows that Naive Bayes has good coverage and good high-risk percentages compared to both risk scoring and Decision Tree.

In 2018, Das et al. [209] introduced an approach for classification of diabetes mellitus disease using data mining technique. The aim of this research is to predict diabetes based on some of the DM techniques like classification and clustering. Out of which, classification is one of the most suitable methods for predicting diabetes. In this study, J48 and Naïve Bayesian techniques are used for the early detection of diabetes. The experimental results based on data from 200 patient records reveal that Naive Bayes algorithm is better than the J48 as the time to build the model is less.

In 2019, Khan et al. [213] introduced a machine learning based intelligent system for predicting diabetes. The objective of this research is to propose an intelligent system based on a machine learning algorithm to improve the accuracy of predicting diabetes. To attain this objective firstly, an algorithm was proposed based on Naïve Bayes with prior clustering. Secondly, the performance of the proposed algorithm was evaluated using 532 data related to diabetic patients. Finally, the performance of the existing Naïve Bayes algorithm was compared with the proposed algorithm. The results of the comparative study showed that the improvement in the accuracy has been made apparent for the proposed algorithm.

In 2021, Jackins et al. [216] proposed an AI based smart prediction of clinical disease using random forest classifier and Naive Bayes. For diabetes data, the Naive Bayes algorithm gives 76.72 and 74.46 accuracies for training and test data, respectively. Random forest algorithm gives 98.88 and 74.03 for training and test data, respectively. Performance analysis of the disease data for both algorithms is calculated and compared. The results of the simulations show the effectiveness of the classification techniques on a dataset, as well as the nature and complexity of the dataset used.

In 2020, Rghioui et al. [218] introduced a smart glucose monitoring system for diabetic patient. The proposed work presents an intelligent architecture for the surveillance of diabetic disease that will allow physicians to remotely monitor the health of their patients through sensors integrated into smartphones and smart portable devices. The proposed architecture includes an intelligent algorithm developed to intelligently detect whether a parameter has exceeded a threshold, which may or may not involve urgency. To verify the proper functioning of this system developed a small portable device capable of measuring the level of glucose in the blood for diabetics and body temperature. The evaluation result showed that the system using the J48 algorithm exhibited excellent classification with the highest accuracy of 99.17%, a sensitivity of 99.47% and a precision of 99.32%. Table 4 represents the related work on diabetes based on NB.

3.1.5 Random Forest (RF)

In 2020, Wang et al. [230] introduced a Prediction of medical expenditures of diagnosed diabetics and the assessment of its related factors using a random forest model. In this work data were collected from the US household component of the medical expenditure panel survey, 2000–2015. Random forest (RF) model was performed with the programs of the random Forest in R software. Spearman correlation coefficients (rs), mean absolute error (MAE) and mean-related error (MRE) was computed to assess the prediction of all the models. The experimental result indicated that the RF model was little superior to traditional regression model. RF model could be used in prediction of medical expenditure of diabetics and an assessment of its related factors well.

In 2021, Wang et al. [231] present an exploratory study on classification of diabetes mellitus through a combined Random Forest Classifier. This study explored different supervised classifiers, combined with SVM-SMOTE and two feature dimensionality reduction methods (Logistic stepwise regression and LAASO) to classify the diabetes survey sample data by unbalanced categories and complex related factors. Analysis and discussion of the classification results of 4 supervised classifiers based on 4 data processing methods. Five indicators, including Accuracy, Precision, Recall, F1-Score and AUC are selected as the key indicators to evaluate the performance of the classification model. According to the result, Random Forest Classifier is combining SVM-SMOTE resampling technology and LASSO feature screening method (Accuracy= 0.890, Precision = 0.869, Recall = 0.919, F1-Score = 0.893, AUC= 0.948) proved the best way to tell those at high risk of DM. Besides, the combined algorithm helps enhance the classification performance for prediction of high-risk people of DM.

In 2021, Ooka et al. [232] present a Random forest approach for determining risk prediction and predictive factors of type 2 diabetes for the peoples of Japan. This study included a cumulative total of 42 908 subjects not receiving treatment for diabetes with an HbA1c <6.5%. It used two analytical methods to compare the predictive powers: RF as a new model and multivariate logistic regression (MLR) as a conventional model. The RF model showed a higher predictive power for the change in HbA1c than MLR in all models. The RF model, including change values showed the highest predictive power. Table 5 represents the related work on diabetes based on RF.

3.2 Supervised learning (Regression)

In 2009, Gani et al. [261] introduced a data-driven model based on glucose data from one diabetic subject, and subsequently applied to predict subcutaneous glucose concentrations of other subjects, even of those with different types of diabetes. This work employed three separate studies, each utilizing a different continuous glucose monitoring (CGM) device, to verify the model's universality. The predictive capability of the models was found not to be affected by diabetes type, subject age, CGM device, and inter individual differences.

In 2012, Georga et al. [264] present a predictive modeling of subcutaneous (s.c.) glucose concentration in type 1 diabetes. In this work support vector regression (SVR) technique is utilized. The proposed method is evaluated using a dataset of 27 patients in free-living conditions. Tenfold cross validation is applied to each dataset individually to both optimize and test the SVR model. In the case, where all the input variables are considered, the average prediction errors are 5.21, 6.03, 7.14, and 7.62 mg/dl for 15-, 30-, 60-, and 120-min prediction horizons, respectively. The results clearly indicate that the availability of multivariable data and their effective combination can significantly increase the accuracy of both short-term and long-term predictions.

In 2015, Paul et al. [268] proposed a technique of linear auto-regressive (AR) and state space, time series models to analyze the glucose profiles for predicting upcoming glucose levels. However, these modelling approaches have not adequately addressed the inherent dependencies and volatility aspects in the glucose profiles. The prediction performances of GARCH approach were compared with other contemporary modelling approaches such as lower and higher order AR, and the state space models. The GARCH approach appears to be successful in both realizing the volatility in glucose profiles and offering potentially more reliable forecasting of upcoming glucose levels.

In 2018, Wu et al. [279] introduced a novel model based on data mining techniques for predicting type 2 diabetes mellitus (T2DM). The model is comprised of two parts, the improved K-means algorithm and the logistic regression algorithm. The Pima Indians Diabetes Dataset and the Waikato Environment for Knowledge Analysis toolkit were utilized to compare the results with the results from other researchers. The conclusion shows that the model attained a 3.04% higher accuracy of prediction than those of other researchers. Moreover, our model ensures that the dataset quality is sufficient.

In 2019, Qiu et al. [280] present an improved prediction method for diabetes based on a feature-based least angle regression algorithm. This work consists of a method based on feature weights to improve diabetes prediction that combines the advantages of traditional least angle regression (LARS) algorithms and principal component analysis (PCA) algorithms. First of all, a principal component analysis algorithm is used to obtain the characteristic independent variables found in typical diabetes prediction regression models. The experimental results show that the algorithm improved the approximation speed for the dependent variables and the accuracy of the regression coefficients.

In 2019, Yao et al. [281] proposed a multivariable logistic regression and back propagation artificial neural network based model to predict diabetic retinopathy. A total of 530 Chinese residents, including 423 with type 2 diabetes (T2D) aged 18 years or older participated in this study. In this work a back propagation artificial neural network (BP-ANN) model is utilized by selecting tan-sigmoid as the transfer function of the hidden layers nodes, and pure-line of the output layer nodes, with training goal of 0.5×10−5. Based on these parameters, the area under the receiver operating characteristic (ROC) curve for the BP-ANN model was significantly higher than that by MLR (0.84 vs. 0.77, $P < 0.001$).

In 2020, Alshamlan et al. [282] introduced a gene prediction function for type 2 diabetes mellitus using logistic regression. In this study the process of feature selection is performed using the Fisher score and chi-square approaches. The total selected number of genes ranges from 1800-2700.The experimental results show that shows that logistic regression produces the highest accuracy with the fisher score for GSE38642 dataset with 90.23% and GSE13760 dataset with 61.90%. Feature selections with logistic regression, classification were used. The obtained accuracy result of logistic regression on two datasets based on fisher score feature selection was higher than Ch-2 feature selection. The accuracy results of two data were 90.23% and 61.90% respectively.

In 2020, Kopitar et al. [283] proposed an early detection of type 2 diabetes mellitus using machine learning based prediction models. This study compares machine learning-based prediction models (i.e. Glmnet, RF, XGBoost, LightGBM) to commonly used regression models for prediction of undiagnosed T2DM. With 6 months of data available, simple regression model performed with the lowest average RMSE of 0.838, followed by RF (0.842), LightGBM (0.846), Glmnet (0.859) and XGBoost (0.881). When more data were added, Glmnet improved with the highest rate (+ 3.4%). The aim of this study was to investigate whether novel machine learning-based approaches offered any advantages over standard regression techniques in the early prediction of impaired fasting glucose (IFG) and fasting plasma glucose level (FPGL) values. Table 6 represents the related work on diabetes based on regression technique.

Table 3 (KNN based) Diabetes Diagnosis summary

Author, Year, References	Algorithms	Experimental Results	Datasets
Lee, 2010 [175]	KNN	The prescription algorithm method based on rules has the disadvantage that is not able to select any treatment method without previous knowledge about diabetes. However, the prescription algorithm using the method based on the KNN selects the optimum treatment method simply by using the given sample data.	300 sets of experimental data.
Chikh, 2012 [176]	Fuzzy KNN	The highest classification accuracy obtained when applying the AIRS2 and MAIRS2 using 10-fold cross-validation was, respectively 82.69% and 89.10%.	UCI machine learning repository
Aslam, 2013 [177]	GP-KNN	GP generated features (GP-KNN, GP-SVM) outperform the original diabetes features (KNN, SVM). The accuracy rate of GP-KNN is 80.5%.	Pima Indian Diabetes Dataset.
Christobel, 2013 [178]	CKNN	The Performance of classification measured with respect to sensitivity, specificity and accuracy has been increased significantly in the case of proposed CkNN algorithm.	Pima Indian Diabetes Dataset.
NirmalaDevi, 2013 [179]	K-mean, K-NN	Experimental results signify the proposed amalgam KNN along with preprocessing produces best result for different k values. If k value is more the proposed model obtained the classification accuracy of 97.4%.	Pima Indian Diabetes Dataset.
Sarwar, 2014 [180]	ANN, Naive Bayes, KNN	The results indicate that the ANN is the best predictor with the accuracy of about 96 % which was followed by Naıve Bayes networks having an accuracy of about 95 % and the KNN came to be the worst predictor having an accuracy of about 91 %.	500 people chosen randomly.
Farahmandian, 2015 [181]	SVM, KNN, ID3	SVM Model was the most accurate of all in diagnosing diabetes, and ID3 was the least accurate in diagnosis.	Pima Indian Diabetes Dataset.

Table 3 (continued)

Author, Year, References	Algorithms	Experimental Results	Datasets
Hidalgo, 2017 [182]	KNN	RF and KNN perform well in the forecasting of blood glucose concentration up to two hours ahead using only past glucose values as well as past and expected future insulin and carbohydrate inputs. 90% of the forecasts are considered correct.	Ten T1DM patients (n = 10) have been selected for observation.
Kumar, 2017 [183]	KNN	As a result, Random Forest algorithm gives better accuracy in comparison with LDA, SVM, CART and k-NN. The accuracy value of Random Forest algorithm is 0.99 which is the best among all other algorithms.	Diagnosis Lab located in Warangal, TS, India.
Aiello, 2018 [184]	KNN	Satisfactory results have been obtained in terms of reduction of the average glucose and of hyperglycemia, and in terms of increment of the time in target with limited increase of hypoglycemia.	Silico population of the UVA/PADOVA simulator.
Mittal, 2019 [185]	KNN	Traditional KNN is simple effective and non parametric widely used for classification but it may not be effective for large scale database or data having many categories. To overcome this problem a clustering based KNN approach is introduced that adopts a dynamic adjustment in each iteration for the neighbor number K. Thus, choosing the value of K dynamically the classification accuracy is enhanced thereby avoiding uneven classification phenomenon.	UCI Machine Learning Repository prepared by University of Wisconsin Hospitals
Dey, 2018 [186]	KNN	The experimental show accuracy rate of KNN is 66.5%.	Pima Indian Diabetes Dataset
Azrar, 2018 [187]	KNN	Accuracy obtained from Decision Tree is highest yet the graph is more dispersed that can be enhanced too. Lowest accuracy is from KNN. KNN is tested with wide range of K values from 1 to 10 and with changing folds from 10 to 20 but still accuracy is not that much.	Pima Indian Diabetes Dataset

Table 3 (continued)

Alehegn, 2019 [188]	KNN	Techniques used for datasets analysis are Random Forest, KNN, Naïve Bayes, and J48. Ensemble approach facilitates in achieving better results. The accuracy of proposed ensemble approach is 93.62% for PIDD and 88.56% for 130_US hospital dataset.	Pima Indian Diabetes Dataset
Aminah, 2019 [189]	KNN	The results show that the accuracy is 85.6%, false-positive rate (FPR) is 11.07%, false negative rate (FNR) 20.40%, specificity 0.889, and sensitivity 0.796.	16 subjects non-diabetic and 11 subjects diabetic.
Faruque, 2019 [190]	KNN	Experimental results show that C4.5 decision tree achieved higher accuracy compared to other machine learning techniques.	Medical Centre Chittagong (MCC), Bangladesh
Dahiwade, 2019 [191]	KNN	The accuracy of general disease prediction by using CNN is 84.5% which is more than KNN algorithm. And the time and the memory requirement is also more in KNN than CNN.	UCI machine learning Website.
El- Sappagh, 2019 [192]	KNN	The resulting framework achieved 90% of accuracy, 90.2% of recall = 90.2%, and 94.9% of precision.	Mansura University Hospitals (Mansura, Egypt)
Ali, 2020 [193]	KNN	The results show that the KNN types (Fine, Weighted, Medium and Cubic) give high accuracy over the Coarse and the Cosine methods.	American diabetes association.
Garcia-Carretero, 2020 [194]	KNN	The K-nearest neighbor's model accurately classified patients in 96% of cases, with a sensitivity of 99%, specificity of 78%, positive predictive value of 96%, and negative predictive value of 94%.	1647 obese, hypertensive Patients.
Gupta, 2020 [195]	KNN	The best performance of KNN is achieved when no of neighbors (K) is either 33, 40 or 45. The accuracy and error rate is same on these K and it is 87.01% and 12.99 % respectively while a little variation is shown in other metric's values.	Pima Indian Diabetes Dataset
Hassan, 2020 [196]	KNN	The results obtained proved that SVM outperforms decision tree and KNN with highest accuracy of 90.23%.	Pima Indian Diabetes Dataset

Table 3 (continued)

Sarker, 2018 [197]	Opt-KNN	This approach determines the optimal number of neighbors with low error rate for providing better prediction outcome in the resultant model. The effectiveness of this machine learning eHealth model is examined by conducting experiments on the real-world diabetes mellitus data collected from medical hospitals.	Five hundred patients from the relevant medical hospitals.
Bhardwaj, 2021 [198]	KNN	The proposed system achieves an overall accuracy of 98.10% by SVM classifier and 100% by kNN classifier.	MESSIDOR
Mohanty, 2021 [199]	KNN	The data should be 98 % accurate for it to be acceptable in realtime diagnostic tool development. The dataset is required to be trained rigorously to make the analysis more efficient.	Dataset through the means of a google form
Patra, 2021 [200]	SDKNN	The proposed approach achieved classification accuracy of 83.2% which is great improvement when it is compared to other technique.	Pima Indian Diabetes Dataset
Shinde, 2021 [201]	KNN	The Random Forest algorithm give highest accuracy of 74.47% with precision of 80.48%, recall 79.83% and 80.16 F-score.	Pima Indian Diabetes Dataset
Suyanto, 2022 [413]	KNN	It obtains a higher accuracy of 95.24% than the DL-based model for predicting diabetes that gives 94.02%.	Pima Indian Diabetes Dataset
Prasad, 2023 [414]	KNN	The proposed method identifies the best proportion of neighborhoods having a reduced inaccuracy risk to improve the predicting performance of the final system.	Pima Indian Diabetes Dataset

3.3 Un-supervised learning (clustering technique)

In 2010, Paul et al. [235] introduced a technique how to use the background knowledge of medical domain in clustering process to predict the likelihood of diseases. To find the likelihood of diseases, it proposed constraint k-Means-Mode clustering algorithm. The proposed method also gives much better accuracy when compared to the k-means and K-Mode with about 77-78% over k-means and about 82-83% over k-mode. The developed algorithm can handle both continuous and discrete data and perform clustering based on anticipated likelihood attributes with core attributes of disease in data point. We have demonstrated its effectiveness by testing it for a real world patient data set.

In 2011, Hazemi et al. [236] proposed a grid-based interactive diabetes system. In this work agglomerative clustering algorithm is utilized as primary algorithm to focus medical researcher in the findings to predict the implication of the undertaken diabetes patient. This focusing was clearly shown that the grouped (red) line; which represented the optimized view of blood sugar changes over the newly selected period of time; was providing netted view of blood sugar measurements than the measurements (blue) line. GIDS was tested to study a basic history of a diabetes patient who was under supervision for less than a month. The test was performed to check two functions provided by GIDS which are changing the basic algorithm that GIDS used (Chronological Clustering Algorithm) and changing the full view of the supervision period in the time domain in the study.

In 2013, Khanna et al. [234] introduced an integrated approach towards the prediction of the likelihood of diabetes. This paper performs classification on diabetes dataset taken from SGPGI, Lucknow (A super specialty hospital in Lucknow, Uttar Pradesh, India). It predicts an unknown class label for a given set of data and helpful to find out whether the class label for the dataset under consideration would be of low risk, medium risk or high risk. The classifier is further trained on the basis of weights assigned to different attributes which are generated by means of expert guidelines. The accuracy of classifier is verified by kappa statistics and accuracy, evaluation criteria for classifiers.

In 2015, Flynt et al. [243] introduced a model-based clustering approach for the likelihood of diabetic. This work consists of model-based clustering, an unsupervised learning approach, to fid latent clusters of similar US counties based on a set of socioeconomic, demographic, and environmental variables chosen through the process of variable selection. Then use Analysis of Variance and Post-hoc Tukey comparisons to examine differences in rates of obesity and diabetes for the clusters from the resulting clustering solution. The results of the cluster analysis can be used to identify two sets of counties with significantly lower rates of diet-related chronic disease than those observed in the other identified clusters.

In 2017, Bhatia et al. [245] proposed a hybrid based clustering technique in diabetic prediction. In this research work, K-means has been used for removal of the inconsistency found in the data and for optimal feature selection, genetic algorithm is used with SVM for the purpose of classification. K-means is an optimized hierarchical clustering method which aims at reduction of computational cost. The application of the proposed hybrid clustering model applied to a Pima Indians Diabetes dataset shows increase in accuracy by 1.351% and in both sensitivity and positive predicted value by 2.0411%. The proposed model attains better results in comparison to the already existing models in the literature.

In 2018, Ahlqvist et al. [247] presents k-means and hierarchical clustering technique in prediction of diabetes. The clusters were based on six variables (glutamate

Table 4 (Naive Bayes) Diabetes diagnosis summary

Author, Year, References	Algorithms	Experimental Results	Datasets
Sopharak, 2010 [202]	Naive Bayes	It is found that the naive Bayes and SVM classifiers perform better than the NN classifier. The overall best sensitivity, specificity, precision, and accuracy are 92.28%, 98.52%, 53.05%, and 98.41%, respectively.	15 features extracted from each of 115,867 sample.
Tama, 2011 [203]	Naive Bayes	This research has four main outcomes regarding to detect T2DM. First, “boosted” techniques with combining two classifiers do not perform well in order to improve performance. IBk with k=1 and J48 have worst performance than naive Bayes, whereas IBk and J48 have the same performance with accuracy 95,34% and 95,45%, respectively.	Mohammad Hoesin public hospital in Southern Sumatera.
Guo, 2012 [204]	Naive Bayes	Classifier was applied to the modified dataset to construct the Naïve Bayes model. Finally weka was used to do simulation, and the accuracy of the resulting model was 72.3%.	Pima Indian Diabetes Dataset.
Leung, 2013 [205]	Naive Bayes	Models generated by support vector machine (svmRadial) and random forest (cforest) had the best prediction accuracy whereas models derived from naïve Bayes classifier and partial least squares regression had the least optimal performance. Using 10 clinical attributes.	119 subjects with DKD and 554 without DKD at enrolment
Lee, 2013 [206]	Naive Bayes	The values of the area under the receiver operating characteristic curve in the predictions by logistic regression and naive Bayes classifier based on the combination of measures were 0.741 and 0.739 in females, respectively, and were 0.687 and 0.686 in males, respectively.	4870 subjects (2955 females and 1915 males)
Huang, 2015 [207]	Naive Bayes	The proposed method yielded better classification results with the decision tree model with accuracy, specificity, and sensitivity reaching 85.27%, 83.32 and 85.24%, respectively.	185 with diabetic nephropathy and 160 without diabetic
Singh, 2017 [208]	Naive Bayes	It is observed that for the NB machine learning algorithm, the PS test method produces better accuracy compared to other methods without preprocessing method. Moreover, the pre-processing method increases the accuracy for the NB machine learning algorithm.	Pima Indian Diabetes Dataset.
Das, 2018 [209]	Naive Bayes	It is easily conclude that Naive Bayes theorem is better than the J48 as the time to build the model is less.	Collected 200 data by preparing
Insani, 2018 [210]	Naive Bayes	The accuracy rate of this system derived from the scenario distribution data 70 training data and 30 testing data that is equal to 100% according to the doctor’s diagnosis.	RSUD Bendan Kota Pekalongan

Table 4 (continued)

Author, Year, References	Algorithms	Experimental Results	Datasets
Uddin, 2019 [211]	Naive Bayes	It is found that the Support Vector Machine (SVM) algorithm is applied most frequently (in 29 studies) followed by the Naïve Bayes algorithm (in 23 studies). However, the Random Forest (RF) algorithm showed superior accuracy comparatively.	Selected 48 articles in total for the comparison
Birjais, 2019 [212]	Naive Bayes	Gradient Boosting has accuracy of the testing data as 86% and Naive Bayes as 77% and Logistic Regression as 79%.	Pima Indian Diabetes Dataset.
Khan, 2019 [213]	Naive Bayes	The result showed that the accuracy of the proposed algorithm was increased by 10.34% which indicates that the percentage of correct assumptions by the system increases by an amount of 10.34%.	Pima Indian Diabetes Dataset
Sonar, 2019 [214]	Naive Bayes	For Decision Tree, the models give precisions of 85%, for Naive Bayes 77% and 77.3% for Support Vector Machine. Outcomes show a significant accuracy of the methods.	Pima Indian Diabetes Dataset.
Nakra, 2019 [215]	Naive Bayes	Results show the values of classifier output for Naive Bayes classifier. It classifies 79.6935% of the correct instances and the 20.3065% of the incorrect instances.	Datasets available in WEKA tool itself.
Jackins, 2021 [216]	Naive Bayes	For diabetes data, Naive Bayes algorithm gives 76.72 and 74.46 accuracies for training and test data, respectively. Random forest algorithm gives 98.88 and 74.03 for training and test data, respectively.	NIDDK
Priya, 2020 [217]	Naive Bayes	In this proposed system using Naïve Bayes Classifier, Output will be the Web Interface showing the Outcome of having diabetes or not by taking the input values like Insulin level, age and so on.	Dataset consists of several medical predictors.
Rghioui, 2020 [218]	Naive Bayes	The evaluation result showed that the system using the J48 algorithm exhibited excellent classification with the highest accuracy of 99.17%, a sensitivity of 99.47% and a precision of 99.32%.	Glucose sensors, temperature sensors, pedometer.
Khanam, 2021 [415]	Naive Bayes	NN model with a different hidden layer with various epochs and observed the NN with two hidden layers provided 88.6% accuracy.	Pima Indian Diabetes Dataset.
Hasan, 2022 [416]	Naive Bayes	The proposed stacked ensemble model has achieved 93.1% accuracy in predicting blood sugar disease.	PIMA
Okikiola, 2023 [417]	Naive Bayes	Achieved better classification of diabetes with F-measure of 87% compared to other related methods	Pima Indian Diabetes Dataset.

decarboxylase antibodies, age at diagnosis, BMI, HbA1c, and homoeostatic model assessment 2 estimates of β-cell function and insulin resistance), and were related to prospective data from patient records on development of complications and prescription of medication. It identified five replicable clusters of patients with diabetes, which had significantly different patient characteristics and risk of diabetic complications. In particular, individuals in cluster 3 (more resistant to insulin) had significantly higher risk of diabetic kidney disease than individuals in clusters 4 and 5, but had been prescribed similar diabetes treatment. Cluster 2 (insulin deficient) had the highest risk of retinopathy. In support of the clustering, genetic associations in the clusters differed from those seen in traditional type 2 diabetes.

In 2020, Nguyen et al. [252] presents a Binning Approach based on Classical Clustering for Type 2 Diabetes Diagnosis. In this study, we propose a method combining K-means clustering algorithm and unsupervised binning approaches to improve the performance in metagenome-based disease prediction. We illustrate by experiments on metagenomic datasets related to Type 2 Diabetes that the proposed method embedded clusters generated by K-means allows to increase the performance in prediction accuracy reaching approximately or more than 70%. Table 7 represents the related work on diabetes based on clustering technique.

3.4 Un-supervised learning (association rule)

In 2000, Hsu et al. [284] proposed a knowledge discovery system for the diabetic patient database, the interesting issues that have surfaced, as well as the lessons we have learnt from this application. The proposed work uses classification with association rule mining (CBA) technique to find all such patterns. It uses minimum support of 1% and minimum confidence of 50% as suggested by the doctors to mine association rules. Approximately 700 rules are generated in total. The result based on 200,000 screening diabetic records suggested that the proposed exploration, mining methodology aims to give the doctors a better understanding of their data and the discovered patterns by helping the doctors to step through the massive amount of information in stages. Table 8 represents the related work on diabetes based on association rule technique.

4 Review of deep learning technique in diabetes prediction

4.1 Convolutional Neural Network (CNN)

A Convolutional Neural Network (ConvNet/CNN) is a Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other. Table 9 represents the related work on diabetes based on CNN technique.

4.2 Recurrent neural networks (RNN)

RNNs are a powerful and robust type of neural network, and belong to the most promising algorithms in use because it is the only one with an internal memory. Because of their internal memory, RNN's can remember important things about the input they received, which allows them to be very precise in predicting what's coming next. This is why they're

Table 5 (Random Forest) Diabetes diagnosis summary

Author, Year, References	Algorithms	Experimental Results	Datasets
Khalilia, 2011 [219]	Random Forest	Predicted eight disease categories. Overall, the RF ensemble learning method outperformed SVM, bagging and boosting in terms of the area under the receiver operating characteristic (ROC) curve (AUC). In addition, RF has the advantage of computing the importance of each variable in the classification process.	Nationwide Inpatient Sample (NIS)
Casanova, 2014 [220]	Random Forest	Results suggest that RF methods could be a valuable tool to diagnose DR diagnosis and evaluate its progression.	Data from 3443 ACCORD-Eye Study participants.
Sabariah, 2014 [221]	Random Forest	Based on the test results, it shows that the addition of trees and attributes splitter can improve the accuracy and reduce the error rate, with the optimal inputs are 50 numbers of trees and 3 number of attributes splitter with 83,8% average accuracy.	600 training data and 100 test data.
Butwall, 2015 [222]	Random Forest	Random Forest Classifier based approach outperforms better with the accuracy of 99.7%.	Pima Indian Diabetes Dataset.
Xu, 2017 [223]	Random Forest	Provide better prediction accuracy using random forest than using the naive Bayes algorithm, ID3 algorithm and AdaBoost algorithm.	School of medicine, University of Virginia.
Kumar, 2019 [224]	GA-Optimized Random Forest (GA-ORF)	In this evaluation, the various performance metrics of classifiers, GA-ORF has achieved accuracy higher than of the previously proposed classifiers for diabetes mellitus.	University of California at Irvine.
VijiyaKumar, 2019 [225]	Random Forest	The proposed model gives the best results for diabetic prediction and the result showed that the prediction system is capable of predicting the diabetes disease effectively, efficiently and most importantly, instantly.	UCI machine learning repository.
Kaur, 2019 [226]	Random Forest	It has been concluded that the proposed technique outperforms existing machine learning techniques in term s of accuracy, f- measure, sensitivity and specificity by 1.92%, 1.45%, 1.28%, and 1.38%, respectively.	https://archive.ics.uci.edu/ml/index.php
Alam, 2019 [227]	Random Forest	Conducted extensive experiments with other classifiers, such as: Support Vector Machine (SVM), Bayes Network, Multilayer Perceptron, etc. and found the better contribution of Random Forest across all datasets as compared to other classifiers.	UCI Machine Learning Repository

Table 5 (continued)

Author, Year, References	Algorithms	Experimental Results	Datasets
Kaur, 2019 [228]	Random Forest	Maximum accuracy of 97.26% has been achieved on Dermatology dataset using Random Forest machine learning technique.	https://archive.ics.uci.edu/ml/datasets.html.
Benbelkacem, 2019 [229]	Random Forest	The results of the experiments show that our approach based on random forest has proved to be more efficient in comparison with other methods of machine learning.	Pima Indian Diabetes Dataset.
Wang, 2020 [230]	Random Forest	It indicated that RF model was little superior to traditional regression model. RF model could be used in prediction of medical expenditure of diabetics and assessment of its related factors well.	US MEPS data, 2000–2015.
Wang, 2021 [231]	Random Forest	According to the result, Random Forest Classifer combining SVM-SMOTE resampling technology and LASSO feature screening method (Accuracy= 0.890, Precision = 0.869, Recall = 0.919, F1-Score = 0.893, AUC= 0.948) proved the best way to tell those at high risk of DM. Besides, the combined algorithm helps enhance the classifcation performance for prediction of high-risk people of DM.	China National Chronic Disease Survey.
Ooka, 2021 [232]	Random Forest	The RF model showed a higher predictive power for the change in HbA1c than MLR in all models. The RF model including change values showed the highest predictive power.	A total of 168 206 data samples from 64 379 people.
Mondal, 2022 [418]	Random Forest	Demonstrated an accuracy of 85% with the Random Forest (RF) algorithm.	768 data counts with eight features
Gündoğdu, 2023 [419]	Random Forest	The results show that the proposed system achieves an accuracy of 99.2%, an AUC of 99.3%.	Diabetic hospital data in Sylhet, Bangladesh.

the preferred algorithm for sequential data like time series, speech, text, financial data, audio, video, weather and much more. Recurrent neural networks can form a much deeper understanding of a sequence and its context compared to other algorithms. Table 10 represents the related work on diabetes based on RNN technique.

4.3 Artificial Neural Network (ANN)

An Artificial Neural Network is an information processing technique. It works like the way human brain processes information. ANN includes a large number of connected processing units that work together to process information. Table 11 represents the related work on diabetes based on ANN technique.

4.4 Long Short-Term Memory Networks (LSTMs)

Long Short Term Memory is a kind of recurrent neural network. In RNN output from the last step is fed as input in the current step. It tackled the problem of long-term dependencies of RNN in which the RNN cannot predict the word stored in the long term memory, but can give more accurate predictions from the recent information. Table 12 represents the related work on diabetes based on LSTM technique.

4.5 Multilayer Perceptron (MLP)

Multilayer Perceptron (MLP) is a class of feed-forward artificial neural networks. An MLP consists of three main layers of nodes — an input layer, a hidden layer, and an output layer. In the hidden and the output layer, every node is considered as a neuron that uses a nonlinear activation function. MLP uses a supervised learning technique called back propagation for training. When a neural network is initialized, weights are set for each neuron. Back propagation helps in adjusting the weights of the neurons to obtain output closer to the expected. Table 13 represents the related work on diabetes based on MLP technique.

4.6 Autoencoder (AE)

Autoencoder is a type of neural network where the output layer has the same dimensionality as the input layer. In simpler words, the number of output units in the output layer is equal to the number of input units in the input layer. An autoencoder replicates the data from the input to the output in an unsupervised manner and is therefore sometimes referred to as a replicator neural network. Table 14 represents the related work on diabetes based on AE technique.

4.7 Radial Basis Function (RBF)

A Radial Basis Function (RBF) neural network has an input layer, a hidden layer and an output layer. The neurons in the hidden layer contain Gaussian transfer functions whose outputs are inversely proportional to the distance from the center of the neuron. RBF networks are similar to K-Means clustering and PNN/GRNN networks. The main difference is that PNN/GRNN networks have one neuron for each point in the training file, whereas RBF networks have a variable number of neurons that is usually much less than the number

Table 6 (Regression) Diabetes diagnosis summary

Author, Year, References	Algorithms	Experimental Results	Datasets
Takahashi, 2006 [255]	Logistic Regression	The optimum cut-off point of hemoglobin A1C to predict diabetic based on ROC are sensitivity = 84.2% and specificity=92.1%.	2818 people become the subject of this study.
Sparacino, 2007 [256]	Autoregressive (AR)	Results demonstrate that, even by using these simple methods, glucose can be predicted ahead in time, e.g., with a prediction horizon of 30 min crossing of the hypoglycemic threshold can be predicted 20–25 min ahead in time.	28 type 1 diabetic volunteers for 48 h
Eren-Oruklu, 2009 [257]	Recursive Linear	Prediction errors are significantly reduced with recursive identification of the models, and predictions are further improved with inclusion of a parameter change detection method. CG-EGA analysis results in accurate readings of 90% or more.	(sample size n ¼ 22, 43.50 10.4 years old, body mass index [BMI] ¼ 35.02 3.4 kg =m2)
Eren-Oruklu, 2009 [258]	Time-series model	The algorithm has been tested with GPC and LQC methods to provide effective blood glucose regulation in response to multiple meal challenges with simultaneous challenge on subject's insulin sensitivity.	Patient data is obtained using GlucoSim.
Gani, 2008 [259]	Autoregressive (AR)	The continuous measurement of glucose concentration via CGM devices together with data-driven AR models provides a potential, practically useful combination of technologies for accurate near-future prediction of glucose concentrations.	4000 min (i.e., the first 4000 data points) of the available data for each subject.
Estrada, 2010 [260]	Recursive predictive	Normalized least mean square method with adaptive gain (NLMS2) technique achieved the best results as compared to the others.	4 different type 1 diabetic Patients

Table 6 (continued)

Author, Year, References	Algorithms	Experimental Results	Datasets
Gani, 2009 [261]	Autoregressive (AR)	The predictive capability of the models was found not to be affected by diabetes type, subject age, CGM device, and inter individual differences.	4000 min (i.e., the first 4000 data points)
Lu, 2011 [262]	Autoregressive (AR)	Results suggest that the proposed real-time approach can yield ~10-min-ahead predictions with clinically acceptable accuracy and, hence, could be useful as a tool for warning against impending glucose deregulation episodes.	Involving 34 type 1 and 2 diabetic patients
Zhao, 2014 [263]	Autoregressive (AR)	Global AR model based on the frequency separation approach and training data for a single subject could be used to predict future glucose concentrations for other subjects without adjusting model parameters.	Two groups of de-identified ambulatory clinical data..
Georga, 2012 [264]	Support vector regression	The results clearly indicate that the availability of multivariable data and their effective combination can significantly increase the accuracy of both Short-term and long-term predictions.	dataset of 27 patients in Free-living condition.
Bayrak, 2013 [265]	Autoregressive	The early alarm systems based on RARPLS shows good performance. A sensitivity of 86% and a false alarm rate of 0.42 false positive/day are obtained for the early alarm system based on six-step-ahead predicted glucose values with an average early detection time of 25.25 min.	University of Illinois Chicago, College of Nursing, and Iowa State University.

Table 6 (continued)

Author, Year, References	Algorithms	Experimental Results	Datasets
Yu, 2014 [266]	Autoregressive	The results indicate that the prediction accuracy of the rapid modeling method is comparable to that for subject-dependent modeling method for some cases. Also, model migration method presents better generalization ability.	864 samples from each case are used as training data.
Zhao, 2015 [267]	Autoregressive	The prediction accuracy of the rapid modeling method is comparable to that for subject-dependent modeling method for some cases. Also, it can present better generalization ability.	FDA-accepted University of Virginia/University
Paul, 2015 [268]	Linear auto-regressive (AR)	The prediction performances of GARCH approach were compared with other contemporary modelling approaches such as lower and higher order AR, and the state space models. The GARCH approach appears to be successful in both realizing the volatility in glucose profiles and offering potentially more reliable forecasting of upcoming glucose levels.	172 patients participating in a randomized clinical trial.
Bagherzadeh-Khiabani, 2016 [269]	Logistic model	Experiment showed that the variable selection methods used in data mining could improve the performance of clinical prediction models. An R program was developed to make these methods more feasible and visualize the results.	55 variables in 803 females, aged 20 years, followed for 10 e12 years.

Table 6 (continued)

Agarwal, 2016 [270]	Logistic regression	Proposed models for Type 2 diabetes mellitus and myocardial infarction achieve precision and accuracy of 0.90, 0.89, and 0.86, 0.89, respectively. Local implementations of the previously validated rule-based definitions for Type 2 diabetes mellitus and myocardial infarction achieve precision and accuracy of 0.96, 0.92 and 0.84, 0.87, respectively.	Stanford clinical data warehouse SCDW.
Lee, 2016 [271]	Binary Logistic regression	Among all of the variables, the presence of HW was most strongly associated with type 2 diabetes (p < 0.001, adjusted odds ratio (OR) = 2.07 [95% CI, 1.72–2.49] in men; p < 0.001, adjusted OR = 2.09 [1.79–2.45] in women).	11 937 subjects participated in this study.
Rahimloo, 2016 [272]	Logistic regression	Logistic regression, the most influential factor on diabetes predict is diagnosed the hemoglobin A1C that in people who have diabetes is (96%) and those who are on the verge of the disease is (57%).	City of Urmia that dataset contains 180 samples
Rau, 2016 [273]	Logistic regression	The result shows that ANN based model provide better response as compared to LR based model. The sensitivity, specificity and AUC of ANN model are 0.757, 0.755 and 0.837 respectively.	National Health Research Database ,Taiwan
Usman, 2016 [274]	Logistic regression	From the result, all the three predictor variables (age, level of HbA1c and auc-PPG) significantly estimate the risk of having increased arterial stiffness among diabetic patients.	Endocrine Clinic at UKM Medical Centre.

Table 6 (continued)

Zhao, 2016 [275]	Objective-oriented regression (OOR)	The resulted predictive model has an area under curve of 0.92 in the training set, and 0.89 in the validating set indicating that this methodology is useful to build predictive models with complex HLA genotypes.	In total, there are 499 exemplars in the training set.
Bajestani, 2018 [276]	Fuzzy regression	The results of the present work shall prevent unnecessary testing of diabetic patient. This study also aims to assist patients and the healthcare community to reduce the cost of diabetes control and treatment by optimizing the number of check-ups.	Bahman Hospital and parsian clinic of Mashad.
Hassan, 2017 [277]	Logistic Regression	The comparison of prediction models in this study showed that accuracy in predicting in diabetes was significantly higher in the ANN model than in the LR model (p 0.001).	Pima Indian Diabetes Dataset.
Zheng, 2017 [278]	Logistic Regression	LR has the highest accuracy (0.99) at the third level of features with several other models closely following its performance, such as RF and SVM (with 0.98 in accuracy).	300 patient samples
Wu, 2018 [279]	Logistic Regression	The conclusion shows that the model attained a 3.04% higher accuracy of prediction than those of other researchers. Moreover, this model ensures that the dataset quality is sufficient.	Pima Indian Diabetes Dataset.
Qiu, 2019 [280]	Least angle regression (LARS) algorithms	The experimental results show that the algorithm improved the approximation speed for the dependent variables and the accuracy of the regression coefficients.	Pima Indian Diabetes Dataset.

Table 6 (continued)

Yao, 2019 [281]	Multivariable Logistic regression	The area under the receiver operating characteristic (ROC) curve for the BP-ANN model was significantly higher than that by MLR (0.84 vs. 0.77, P < 0.001).	530 Chinese residents
Alshamlan, 2020 [282]	Logistic Regression	Logistic regression produces the highest accuracy with fisher score for GSE38642 dataset with 90.23% and GSE13760 dataset with 61.90%.	GSE38642 and GSE13760
Kopitar, 2020 [283]	Simple Regression	Simple regression model performed with the lowest average RMSE of 0.838, followed by RF (0.842), LightGBM (0.846), and XGBoost (0.881).	27,050 adult individuals.

of training points. For problems with small to medium size training sets, PNN/GRNN networks are usually more accurate than RBF networks, but PNN/GRNN networks are impractical for large training sets. Table 15 represents the related work on diabetes based on RBF technique.

5 Discussion and comparison

In this section, we mainly focus on comparative analysis of several machines and deep learning based diabetic prediction approach , including SVM,KNN, and DT (machine learning) with CNN,RNN and MLP (deep learning) .It also discusses the performance of various machines and deep learning approach for prediction of diabetic disease. Deep learning is computer software that mimics the network of neurons in a brain. It is a subset of machine learning and is called deep learning because it makes use of deep neural networks. The machine uses different layers to learn from the data. The depth of the model is represented by the number of layers in the model. Deep learning is the new state of the art in term of AI.

5.1 SVM vs. CNN

In 2016, Abdillah et al. [45] proposed a machine learning approach using support vector machines with kernel radial basis function (SVM-RBF) to predict diabetes. The Pima Indian diabetes dataset was used to validate the effectiveness of the proposed work. In order to achieve high classification performance, 10-fold cross-validation was used to build a model and search for the optimal parameters. The results of SVM-RBF using 10-fold cross validation was obtained from 500 training data with optimal parameter , which yields accuracy, sensitivity, specificity, and AUROC of 80.22%, 82.56%, 79.12%, and 0.8084 respectively. In the same year Zhu et al. [305] presents a deep learning approach based on CNN to find a patient similarity evaluation framework based on temporal matching of longitudinal patient EHRs (Electronic Health Records). The results of clustering, the deep model with feature embedding is clearly superior to others. On DATASET-I, the deep embedding model achieves an average Rand index of 0.9887, comparing with the second best one with 0.6796. Measured by Purity and NMI, it can achieve the performances of 0.9882 and 0.9516, separately, which also outperforms others with a margin.

In 2018, Dagliati et al. [59] proposed a machine learning approach based on SVM to predict diabetic complications. As far as the choice of the classification method is concerned, AUC values are higher for SVMs and RF when the data sets are balanced. However, SVMs and RF models are harder to interpret, especially considering that our final goal is the model application into clinical practice. In the same year Swapna et al. [310] proposed a deep learning based diabetic prediction model based on CNN approach. The proposed work consists of long short-term memory (LSTM), convolutional neural network (CNN) and its combinations for extracting complex temporal dynamic features of the input HRV data. These features are passed into a support vector machine (SVM) for classification. It has obtained the performance improvement of 0.03% and 0.06% in CNN and CNN-LSTM architecture respectively, compared to earlier work without using SVM. The classification system proposed can help the clinicians to diagnose diabetes using ECG signals with a very high accuracy of 95.7%.

In 2019, Alirezaei et al. [66] present a machine learning approach based on SVM to predict diabetic complications. In this work, a method based on the k-means clustering algorithm is first utilized to detect and delete outliers. Then in order to select significant and effective features, four bi-objective meta-heuristic algorithms are employed to choose the least number of significant features with the highest classification accuracy using support vector machines (SVM). The results, based on PIMA Indian Type-2 diabetes dataset concluded that the multi-objective firefly (MOFA) and multi-objective imperialist competitive algorithm (MOICA) with 100% classification accuracy, outperform the non-dominated sorting genetic algorithm (NSGA-II) and multi-objective particle swarm optimization (MOPSO) with the accuracies of 98.2% and 94.6%, respectively.Sun et al. [313] proposed a Neural Network Method (CNN) based deep learning approach to build a diagnostic model, in this work, the CNN model is combined with the BN layer to prevent the dispersion of the gradient, speed up the training speed and improve the accuracy of the model. The experiments show that this method can achieve a training accuracy of 99.85% and a testing accuracy of 97.56%, which is more than 2% higher than that of using logistic regression.

In 2020, Harimoorthy et al. [79] proposed a multi disease prediction model using machine learning approach based on improved SVM-Radial bias technique. In this work, a general architecture has proposed for predicting the disease in the healthcare industry. This system was experimented using with reduced set features of Chronic Kidney Disease, Diabetes and Heart Disease dataset using improved SVM-Radial bias kernel method, and also this system has compared to other machine learning techniques such as SVM-Linear, SVM-Polynomial, Random forest and Decision tree in R studio. The performance of all these machine learning algorithms has evaluated with accuracy, misclassification rate, precision, sensitivity and specificity. From the experiment results, improved SVM-Radial bias kernel technique produces accuracy as 98.3%, 98.7% and 89.9% in Chronic Kidney Disease, Diabetes and Heart Disease dataset respectively. Ismail et al. [316] proposed a Remote health monitoring application with the advent of Internet of Things (IoT) technologies. The proposed work uses a deep learning approach based on CNN. In this framework, develop a CNN-regular pattern discovery model for data classification. First, the most important health-related factors are selected in the first hidden layer, then in the second layer, a correlation coefficient analysis is conducted to classify the positively and negatively correlated health factors. Moreover, regular patterns' behaviours are discovered through mining the regular pattern occurrence among the classified health factors. The accuracy of diagnosis and referral of our model reached 80.43%; 80.85%; 91.49%; 82.61%; 95.60% with a testing dataset, respectively.

5.2 KNN vs. RNN

In 2018, Dwivedi et al. [385] introduced a computational intelligence technique for diabetes mellitus prediction. The proposed model uses a machine learning approach based on KNN (K-Nearest Neighbor). Clearly indicates that the ANN and the logistic model predicts the highest number of true positive (430 and 443 out of 500, respectively), where naïve Bayes predicts the highest number of true negative (179 out of 268). Naïve Bayes also predicts lowest number of false positive (89) whereas logistic regression predicts the lowest number of false negative. Overall, naïve Bayes has lowest type I error and logistic regression has lowest type II. Chen et al. [330] proposed a new deep learning technique, which is based on the Dilated Recurrent Neural Network (DRNN)

Table 7 (Clustering) Diabetes diagnosis summary

Author, Year, References	Algorithms	Experimental Results	Datasets
Padmaja, 2008 [233]	K-mean	The results were evaluated in five different clusters and they show that 23% of the women suffering from diabetes fall in cluster-0, 5% fall in cluster-1, 23% fall in cluster-2, 8% in cluster-3 and 25% in cluster-3.	National Institute of Diabetes
Paul, 2010 [235]	Constraint k-Means-Mode	The proposed method also gives much better accuracy when compared to the k-means and K-Mode with about 77-78% over k-means and about 82-83% over k-mode.	50273 instances with 514 attribute.
Hazemi, 2011 [236]	Chronological clustering	This focusing was clearly shown that the grouped (red) line; which represented the optimized view of blood sugar changes over the newly selected period of time; was providing netted view of blood sugar measurements than the measurements (blue) line.	27 data from 5th to 17th of January, 2010.
Khanna, 2013 [234]	K-mean	Three clusters has been identified which further represent three different class levels for the diabetic patients. Here it is also explored that different attributes have different impact in classification process and playing a crucial role for determining the performance of the classifier.	SGPGI, Lucknow
Antonelli, 2013 [237]	Multiple-level clustering	The experimental validation, performed on a real collection of diabetic patients, demonstrates the effectiveness of the approach in identifying groups of patients with a similar examination history and increasing severity in diabetes complications.	Local Health Center (LHC) of the Asti province in Italy.
Al-Hazemi, 2014 [238]	Agglomerative clustering	GIDS is enabling a broader view of diabetes patient's condition and predict the future implication based on current developments in some diabetes factors such as blood sugar, fat, or potassium.	Less than a month data set.

Table 7 (continued)

Author, Year, References	Algorithms	Experimental Results	Datasets
Jeong, 2014 [239]	Chronological clustering	Evaluation results showed that the proposed model can be used to analyze the temporal change of MS status. Using the proposed model, it can effectively manage the patients having metabolic syndrome using the control range of patient status.	Patients data over last 10 years
Kim, 2014 [240]	Hierarchical	Demonstrated that our clustering method retains the benefits of existing diabetes risk models and adds its own advantage through allowing for fine control of detail that is presented to the user. This promises great potential of contributing to clinical practice.	Consists of 52,139 patients.
Vijayarani, 2014 [241]	Hierarchical and partitioning	Through examining the experimental results, it is observed that the CURE with K-Means clustering algorithm performance is more accurate than the BIRCH with K-Means algorithm.	Pima Indian Diabetes Dataset.
Sanakal, 2014 [242]	Fuzzy C-means clustering	FCM clustering applied on Diabetes dataset yields relatively better classification result of 94.3% accuracy.	Pima Indian Diabetes Dataset.
Flynt, 2015 [243]	Model-based clustering	It is found that clusters containing smaller metropolitan to non- metropolitan counties (clusters 1 and 2) have significantly higher diabetes rates than cluster 5 (large and large fringe). Additionally, cluster 2 has a higher diabetes rate than cluster 4.	US states of Pennsylvania (PA) and New York (NY).
Barale, 2016 [244]	K-mean	K-means algorithm combined with artificial neural network classifier and k-means algorithm combined with logistic regression classifier achieve classification accuracy above 98%.	Pima Indian Diabetes Dataset.

Table 7 (continued)

Author, Year, References	Algorithms	Experimental Results	Datasets
Bhatia, 2017 [245]	Hybrid-Clustering	The application of the proposed hybrid clustering model applied on Pima Indians Diabetes dataset shows increase in accuracy by 1.351% and in both sensitivity and positive predicted value by 2.0411%.	Pima Indian Diabetes Dataset
Cheruku, 2017 [246]	K-mean	It is also proved that proposed model performs better compared to familiar classifiers namely probabilistic neural network, feed forward neural network, cascade forward network, time delay network, artificial immune system and GINI classifier.	Pima Indian Diabetes Dataset.
Ahlqvist, 2018 [247]	K-means and hierarchical clustering	Individuals in cluster 3 (most resistant to insulin) had significantly higher risk of diabetic kidney disease than individuals in clusters 4 and 5, but had been prescribed similar diabetes treatment. Cluster 2 (insulin deficient) had the highest risk of retinopathy.	All New Diabetics in Scania (ANDIS).
Rani, 2018 [248]	Association clustering	Time series based prediction has the difference of prediction from 1 to 5%. But ANN based prediction has different of 19%to 22%.	http://archive.ics.uci.edu/ml/machine-learningdatabases/00296/.
Derevitskii, 2019 [249]	K-mean	The proposed method proved to be a useful tool for analyzing diabetes trajectories. However, it should take into consideration the specifics of both chronic diseases in general and specific particular diseases.	7000 diabetic records.
Lasek, 2019 [250]	DBSCAN, K-mean	Proposed a method for visual analysis of clustering results using our developed visualization application which main feature is that it is capable of visualizing results of clustering of high-dimensional datasets so that the graphical representation is not disturbed with additional information.	65,000 of records from a 2010 Canadian Community Health survey

Table 7 (continued)

Raihan, 2019 [251]	K-mean and Hierarchical	It is clear that clustering techniques are impressively useful to predict diabetes and as diabetes is becoming a major problem nowadays it is important to find out proper solution to get rid of diabetes on an early stage and from that perspective.	464 instances and 23 features
Nguyen, 2020 [252]	K-mean with Binning Approach.	Type 2 Diabetes that the proposed method embedded clusters generated by K-means allows to increase the performance in prediction accuracy reaching approximately or more than 70%.	Includes 344 Chinese individuals and 96 western women
Syafaah, 2020 [253]	K-mean	The results obtained are of high quality and the tool can be used correctly. Whereas in terms of quantity, the urinary colour detection error is about 5%.	10 sample with diabetic patients, and 10 samples with a normal patients.
Anwar, 2021 [254]	Hierarchical	In the first part, the risk factors identified among non-diabetic participants showed a significant association with the development of diabetes mellitus, particularly physical inactivity (49.12%), hypertension (41.15%), and high body mass index (19.03%). Likewise, in 11.54% of diabetic patients, elevated body mass index (30.51%), hypertension (27.12%) and physical inactivity (55.93%), this could be associated with diabetic complications.	Five hundred and eleven participants took part in the study
Hassan, 2022 [420]	Elbow, Silhouette, and K-means	On the cluster-based dataset and the complete dataset, the maximum Accuracy (ACC) is 99.57% and 99.03%, respectively.	520 diabetic patients.
Alghamdi, 2023 [421]	K-Mean	The XGBoost classifier showed the most effectiveness, with an accuracy rate of 89%.	PIMA

model, is proposed to predict the future glucose levels for prediction horizon (PH) of 30 minutes. The result reveals that using the dilated connection in the RNN network, it can improve the accuracy of short-time glucose predictions significantly (RMSE = 19.04 in the blood glucose level prediction (BGLP) on and only on all data points provided). Lastly, in order to improve the performance of DRNN model, the first-order linear interpolation and first-order extrapolation are applied to the training and testing set, respectively.

In 2019, Alehegn et al. [188] introduced a MLTs (Machine Learning Techniques) that can act as a savior for early diagnosis and prediction of DM. ML is another side of Artificial Intelligence so that be used for prediction, recommendation and recovery from disease in early stages. The system proposed in this work makes use of two datasets viz. PIDD (Pima Indian Diabetes Dataset) and 130_US hospital diabetes data sets. Techniques used for datasets analysis are Random Forest, KNN, Naïve Bayes, and J48. The ensemble approach facilitates in achieving better results. The accuracy of proposed ensemble approach is 93.62% for PIDD and 88.56% for the 130_US hospital dataset. It is also observed that when dataset becomes large the accuracy of the proposed algorithm is not good relatively. NB and J48 prediction algorithm are better for large datasets analysis. KNN technique is not good for large dataset analysis. Li et al. [406] present a deep learning model that is capable of forecasting glucose levels with leading accuracy for simulated patient cases. This work is based on multi-layer convolutional recurrent neural network (CRNN) architecture. The proposed CRNN method showed superior performance in forecasting BG levels (RMSE and MARD) in the in silico and clinical experiments. The results achieved a mean RMSE = 9.38mg/dL in silico using the proposed method, and it is the best amongst other algorithms, including SVR, LVX and 3rd order ARX.

In 2018, Sarker et al. [197] proposed an optimal K-Nearest Neighbor (KNN) Learning based Diabetes Mellitus Prediction and analysis for eHealth Services. In this model select 5 baseline classification methods, such as Adaptive Boosting (AdaBoost), Logistic Regression (LR), Naive Bayes (NB), Support Vector Machines (SVM), and Decision Tree (DT) that are frequently used to analyze health data. The results show that proposed Opt-KNN based disease prediction model outperforms the traditional KNN based model. Opt-KNN based model gives better prediction accuracy in terms of precision, recall, f-measure, ROC area. This results show that Opt-KNN is more effective that traditional KNN in terms of prediction accuracy and minimize the additional effort for assuming the K-value. Zhou et al. [326] proposed a deep learning approach using RNN that can help not only to predict the occurrence of diabetes in the future, but also to determine the type of the disease that a person experiences. The experimental results show the effectiveness and adequacy of the proposed DTP model. The best result for the diabetes type dataset was 94.021 74% and that for the Pima Indians diabetes dataset was 99.411 2%. The experiments proved that proposed model can perform well on different types of data. The proposed model not only can predict if a person will be diabetic in the future, but also can determine and predict the specific type of the disease, type 1 or type 2.

In 2021, Patra et al. [200] introduced a machine learning based diabetic prediction model using standard deviation K-Nearest Neighbor (SDKNN). The proposed technique is based on a new distance calculation formula to find the nearest neighbor in KNN. The work consists of two segments, in the first segment standard deviation of attributes is used as power for calculating K-nearest neighbor to improved classification accuracy and in the second segment, based on mean of standard deviation attributes ,distance in KNN is processed to further improve the classification accuracy. This concept is applied on Pima Indian Diabetes Dataset (PIDD). The analysis is carried out on data set by splitting 90% of

Table 8 (Association-Rule) Diabetes diagnosis summary

Author, Year, References	Algorithms	Experimental Results	Datasets
Hsu, 2000 [284]	Classification with association rule mining (CBA)	Proposed exploration mining methodology aims to give the doctors a better understanding of their data and the discovered patterns by helping the doctors to step through the massive amount of information in stages.	200,000 screening records.
Stilou, 2001 [285]	Apriori algorithm	Because of the small confidence (50%) the rules that are generated are not so strong. The support parameter is also small (10%), because of the small number of categories that were generated. The specific data set creates 2 or at most 3 categories for each field.	127,886 beneficiaries Participated.
Zorman, 2002 [286]	Association Rule	For comparison with decision trees, only those rules were selected, where the consequence attribute was from the set of preselected 5 attributes. The rule size was limited only to rules with less than five attributes.	Osaka Medical College Hospital.
Duru, 2005 [287]	Apriori algorithm	The use of this software made possible to generation of two, three and even four item sets. It concluded that, developed software and the methodology have served the purpose and worked well.	Faculty of medicine of Kocaeli University
Mao, 2009 [288]	Apriori-Gen	The result shows the interaction among 5 SNPs with support s of 50% and confidence α of 60%. The risk rate RR and odds ratio OR are 2.14 and 2.92, respectively.	19 tag SNPs out of 92 for the type 2 diabetes data set.
Patil, 2010 [289]	Apriori algorithm	Generated the association rules which are useful to identify general associations in the data, to understand the relationship between the measured fields whether the patient goes on to develop diabetes or not.	Pima Indian Diabetes Dataset.

Table 8 (continued)

Author, Year, References	Algorithms	Experimental Results	Datasets
Patil, 2011 [290]	Predictive Apriori algorithms	Generate the rules by using both the algorithm standard Apriori and predictive Apriori. Predictive Apriori is able to mine a high quality set of association rules. However, experiments it shows that the time complexity of predictive Apriori is worse. Predictive Apriori can improve classification using association rules when it is used to generate a small set of rules.	Pima Indian Diabetes Dataset.
Kasemthaweesab, 2012 [291]	Apriori algorithm	The primary objective is to provide a useful medical and healthcare information that can be applied for a treatment of elderly-adult patients, an improvement of healthcare service including a provision of practical instruction for diabetes patients without complications.	65,535 raw data samples were normalized
Kim,2012 [292]	Association rule mining (ARM)	Six association rules were found among three comorbid diseases. Among them, essential hypertension was an important node between T2DM and stroke (support, 4.06%; confience, 8.12%) as well as between T2DM and dyslipidemia (support, 3.44%; confience, 6.88%).	Keimyung University Medical Center.
Simon, 2013 [293]	Survival Association Rule (SAR)	The proposed approach identified clinically relevant association rules, and estimated the risk associated with the risk factors in the rules more correctly than traditional association rules in a manner that makes interpretation even simpler.	Mayo Clinic
Schrom, 2013 [294]	Association rule mining (ARM)	This method found that statins statistically significantly increase the risk of diabetes between 13% and 41% among various phenotypes. However, this method of calculating relative risk assumes that the treatment and control groups are generally similar.	Data set consists of 18,958 patients

Table 8 (continued)

Author, Year, References	Algorithms	Experimental Results	Datasets
Lakshmi, 2014 [295]	Apriori algorithm	A comparative study for the number of rules generated for different values of support and confidence is also conducted. The number of rules generated decreases as the value of confidence is increased.	Consists of more than 1000 medical transcripts.
Karthikeyan, 2015 [296]	Association rule mining (ARM)	The experimental observations reveal that this framework provides a better accuracy of 95% when evaluated against the existing techniques.	Pima Indians Diabetes Data Set
Ramezankhani, 2015 [297]	Association rule mining (ARM)	Proposed study showed that ARM is a useful approach in determining which combinations of variables or predictors occur together frequently, in people who will develop diabetes.	A total of 6647 persons participated.
Simon, 2013 [298]	Association rule mining (ARM)	Used distributional association rule mining to identify sets of risk factors and the corresponding patient subpopulations who are at significantly increased risk of progressing to diabetes.	Mayo Clinic patient data.
Kamalesh, 2016 [299]	Association rule mining (ARM)	Four summarization techniques (APRX-Collection, RPGlobal, Top-K, and BUS) were analyzed and found that bottom-up summarization (BUS) produced optimal results.	Pima Indian Diabetes Dataset.
Alam, 2019 [300]	Apriori method	Using association rule mining, the results have shown that there is a strong association of BMI and glucose with diabetes. The association of blood glucose, blood pressure, age, and BMI with diabetes also depended on socio economic, geographic, and clinical factors.	National Institute of Diabetes
Lu, 2021 [301]	Apriori algorithm	Results showed that the most associated rules were {BL23, BL18} ≥ {SP6}, {BL20, BL18} ≥ {PC6}, {PC6, BL18} ≥ {BL20}, and {SP6, BL18} ≥ {BL23} in the database.	Taipei Tzu Chi Hospital
Khafaga, 2022 [422]	Apriori algorithm	KNN provided the highest accuracy of 97.36% compared to the other applied algorithms.	University of California

Table 9 (CNN based) Diabetes diagnosis summary

Author, Year, References	Algorithms	Experimental Results	Datasets
Cheng, 2016 [302]	CNN	Different temporal fusion mechanisms are also investigated to explore temporal smoothness of patient EHRs in the proposed framework. Diabetes and hypertensions are also observed because those chronic conditions usually co-exist with each other.	Records of 319,650 patients over 4 years.
Pratee, 2016 [303]	CNN	The results demonstrate impressive results, particularly for a high-level classification task. On the data set of 80,000 images used our proposed CNN achieves a sensitivity of 95% and an accuracy of 75% on 5,000 validation images.	Kaggle dataset
Shi, 2016 [304]	CNN	The experiment has been conducted with data of patients with cerebral infarction, patients with pulmonary infection and patients with coronary atherosclerotic heart disease obtained from a second grade.	31,919 patients, and 20320848 recorders
Zhu, 2016 [305]	CNN	The experimental results show that proposed model achieves significantly better representations over the baselines, which enables more accurate patient cohort discovery.	EHR database of 218,680 patients
Lekha, 2017 [306]	Modified CNN	It has been found that the algorithm substantially reduces the mean square errors and optimizes the overall performance of the classifier.	25 signals containing a type 1 diabetic Samples.
Mohebbi, 2017 [307]	CNN	It contrasts a standard logistic regression baseline to Multilayer Perceptrons (MLPs) and Convolutional Neural Networks (CNNs). The best classification performance with an average accuracy of 77 .5% was achieved with CNN.	10800 days of CGM total of 97200 CGM days.
Kwasigroch, 2018 [308]	CNN	The best tested model achieved an accuracy of about 82% in detecting the retinopathy and 51% in assessing its stage. Moreover, system obtained decent Kappa score equal 0.776.	88000 retina photographs
Swapna, 2018 [309]	CNN	The maximum accuracy obtained for test data is 90.9% using CNN-LSTM. Using 5 fold cross-validations, CNN gave an accuracy of 93.6% while CNN-LSTM combination gave the maximum accuracy of 95.1%.	(ECG) of 20 diabetes patients
Swapna, 2018 [310]	CNN	Obtained the performance improvement of 0.03% and 0.06% in CNN and CNNLSTM architecture respectively compared to our earlier work without using SVM.	Electrocardiograms (ECG) of 20 people.
Butt, 2019 [311]	CNN	The proposed system is tested on a DR Dataset consisting of 35,126 images provided by EyePACS. Experimental results indicate that the accuracy of 97.08% is achieved through the model that outperforms those achieved through other methods in recent studies.	Dataset consisting of 35,126 images.

Table 9 (continued)

Khan, 2019 [312]	CNN	The proposed method uses pre-trained CNN models i.e. AlexNet, VGG-16 and SqueezeNet, which gave the classification accuracy of 93.46%, 1.82% and 94.49% respectively.	The MESSIDOR dataset.
Sun, 2019 [313]	CNN	The experiments show that this method can achieve a training accuracy of 99.85% and a testing accuracy of 97.56%, which is more than 2% higher than that of using logistic regression.	Information of 301 hospitalized patients
Raj, 2020 [314]	CNN	For this research work uses Diabetic Retinopathy dataset provided by Kaggle Community. Finally, CNN to predict the Diabetic Retinopathy (DR). Proposed methodology, achieved 95.41% accuracy.	88,702 sets of High-resolution.
Rahman, 2020 [315]	CNN	The Conv-LSTM-based model classified the diabetes patients with the highest accuracy of 91.38 %. In later, using cross-validation technique the Conv-LSTM model achieved the highest accuracy of 97.26 % and outperformed the other three models along with the state-of-the-art models.	Pima Indians Diabetes Database.
Ismail, 2020 [316]	CNN	The RMSE of the general CNN model is about 0.87 with 1.1 complexities. However, the maximum RMSE of the proposed model prediction is 0.2562 for the presence of diabetes.	Korea National Health survey
Islam, 2021 [317]	CNN	Using a relatively small-sized dataset, it develop a multi-stage convolutional neural network (CNN)-based model DiaNet that can reach an accuracy level of over 84% on this task, and in doing so, successfully identifies the regions on the retina images that contribute to its decision-making process, as corroborated by the medical experts in the field.	Retinal images from a diabetes cohort of size 246 and a control group of size 246.
Madan, 2022 [423]	CNN	CNN-Bi-LSTM surpasses the other deep learning methods in terms of accuracy (98%), sensitivity (97%), and specificity (98%).	PIMA Indian dataset
Aslan, 2023 [424]	CNN	The accuracy of the classification using the SVM/cubic model with 500 selected features was 92.19%.	PIMA Indian dataset

training data and 10% of testing data. The proposed approach achieved an accuracy rate of 83.2%, which shows better improvement as compared to the other technique.

Rabby et al. [328] presents a novel approach to predicting the blood glucose level with a stacked long short term memory (LSTM) based deep recurrent neural network (RNN) model considering sensor fault. In this work Kalman smoothing technique is used for the correction of the inaccurate CGM readings due to sensor error. Results demonstrate that the RNN model with stacked LSTM layered architecture performs better than RNN with a single LSTM layer for all of the cases based on OhiT1DM dataset. The proposed approach is more generalized as the prediction RMSE for all six patients is uniformly improved. As a consequence, it does not further experiment with traditional machine learning approaches. Proposed approach provides more reliable predictions than traditional methods while it assumed fngerstick BG readings as the ground truth in our experiment.

5.3 DT vs. MLP

In 2018, Zou et al. [155] proposed a diabetes mellitus prediction model using different machine learning technique. In this study, it uses decision tree, random forest and neural network to predict diabetes mellitus. The dataset is the hospital physical examination data in Luzhou, China. It contains 14 attributes. In this study, five-fold cross validation was used to examine the models. Principal component analysis (PCA) and minimum redundancy maximum relevance (mRMR) to reduce the dimensionality. The results showed that prediction with random forest could reach the highest accuracy (ACC = 0.8084) when all the attributes were used. In the Luzhou dataset, J48 (Decision Tree) has the best performance. But the results are not better than using all features. In the Pima Indians dataset, this method, which used RF as the classifier, has the best performance. Alfian et al. [371] developed a personalized healthcare monitoring system for diabetic patients by utilizing deep learning approach based on multi-layer perceptron (MLP). The results show that MLP achieved the highest accuracy (77.083%) compared to 73.046%, 76.6927%, 76.562%, and 76.0417% for Random Forest, Naïve Bayes, SVM, and Logistic Regression, respectively. These results show that for a small number of features (2 h glucose tolerance, diastolic blood pressure, body mass index, and age), the MLP algorithm achieved the highest accuracy of prediction compared to other models.

In 2019, Hebbar et al. [161] introduced a decision tree (DT) based prediction model for diabetic patients. Decision tree and random forest algorithms are applied on data to learn the class model. The optimal model is selected. Then the chosen model is promoted for testing with the test set of data. DRAP makes the classification and prediction based on the feature set mainly consisting of BMI, age, blood pressure, insulin level, and glucose level. The model used to modify decision tree, and random forest algorithm for learning, classification and prediction. Experimental study of the real-life data set has shown promising results and DRAP - yield the accuracy of 72% and 75% for decision tree, and random forest respectively. Mohapatra et al. [372] introduced a deep learning based diabetes prediction model using multi-layer perceptron (MLP). In this work, MLP is used for classification of pregnant women. Proposed technique has been applied on the diabetes database of PIMA for Indian people and is collected from University of California (UCI). Total 1768 lady patients are considered for the experiment It is found that 268 are suffering from diabetes and rest 500 cases are in healthy. Also, it is verified for missing data and found no missing is there. The experiment results achieved 77.5% of classification accuracy, using the proposed approach.

Table 10 (RNN based) Diabetes diagnosis summary

Author, Year, References	Algorithms	Experimental Results	Datasets
Allam, 2011 [318]	RNN	The results of the proposed technique is evaluated and compared relative to that obtained from a feed forward neural network prediction model (NNM). Results indicate that, the RNN is better in prediction than the NNM for the relatively long prediction horizons.	Data set consists of 4916 samples.
Chu, 2018 [319]	RNN	The best performance among the state-of the-art proposals and reduces workload of manual annotation significantly with fewer omissions in detecting the three types of AMEs.	8845 medical records of patients
Wang, 2018 [320]	RNN	The experimental results showed that the recurrent model can get a better effect than other non-sequential methods in possibility and diagnosis prediction of clinical visits.	January 1, 2011 and July 29, 2015, total 55 months.
Wu, 2018 [321]	RNN	It is found that the inclusion of relative timing can meaningfully improve performance, especially when concatenating one history, element of relative time; but, it is also a noisy signal, and further work needs to be done to engineer systems using this paradigm.	Olmsted County Birth Cohort
Martinsson, 2018 [329]	RNN	Present an approach for predicting blood glucose levels for diabetics up to one hour into the future. The approach is based on recurrent neural networks trained in an end-to-end fashion, requiring nothing but the glucose level history for the patient.	Ohio T1DM Dataset for Blood Glucose Level.
Chen, 2018 [330]	RNN	The result reveals that using the dilated connection in the RNN network, it can improve the accuracy of short-time glucose predictions significantly (RMSE = 19.04 in the blood glucose level prediction (BGLP) on and only on all data points provided).	Dataset of 575 subjects.
Dong, 2019 [322]	RNN	We firstly designed a prediction model based on GRU and then proposed pre-training and fine-tune processes incorporating into the GRU based model. Numerical results suggest that the proposed approach outperforms existing SVR and Md3RNN methods on both 30 min and 45 min BG prediction tasks.	Collected historical type 1 BG data set of 40 patients.

Table 10 (continued)

Author, Year, References	Algorithms	Experimental Results	Datasets
Dong, 2019 [323]	RNN	Numerical results suggest that the proposed Clu-RNN approach utilizes more than one cluster for both type I and type II diabetes and has gained improvements compared with support vector regression (SVR) and other RNN methods in terms of BG prediction accuracy.	80 diabetic patients.
Jang, 2019 [324]	RNN	To verify the results of the prediction model, it compared the accuracy with the existing machine learning methods, LR, k-NN, and SVM. Proposed rediction model accuracy was 0.92 and the AUC was 0.92, which were higher than the other. Therefore predicting the onset of T2DM by using the proposed diabetes prediction model in this study, it could lead to healthier lifestyle and hyperglycemic control resulting in lower risk of diabetes by alerted diabetes occurrence.	Korean Genome and Epidemiology study (Ansan, Anseong Korea).
Li, 2019 [406]	CRNN	Achieved a mean RMSE = 9.38mg/dL in silico using the proposed method, and it is the best amongst other algorithms, including SVR, LVX and 3rd order ARX.	T1DM subjects in a 6 month clinical trial
Munoz-Organero, 2020 [325]	RNN	The differential equations for carbohydrate and insulin absorption in physiological models are modeled using a Recurrent Neural Network (RNN) implemented using Long Short-Term Memory (LSTM) cells. The results show Root Mean Square Error (RMSE) values under 5 mg/dL for simulated patients and under 10 mg/dL for real patients.	D1NAMO dataset.
Zhou, 2020 [326]	RNN	The experimental results show the effectiveness and adequacy of the proposed DTP model. The best result for the diabetes type dataset was 94.021 74% and that for the Pima Indians diabetes dataset was 99.411 2%.	PIMA indian dataset.
Zhu, 2020 [327]	RNN	With a properly trained model, it conducted an evaluation with a new set of data and compared its performance with the results of the NNPG, SVR and ARX methods. Our results show that the DRNN model achieves the best performance with the smallest RMSE, MARD and time lag.	UVA/Padova T1D simulator to generate 10 virtual T1DM subjects.

Table 10 (continued)

Author, Year, References	Algorithms	Experimental Results	Datasets
Rabby, 2021 [328]	RNN	Achieved an average RMSE of 6.45 and 17.24 mg/dl for 30 min and 60 min of prediction horizon (PH), respectively.	OhioT1DM (2018) dataset
Srinivasu, 2022 [425]	RNN	Proposed research showed that the suggested model could be used in real-world scenarios	PIMA
Kiruthiga, 2023 [426]	RNN	This study proposes an approach using the deep spectral recurrent neural network (DSRNN) algorithm.	PIMA

Table 11 (ANN based) Diabetes diagnosis summary

Author, Year, References	Algorithms	Experimental Results	Datasets
Jaafar, 2005 [331]	ANN	The network with eight inputs and four inputs are then tested and results obtained are compared in terms of error. The outcome of this study is whether someone is the diabetes sufferer or not. Accurate results have been obtained which proves the effectiveness of the proposed ANN.	768 cases were selected for analysis.
Mougiakakou, 2006 [332]	ANN	Results obtained from the FFNN and the RNN trained with the on-line RTRL-TF are superior to those obtained by the on-line RTRL-FR trained RNN for all diabetes patients.	4 children with Type 1 diabetes
Dey, 2008 [333]	ANN	The results presented for the diabetes classification problem validates the fact that, the network is able to classify diabetic and non-diabetic patients with the network performance of 92.5%.	Sikkim Manipal Institute
Pappada, 2008 [334]	ANN	Overall, the neural network models perform adequately at predicting at normal (>70 and <180 mg/dl) and hyperglycemic ranges (≥180 mg/dl); however, glucose concentrations in areas of hypoglycemia were commonly Overestimated.	Private endocrine practice in Warren, OH.
Zainuddin, 2009 [335]	ANN	Comparisons of the diagnostic accuracy with other neural network models, which use the same dataset are made. The comparison results showed overall improved accuracy, which indicates the effectiveness of this proposed system.	One patient covering a period of 77 days.
Perez-Gandia, 2010 [336]	ANN	A comparison with a previously published technique, based on an autoregressive model (ARM), has been performed. The comparison shows that the proposed NNM is more accurate than the ARM, with no significant deterioration in the prediction delay.	Guardian Real-Time CGM System
Pappada, 2010 [337]	ANN	Proposed approach reveals that the patient specific model generated prediction with a high degree of accuracy of 95.1%.	2,923 data points of CGM
Chakraborty, 2010 [338]	ANN	The average classifiation rate using BPMLP neural network architecture is found to be 78.20% with the standard deviation of 2.57% whereas PNN neural network architecture gives 85.6% as the average classifiation rate with the standard deviation of 2.06%.	Total 500 data set of the patients
Allam, 2011 [339]	ANN	The results of the proposed research indicate that the NNM can be used to accurately predict future glucose values for prediction horizons of 30 minutes or less without time delay between the predicted output and the real glucose samples.	Medtronic [Northridge, CA]
Pappada, 2011 [340]	ANN	The model predicts 88.6% of normal glucose concentrations (>70 and <180 mg/dL), 72.6% of hyperglycemia (>=180 mg/dL), and 2.1% of hypoglycemia (<=70 mg/dL).	Private endocrine Warren, OH.

Table 11 (continued)

Author, Year, References	Algorithms	Experimental Results	Datasets
Robertson, 2011 [341]	ANN	This is probably due to the reduced effect of meals and insulin injections during this period, although due to time constraints, a more rigorous investigation of this hypothesis could not be conducted.	diabetes simulator, AIDA.
Ali, 2018 [342]	ANN	Experimental results show that the proposed ANN is accurate, adaptive, and very encouraging for a clinical implementation.	CGM data of 13 patients
Kathiroli, 2018 [343]	ANN	Proposed algorithm has high accuracy when compared with other algorithm except decision table. The accuracy of DT is 0.1% greater than the proposed algorithm but the construction of table is time consuming and the small change in input dataset can largely vary the outcome.	Pima Indians Diabetes Dataset
Senturk, 2020 [344]	ANN	ANN based detection approach gave the best results. 88.52% sensitivity has been obtained using the features of Messidor dataset.	Messidor dataset.
Bukhari, 2021 [427]	ANN	ABP-SCGNN model, containing 20 neurons, attains 93% accuracy on the validation set.	PIMA
Karthik, 2022 [428]	ANN	Proposed method achieves a higher degree of accuracy with 99% than other existing methods.	PIMA
Al Sadi, 2023 [429]	ANN	The proposed model shows significant performance in selecting the most accurate predictors of diabetes.	Oman Dataset

Table 12 (LSTMs) Diabetes diagnosis summary

Author, Year, References	Algorithms	Experimental Results	Datasets
Sun, 2018 [345]	LSTM	The LSTM network outperformed the baseline methods ARIMA and SVR. In comparison with the baselines and in all prediction horizons, the LSTM network reduced the RMSE and TL, while CC and Fit were increased.	26 datasets from 20 real patients.
Farias, 2018 [346]	LSTM	The best results were obtained with a bidirectional LSTM obtaining a mean accuracy of 0.87, mean specificity of 0.89 and mean F1 score of 0.86.	Philips Ingenia 3T.
Bahadur, 2019 [347]	LSTM	Proposed LSTM model achieves a classification accuracy of 97.08% having a standard deviation of 0.23% among 20 fold of testing dataset.	24,116 instances of thirteen activities.
Bois, 2019 [348]	LSTM	Proposed approach outperforms all baseline results. More precisely, it trades a loss of 4.3% in the prediction accuracy for an improvement of the clinical acceptability of 27.1%.	Ohio T1DM and the IDIAB dataset
Bois, 2019 [349]	LSTM	The proposed models do not necessarily exhibit a good clinical acceptability, measured by the CG-EGA. Only the LSTM, SVR and GP-DP models have overall acceptable results.	T1DMS software
Massaro, 2019 [350]	LSTM	A percentage improvement of test set accuracy of 6.5% has been observed by applying the LSTM-AR- approach, comparing results with up-to-date MLP works. The LSTM-AR-neural network can be applied as an alternative approach for all homecare platforms where not enough training sequential dataset is available.	Dataset made by 768 records.
Padmapritha, 2019 [351]	LSTM	The LSTM method showed the superior performance in forecasting the blood glucose level with 1.79mmol/l (18.79 mg/dl) RMSE and the maximum error between predicted and observed blood glucose value is 1.94 mmol/l (30mg/dl).	200 blood glucose data observed.
Idrissi, 2019 [355]	LSTM	The results show that our LSTM NN is significantly more accurate; in fact, it outperforms the existing LSTM model for all patients and outperforms the AR model in 9 over 10 patients	10 patients' datasets
Carrillo-Moreno, 2020 [352]	LSTM	Results show that predictors with a PH of 30 min provide the best compromise between the amount of time that a patient has to modify the treatment and the performance of the model predictions.	Guardian Real Time CGM sensor.
Wang, 2020 [356]	LSTM	The experimental results showed that the proposed VMD-IPSO-LSTM model could achieve high prediction accuracy at 30min, 45min and 60min in advance. The improved accuracy of blood glucose prediction and longer prediction time can provide sufficient time for physicians and patients to control blood glucose concentrations and improve the effectiveness of diabetes treatment.	RT_CGM dataset.

Table 12 (continued)

Author, Year, References	Algorithms	Experimental Results	Datasets
Amalia, 2021 [353]	LSTM	Detection and description generation of DR using two deep learning architectures, i.e., CNN and LSTM, have been done in this paper with an accuracy of around 90%. The description sentence obtained from the model can help radiologists as a consideration in the diagnosis of the class of DR.	MESSIDOR data set
Idrissi, 2021 [354]	LSTM	The results show that the proposed CNN outperformed significantly the LSTM model for both one-step and multi-steps prediction and no MSF strategy outperforms the others for CNN.	10 patients' datasets
Alex, 2022 [430]	LSTM	Proposed method achieved the highest prediction accuracy of 99.64%.	PIMA
Prendin, 2023 [431]	LSTM	p-LSTM and np-LSTM) with similar prediction accuracy could lead to different therapeutic decisions.	OhioT1DM dataset

Table 13 (MLP) Diabetes diagnosis summary

Author, Year, References	Algorithms	Experimental Results	Datasets
Kayaer, 2003 [365]	MLP	The performance of RBF was worse than the MLP for all spread values tried. Although the Levenberg-Marquardt training algorithm of MLP gives the best result for the training data. The best result achieved on the test data is the one using the GRNN structure (80.21%).	Pima Indians Diabetes
Ergun, 2004 [366]	MLP	Correct classification size in RBF neural network is 88.4% and the same size in MLP network is 94.2%. As a conclusion, while classifying Doppler signals of the diabetes and control group, MLP neural network is more successful than RBF neural network.	Firat University from the 104 adults
Quchani, 2007 [367]	MLP	The performance of Elman Recurrent Neural Network is better than the MLP Neural Network for blood glucose levels prediction	10 Iranian type-1 diabetic
Bhatkar, 2015 [368]	MLP	It was observed that MLPNN classifier performing well with 11 hidden PEs, learning rule momentum, transfer function tanh and step size 0.1. System obtained relatively high accuracy of 100% for training and cross validation datasets.	64-point DCT along with 09 statistical parameters
Kumar, 2015 [400]	MLP	Various data mining techniques like C4.5, random forest (RF), Bayes Net and Multi-Layer Perceptron (MLP) are trained using randomly training data set and after that the testing of the trained models is done using randomly tested data set. Partitions of data plays very important role in accuracy of models.	Pima Indians Diabetes
Ambilwade, 2016 [369]	MLP	The proposed model is tested on 385 patient's data and gives 91% of classification accuracy, specificity of 94% and sensitivity of 91%.This proposed model will also help doctors for diagnosis of diabetic patients.	385 patient's data.
Choubey, 2016 [370]	MLP	The experimental results obtained classification accuracy (79.1304%) and ROC (0.842) show that GA and MLP NN can be successfully used for the diagnosing of diabetes disease.	Pima Indian Diabetes Database
Alfian, 2018 [371]	MLP	These results show that for a small number of features (2 h glucose tolerance, diastolic blood pressure, body mass index, and age), the MLP algorithm achieved the highest accuracy of prediction compared to other models.	Pima Indian Diabetes Database
Mohapatra, 2019 [372]	MLP	The common MLP classifier is utilized for attributes and the experiment is learned with R studio platform. The performance found to be better as compared to earlier methods and verified in MATLAB platform as well. In this experiment, 77.5% classification accuracy is obtained.	Pima Indian Diabetes Database
Bani-Salameh, 2020 [373]	MLP	The results indicate that MLP correctly predicts the probability of being diseased or not, and the performance can be significantly increased compared with both SVM and KNN. This shows MLP's effectiveness in early disease prediction.	Pima Indian Diabetes Database

Table 13 (continued)

Author, Year, References	Algorithms	Experimental Results	Datasets
Guldogan, 2020 [374]	MLP	The performance values obtained with MLP from the applied models were accuracy 78.1%, specificity 81.2%, AUC 0.848, sensitivity 71%, positive predictive value 61.7%, negative predictive value 86.8% and F-score 66%.	Pima Indian Diabetes Database
Mishra, 2020 [375]	MLP	The proposed classification approach was named as Enhanced and Adaptive-Genetic Algorithm-Multilayer Perceptron (EAGA-MLP). The results show a maximum accuracy rate of 97.76% and 1.12 s of execution time.	Pima Indian Diabetes Database
Bani- Salameh, 2021 [432]	MLP	Correct classification rate (CCR) of 77.6% for diabetes and 68.7% for hypertension.	Collected from 768 individuals.
Sivasankari, 2022 [433]	MLP	MLP outperforms the competition in terms of accuracy, with an accuracy rate of 86.08%.	PIMA
Ali, 2023 [434]	MLP	The proposed MLP-progressive model outperforms existing methods and attains a classification accuracy of 97.14%.	SCADI dataset

Table 14 (AE) Diabetes diagnosis summary

Author, Year, References	Algorithms	Experimental Results	Datasets
BEAULIEU-JONES, 2017 [357]	AE	Autoencoder showed strong performance for imputation accuracy and contributed to the strongest disease progression predictor. Finally, we show that despite clinical heterogeneity, ALS disease progression appears homogenous with time from onset being the most important predictor.	ALS Clinical Trials Database (PRO-ACT)
Hwang, 2018 [358]	AE	The accuracy of filling in the missing values with a stacked autoencoder is higher than that of simply filling in missing values with the mean value. In addition, under the same conditions without oversampling methods, which consume additional memory, generative adversarial networks outperform other disease prediction methods.	Breast Cancer Wisconsin (Diagnostic) dataset.
Babu, 2018 [359]	AE	The performance of GWO+RNN method is calculated in terms of different evaluation metrics like specificity, sensitivity and accuracy. The GWO+RNN method achieved 16.825%of improved accuracy in Cleveland dataset for disease prediction.	Pima Indians Diabetes
DEPERLİOĞLU, 2018 [401]	AE	The obtained classification accuracy is 97.7% and higher than the previously mentioned classification methods. The obtained evaluations show that the proposed method is very efficient and increases the classification success.	Pima Indians Diabetes
KATSUKI, 2018 [402]	Stacked convolutional autoencoder (SCAE)	Proposed approach confirmed that our approach performed better than baseline methods and that the extracted features were promising for understanding the disease.	30,810 patients in a Japanese hospital.

Table 14 (continued)

Author, Year, References	Algorithms	Experimental Results	Datasets
Kannadasan, 2019 [360]	AE	From the results it is evident that proposed model outperforms other model with an accuracy of 86.26%. Furthermore, the model gives precision value of 90.66% and recall of 87.92% which is quite good for the ideal classification model.	Pima Indians Diabetes
Kumar, 2019 [361]	AE	The performance of proposed approach is obtained as 94.5% in terms of prediction accuracy.	Pima Indians Diabetes
Makino, 2019 [403]	AE	AI could predict DKD (diabetic kidney diseases) aggravation with 71% accuracy. Furthermore, the group with DKD aggravation had a signifcantly higher incidence of hemodialysis than the non-aggravation group, over 10 years (N = 2,900).	(EMR) of 64,059 diabetes patients.
Sahoo, 2020 [362]	AE	The experiment results prove that convolution neural network based deep learning method provides the highest accuracy than other machine learning algorithms.	Pima Indians Diabetes
Zhang, 2020 [363]	Stacked sparse Autoencoder	Experimental results on a dataset containing 450 healthy samples, 284 diabetes and 175 lung cancer patients produced the F1-score of 93.57%, 97.54%, 81.56% for detecting healthy, diabetes and lung cancer, respectively, validating the effectiveness of the proposed method.	450 healthy samples, 284 diabetes patient.

Table 14 (continued)

Author, Year, References	Algorithms	Experimental Results	Datasets
Deepthi, 2020 [404]	AE	Conducted 5-fold and 10-fold cross-validation experiments to assess the performance; AE-DNN could achieve AUC scores of 0.9392 and 0.9431, respectively. Experimental results and case studies indicate the robustness of our model in circRNA-disease association prediction.	CircR2Disease database http://bioinfo.snnuedu.cn/CircR2Disease/
García- Ordás, 2021 [364]	Sparse autoencoder (SAE)	A 92.31% of accuracy was obtained when CNN classifier is trained jointly the SAE for featuring augmentation over a well-balanced dataset. This means an increment of 3.17% of accuracy with respect the state-of-the-art.	Pima Indians Diabetes
Tran, 2021 [405]	Hierarchical Autoencoder	Demonstrate that scDHA outperforms state-of-the-art techniques in many research subfields of scRNA-seq analysis, including cell segregation through unsupervised learning, visualization of transcriptome landscape, cell classification, and pseudo-time inference.	34 scRNA-seq data sets
Bodapati, 2022 [435]	stacked convolutional Auto-Encoder	The proposed approach outperforms several existing models by achieving an accuracy of 84.17%.	Kaggle APTOS19
Ismael, 2023 [436]	CAER-DNN	f1-score 97.38%, accuracy, recall 97.25%, specificity 97.59%, precision 97.53%	PIMA

In 2020, Maniruzzaman et al. [168] proposed a decision tree (DT) based machine learning approach for diabetic prediction. Moreover, LR-based model has been adopted to determine the high risk factors of diabetes disease. The high risk factors have been selected based on p-values and odds ratio (OR). Moreover, four classifiers have been also adapted and compared their performance based on ACC, SE, PPV, NPV, FM, and AUC, respectively. The dataset consists of 6561 respondents with 657 diabetic and 5904 controls. The overall ACC of ML-based system is 90.62%. The combination of LR-based feature selection and RF-based classifier gives 94.25% ACC and 0.95 AUC for K10 protocol. Bani-Salameh et al. [373] introduced a model based on the Multi-Layer Perceptron Neural Network (MLP). The main objective of this research is to benefit from ANN's prediction capabilities. Examine whether an MLP neural network can help to precisely predict if patients are diabetes and/or suffer from blood pressure problems. Also, help determine the factor which has a high influence on these diseases. This study presents a prediction method for both diabetes and blood pressure by using ANNs. Python programming language was used to build the neural network model, test its accuracy, and compare it with other neural networks and classifiers. The model predicted the two diseases with correct classification rate (CCR) of 77.6% for diabetes and 68.7% of hypertension. The results indicate that MLP correctly predicts the probability of being diseased or not, and the performance can be significantly increased compared with both SVM and KNN. This shows MLP's effectiveness in early disease prediction.

In 2021, Taser et al. [172] introduced the application of Bagging and Boosting Approaches Using Decision Tree-Based Algorithms in Diabetes Risk Prediction. The proposed model consists of six different decision tree based (DTB) classifiers were implemented on experimental data for diabetes prediction. This work also compares applied individual implementation, bagging, and boosting of DTB classifiers in terms of accuracy rates. The results indicate that the bagging and boosting approaches outperform the individual DTB classifiers, and real Adaptive Boosting (AdaBoost) and bagging using Naive Bayes Tree (NBTree) present the best accuracy score of 98.65%. In this study, bagging and boosting approaches using DTB algorithms were implemented in the experimental data to predict diabetes risk at an early stage. Thyde et al. [386] introduced a model which detects Type 2 Diabetes Patients using deep learning (MLP) approach. In this study, it explores how deep learning (DL) based on CGM data can be used for detecting adherence to once-daily basal insulin injections. It further considered a multilayered feed-forward neural network based on multilayer perceptrons (MLPs) and CNNs, the latter based on the raw CGM as input to the model. The T2D modified version of the MVP model was successfully used to simulate a large amount of realistic CGM data. The data were used to develop methods for treatment adherence detection. The automatically extracted features based on DL methods with added expert-dependent features performed best with an accuracy of 79.8% ± 0.5% 16 hours after TOI.

5.4 Discussion

5.4.1 Machine learning based

This section mainly focuses on a comparison of results for different machine learning approach for prediction of diabetic disease, including SVM, DT, KNN, NB and RF. There are various types of machine learning approach have been utilized by different researcher for prediction of diabetic disease. The machine learning technique is

Table 15 (RBF) Diabetes diagnosis summary

Author, Year, References	Algorithms	Experimental Results	Datasets
Om, 1998 [376]	RBF	The proposed network can replace the conventional statistical method for analyzing relationship between a target disease and health examination parameter. The proposed statistical RBF is expected to applied to almost all statistical problems.	Sample of 30 Data.
Nabney, 1999 [377]	RBF	There are 200 training examples, and 332 test examples. The default classifier (assigining every subject to the healthy class) has an error rate of 33%.	Pima women
Venkatesan, 2006 [378]	RBF	RBF neural networks performed best at ten centres and maximum number of centres tried was 18. Root mean square error using the best centres was 0.3213. Sensitivity of the RBF neural network model was 97.3%, specificity was 96.8% and the percentage correct prediction was 97%.	Of the 1200 cases
Sadi, 2015 [379]	RBF	Naive Bayes has the lowest error in RMSE criterion compared to J48 and RBF Network algorithms. RBF Network algorithm has a lower error in the RMSE criterion compared to J48 algorithm. It can be said that the error rate of RMSE criterion in J48 and RBF Network algorithms are equal because they both are 0.4231 and 0.4239, respectively.	Pima Indians Dataset
Ashiquzzaman, 2018 [380]	RBF	The proposed method has the advantages of dropout as the regu-larization, which gives the network a considerable boost in performance. As a result, the overfitting issue that has been plaguing the other methods has minimal effect on the proposed method.	Pima Indians Dataset
Chetoui, 2018 [381]	RBF	The experimental results show that LESH is the best performing technique with an obtained accuracy of 0.904 using SVM with a Radial Basis Function kernel (SVMRBF). Similarly, the analysis of the ROC curve shows that LESH with SVM-RBF gives the best AUC (Area Under Curve) performance with 0.931.	1200 color images of the eyes retinal fundus.
Adegoke, 2019 [382]	RBF	The model outputs an accuracy of 96% when EKF-RBFN was applied as a base classifier compared to 94% when Decision Stump was applied and AdaBoost as an ensemble technique in both cases.	UCI repository Dataset
Hosseini, 2019 [383]	RBF	The dataset used is extracted from the UCI database. The accuracy of the proposed method is 97.14% which is significantly higher than other models of diabetes diagnosis.	UCI database
Kamble, 2020 [384]	RBF	From the experimental analysis it is observed that the intended system yielded 0.83 Sensitivity & 0.043 Specificity for DIARETDB0, whereas for 0.94 Sensitivity and 0.16 Specificity for DIARETDB1 has been observed.	DIARETDB0 and DIARETDB1 dataset

Table 15 (continued)

Author, Year, References	Algorithms	Experimental Results	Datasets
Rashmi, 2021 [437]	RBF	RBFNN neural model producing accuracies of 67% and 72%.	PIMA
Sivaraman, 2023 [438]	RBF	The result shows that the RBFNN algorithms give good accuracy than SVM in predicting diabetics.	PIMA

able to solve these issues which are faced by the doctors to diagnose properly the diabetic patients, it also helps the patient for early detection of diabetics, so that by taking the prior precaution the disease of diabetes can be minimized. The related works on machine learning based diabetic prediction model was proposed by [92, 171, 198, 231, 218]. The related work on diabetics using different ML and DL techniques from year 2014 to 2023 is shown in comparative graph which is depicted in fig 5.8. It is evident from the comparative graph of the year 2023 that ML techniques ([409] for SVM, [412] for DT, [414] for KNN, [417] for NB, and [419] for RF) as well as DL techniques ([429] for ANN, [426] for RNN, [424] for CNN, [434] for MLP, and [438] for RBF) were utilized for diabetic prediction.

In 2021, Dinesh et al. [92] proposed a Diabetes Mellitus Prediction System Using Hybrid KPCA-GA-SVM Feature Selection Techniques. The proposed work implements the Kernel Principal Component Analysis for dimensionality reduction. Genetic Algorithm to select the relevant and optimal features from the dataset. Then at the last Support Vector Machine is used as a classifier to classify the diabetes mellitus data. The proposed KPCA-GA-SVM obtains accuracy of 99.53% and also reduced feature size compared to GA-SVM of 98.79% accuracy. It is also proved that proposed algorithm performs better in terms of sensitivity (96.4%), specificity (94%), accuracy (97.3%) and MCC (89.3%) compared to other classification algorithms.

In 2020, Haq et al. [171] introduced an intelligent machine learning approach for effective recognition of diabetes in E- Healthcare uses clinical data. In this work, a filter method based on the Decision Tree (Iterative Dichotomiser 3) algorithm for highly important feature selection. Two ensemble learning algorithms, AdaBoost and Random Forest, are also used for feature selection and also compared the classifier performance with wrapper based feature selection algorithms. Classifier Decision Tree has been used for the classification of healthy and diabetic subjects. The experimental results show that the proposed feature selection algorithm selected features improve the classification performance of the predictive model and achieved optimal accuracy. The proposed method DT (ID3) +DT achieved 99% test accuracy, 99.8% accuracy with k-floods and 99.9% accuracy with LOSO validation. The accuracy rate achieved by the proposed method is higher than the accuracy rate achieved by authors in [92].

In 2021, Bhardwaj et al. [198] introduced a hierarchical severity level grading (HSG) system for the detection and classification of diabetic retinopathy (DR). In this work, SVM and KNN classification algorithm are utilized for the prediction of diabetic retinopathy. The proposed system achieves an overall accuracy of 98.10% by SVM classifier and 100% by kNN classifier. Hierarchal discrimination into further grades of abnormalities resulted in accuracy values of 95.68% and 92.60% with SVM classifier using Gaussian kernel and, 97.90% and 95.30% employing fine kNN classifier. Gaussian RBF kernel for SVM classifier and fine kNN classifier provides better performance in terms of diferent indices due to the robustness of these classifiers for non-linear classifcation problems. The accuracy rate achieved by the proposed method is higher than the accuracy rate achieved by authors in [171] and as well as by authors in [92].

In 2020, Rghioui et al. [218] presents an intelligent architecture for the surveillance of diabetic disease that will allow physicians to remotely monitor the health of their patients through sensors integrated into smartphones and smart portable devices. The classification algorithms used in the study were the naive Bayes (NB), J48, random tree, ZeroR, SMO (sequential minimal optimization), and OneR algorithms. Results demonstrated that the J48 algorithm exhibited excellent classification, with the highest accuracy of 99.17%, a sensitivity of 99.47% and a precision of 99.32%.NB achieved an accuracy of 85.16%, which is lower than the accuracy rate achieved by authors in [171] and as well as by authors in [92] and [198].

In 2021, Wang et al. [231] present an exploratory study on classification of diabetes mellitus through a combined Random Forest Classifier. This study explored different supervised classifiers, combined with SVM-SMOTE and two feature dimensionality reduction methods (Logistic stepwise regression and LAASO) to classify the diabetes survey sample data by unbalanced categories and complex related factors. Analysis and discussion of the classification results of 4 supervised classifiers based on 4 data processing methods. According to the result, Random Forest Classifier is combining SVM-SMOTE resampling technology and LASSO feature screening method (Accuracy= 0.890, Precision = 0.869, Recall = 0.919, F1-Score = 0.893, AUC= 0.948) proved the best way to tell those at high risk of DM. Besides, the combined algorithm helps enhance the classification performance for prediction of high-risk people of DM. The results showed that the Random Forest classifier combining with SVM-SMOTE and LASSO feature reduction method performs best in telling high-risk patients of DM from ordinary individuals. The proposed approach achieved a higher accuracy level than authors in [218], but it is less than the accuracy level of authors proposed by [198] and authors in [171] and [92].

5.4.2 Disadvantage of existing (machine learning) techniques

This section mainly focuses on disadvantages of existing work for different machine learning approach for prediction of diabetic disease, including SVM, DT, KNN, NB and RF. Diabetes Mellitus Prediction System Using Hybrid KPCA-GA-SVM Feature Selection Techniques proposed by Dinesh et al. [92] shows promising results, the hybrid approach combining Kernel Principal Component Analysis (KPCA), Genetic Algorithm (GA), and Support Vector Machine (SVM) introduces increased complexity to the model. This complexity may lead to higher computational costs, longer training times, and increased resource requirements, especially for large datasets. The hybrid approach may sacrifice interpretability for improved performance. With multiple layers of feature selection and dimensionality reduction techniques, it may become challenging to interpret how individual features contribute to the prediction model. The high accuracy reported by the proposed method could potentially indicate overfitting, especially if the model is trained and evaluated on the same dataset. It is essential to validate the model's performance on unseen data to ensure that it generalizes well to new instances.

While the intelligent machine learning approach introduced by Haq et al. [171] for the recognition of diabetes in E-Healthcare demonstrates impressive results, while the feature selection algorithms like Decision Tree (Iterative Dichotomiser 3), AdaBoost, and Random Forest are powerful techniques for selecting relevant features, they may introduce bias in feature selection. The choice of features and the criteria used to select them can influence the model's performance and may not always capture the most informative features for diabetes recognition across diverse datasets or populations. Ensemble learning methods such as AdaBoost and Random Forest can be computationally expensive, especially when dealing with large datasets or a high number of features. The performance of the proposed method heavily relies on the quality and representativeness of the clinical data used for training and evaluation. Inaccurate or incomplete data, common in healthcare datasets, can lead to biased model predictions and undermine the effectiveness of the proposed approach.

While the hierarchical severity level grading (HSG) system proposed by Bhardwaj et al. [198] for the detection and classification of diabetic retinopathy (DR) shows impressive results, the reported high accuracy rates, especially 100% accuracy by the kNN classifier, raise concerns about potential overfitting or bias in the model. It is essential to validate the model's performance on unseen data to ensure that it generalizes well to new instances and diverse populations. The complex decision boundaries learned by SVM with Gaussian RBF kernel and kNN may make it challenging to understand the underlying factors contributing to the classification decisions, limiting the clinical interpretability of the model's predictions. SVM with Gaussian RBF kernel and kNN classifiers can be computationally expensive, especially when dealing with large datasets or high-dimensional feature spaces.

While the intelligent architecture presented by Rghioui et al. [218] for the surveillance of diabetic disease using smartphone and smart portable devices shows promising results, the effectiveness of the surveillance system heavily relies on the features extracted from sensors integrated into smartphones and smart portable devices. The limited set of features may not capture all relevant aspects of diabetic health, potentially leading to incomplete or inaccurate monitoring of the disease. The accuracy and reliability of the sensor data collected from smartphones and smart portable devices are critical for the effectiveness of the surveillance system. Inaccurate or noisy sensor data may lead to erroneous classifications and misinterpretations of diabetic health status, compromising the reliability of the system. The performance of the surveillance system may vary across different patient populations and demographic groups. The effectiveness of the classification algorithms and the accuracy of the surveillance system may depend on factors such as age, gender, ethnicity, and comorbidities, which may not be adequately addressed in the study. Remote monitoring of patients' health using smartphones and smart portable devices raises privacy and security concerns regarding the collection, storage, and transmission of sensitive health data. Ensuring compliance with data protection regulations and implementing robust security measures is crucial to maintain patient confidentiality and prevent unauthorized access to health information.

While the study by Wang et al. [231] on the classification of diabetes mellitus through a combined Random Forest Classifier offers promising results, dealing with unbalanced categories in the dataset poses challenges for classification algorithms. While methods like SVM-SMOTE can help address class imbalance by oversampling minority classes, they may also introduce biases or overfitting, particularly if not applied carefully or if the underlying assumptions of the data distribution are not met. The combined approach of Random Forest Classifier with SVM-SMOTE and LASSO feature reduction method may increase the computational complexity of the classification model, especially for large datasets or high-dimensional feature spaces. This could limit the scalability of the model, particularly in real-time or resource-constrained environments. While the reported performance metrics (e.g., Accuracy, Precision, Recall, F1-Score, AUC) indicate strong classification performance, there is a risk of overfitting, particularly if the model is not properly validated on unseen data. It is essential to assess the generalizability of the model across different datasets and patient populations to ensure its reliability in diverse clinical settings.

5.4.3 Deep learning based

This section mainly focuses on a comparison of results of different deep learning approach for prediction of diabetic disease, including CNN, RNN, LSTM, MLP and RBF. There are various types of deep learning approach have been utilized by different researcher for prediction of diabetic disease. Deep learning technique is able to solve these issues which are faced by the doctors to diagnose properly the diabetic patients, it also helps the patient for early detection of diabetics, so that by taking the prior precaution the disease of diabetes can be minimized. The related works on deep learning based diabetic prediction model was proposed by [82, 316, 327, 352, 373].

In 2020, Ismail et al. [316] proposed a remote health monitoring system using a deep learning approach based on CNN. In this work, first, the most important health-related factors are selected in the first hidden layer, then in the second layer, a correlation coefficient analysis is conducted to classify the positively and negatively correlated health factors. By exploiting such knowledge of the regular correlated algorithm, the proposed model demonstrated competitive analysis performance on 4,759,777 medical records. The accuracy of diagnosis and referral of our model reached 80.43%; 80.85%; 91.49%; 82.61%; 95.60% with a test dataset, respectively. Regarding the performance study of the proposed model, it provides knowledge related to regular-correlated health parameters of obesity, high blood pressure, and diabetes.

In 2020, Zhu et al. [327] introduce a deep learning model based on a dilated recurrent neural network (DRNN) to provide 30-min forecasts of future glucose levels. The proposed approach outperforms existing glucose forecasting algorithms, including autoregressive models (ARX), support vector regression (SVR) and conventional neural networks for predicting glucose (NNPG) (e.g. RMSE = NNPG, 22.9 mg/dL; SVR, 21.7 mg/dL; ARX, 20.1 mg/dl; DRNN, 18.9 mg/dL on the OhioT1DM dataset). The results suggest that dilated connections can improve glucose forecasting performance efficiency. Compared with the standard RNNs, the recurrent layers in the DRNN model exponentially increase dilation to expand their receptive fields and improve the prediction accuracy. Proposed model results show that the DRNN model achieves the best performance with the smallest RMSE, MARD and time lag. Therefore, it is believed the DRNN model is a promising approach to achieve good BG prediction and has great potential for future research in diabetes management.

In 2020, Carrillo-Moreno et al. [352] a glucose predictor based on long short-term memory (LSTM) neural networks is designed. Different prediction times and input dimensions have been evaluated in order to provide the best prediction to patients. The main goal of this paper is to design and implement a set of predictors with the aim of predicting accurately the glucose level of type 1 diabetes and improving previous results in the literature. First, a main predictor model development will consist of an LSTM fed with previous values of glucose, insulin bolus and meal intake. Based on the main predictive model, a set of predictors specialized in forecasting for different PHs will be deployed. Twelve models have been deployed to achieve the objective of predicting glucose concentration in patients with type 1 diabetes. These models may be grouped by their PH: 5 min, 15 min, 30 min and 45 min. The predictors with a PH of 5 min are the most accurate models, but they do not provide enough time to anticipate therapeutic actions to predict adverse events (hypoglycemia or hyperglycemia) due to the delayed action of insulin infusions. Consequently, the predictors with a PH of 5 min are not useful in a clinical scenario.

In 2020, Bani-Salameh et al. [373] present a Prediction model of Diabetes and Hypertension using Multi-Layer Perceptron (MLP) Neural Networks. The inputs of the network were the factors for each disease, while the output was the prediction of the disease's occurrence. The model performance was compared with other classifiers Support Vector Machine (SVM) and K-Nearest Neighbors (KNN). It used performance metric measures to assess the accuracy and performance of MLP. The model predicted the two diseases with correct classification rate (CCR) of 77.6% for diabetes and 68.7% of hypertension. The results indicate that MLP correctly predicts the probability of being diseased or not, and the performance can be significantly increased compared with both SVM and KNN. This shows MLP's effectiveness in early disease prediction. The results indicate 77.6% accuracy in diabetes and 68% in hypertension. Also, F1-score and MCC values show improvement for MLP compared with both of the two algorithms. This shows that the used method is effective and improves disease prediction.

In 2020, Nnamoko et al. [82] presents a RBF based deep learning approach for prediction of diabetes. Experiments with Naïve Bayes, SVM-RBF, C4.5 and RIPPER show that proposed selective data preprocessing method applied to C4.5 decision tree produced better results than the other three classifiers with 89.5% Accuracy, 90% Precision, 89.4% Recall, 89.5% F-score and 83.5% Kappa. These results are also better than baseline experiments conducted with AdaBoostM1 and Random Forest. The experiment results show that SVM-RBF trained with IQRd +SMOTEd produced the best results. To experimentally demonstrate the significance of this improvement, over the best performing models from the other classifiers, including the baseline models – AdaBoostM1 and Random Forest; it conducted a McNemar's test to compare their predictions. The performance of SVM-RBF trained with IQRd +SMOTEd data led to significant improvement in all but one classifier, i.e., Random Forest. Nevertheless, the result is a clear indication that given the right classifier, models trained in the selective data preprocessing method presented in this study generally responds positively to class imbalance and outliers.

5.4.4 Disadvantage of existing (deep learning) techniques

This section mainly focuses on disadvantages of existing work for different deep learning approach for prediction of diabetic disease, including CNN, RNN, LSTM, MLP and RBF. While the remote health monitoring system proposed by Ismail et al. [316] using a deep learning approach based on CNN demonstrates competitive analysis performance, Deep learning models, such as convolutional neural networks (CNNs), are often considered black-box models due to their complex architectures and numerous parameters. While the model may achieve high accuracy, understanding the underlying factors and features driving the predictions can be challenging. Deep learning models, particularly CNNs, can be computationally expensive and resource-intensive, especially when dealing with large datasets and complex architectures. While the reported accuracy rates on the test dataset are promising, it is essential to validate the performance of the model on independent datasets and real-world clinical scenarios. While the reported accuracy rates on the test dataset are promising, it is essential to validate the performance of the model on independent datasets and real-world clinical scenarios.

While the dilated recurrent neural network (DRNN) proposed by Zhu et al.[327] shows promising results in forecasting glucose levels, like any model, deep learning models, especially those incorporating advanced architectures like DRNNs, can be computationally expensive to train and deploy. Deep learning models, including DRNNs, often require large amounts of data to generalize well and make accurate predictions. Limited data availability or poor data quality could hinder the performance of the model. Deep learning models are often criticized for their lack of interpretability. It can be difficult to understand how the model arrives at its predictions, which may be crucial in healthcare applications where interpretability is important for trust and acceptance.

The glucose predictor based on Long Short-Term Memory (LSTM) neural networks designed by Carrillo-Moreno et al.[352] presents several advantages in predicting glucose levels for type 1 diabetes patients. However, there are also some disadvantages or limitations noted in the study, the predictors with a prediction horizon (PH) of 5 minutes, which are the most accurate models, are deemed not useful in a clinical scenario because they do not provide enough time to anticipate therapeutic actions for adverse events such as hypoglycemia or hyperglycemia. The model relies on inputs such as previous glucose levels, insulin bolus, and meal intake. While these features are important for predicting glucose levels, inaccuracies or missing data in these inputs could affect the reliability and accuracy of the predictions. LSTM neural networks, while powerful for sequence prediction tasks, can be complex and difficult to interpret. Understanding how the model arrives at its predictions, especially in healthcare settings where interpretability is crucial, may pose challenges.

While the prediction model of Diabetes and Hypertension using Multi-Layer Perceptron (MLP) Neural Networks presented by Bani-Salameh et al. [373] demonstrates promising results, MLP neural networks are often criticized for their lack of interpretability. Understanding how the model makes predictions can be challenging, especially in clinical settings where interpretability is crucial for decision-making and trust in the model's output. The performance of the MLP model heavily relies on the quality of the input data and the selection of relevant features. Inadequate or biased data, as well as irrelevant features, could lead to inaccurate predictions and diminish the model's effectiveness. Training MLP neural networks, especially with large datasets and complex architectures, can be computationally expensive and time-consuming. This may pose challenges in resource-constrained environments or real-time prediction scenarios. MLP models have several hyper parameters that need to be tuned to achieve optimal performance. Finding the right combination of hyper parameters requires extensive experimentation and computational resources. The performance of the MLP model may heavily depend on the distribution and diversity of the training data. Models trained on limited or biased datasets may not generalize well to new patients or different populations.

While the RBF-based deep learning approach presented by Nnamoko et al. [82] shows promising results for predicting diabetes, RBF-based deep learning models, like other deep learning approaches, can be complex and difficult to interpret. Understanding the underlying mechanisms and decision-making processes of such models may pose challenges, especially in clinical settings where interpretability is crucial. RBF-based deep learning models have several hyper parameters that need to be tuned to achieve optimal performance. Finding the right combination of hyper parameters requires extensive experimentation and computational resources. Predictive models for diseases such as diabetes raise ethical considerations related to patient privacy, consent, and the potential consequences of false positives or false negatives. Ensuring the responsible use of predictive models in healthcare is essential. The performance of the RBF-based deep learning model may vary across different populations or healthcare settings. Models trained on specific datasets may not generalize well to new patients or diverse demographic groups. While the RBF-based deep learning model demonstrates improved prediction accuracy, its integration into existing clinical workflows and electronic health record systems may pose technical challenges. While the model may achieve high accuracy and performance metrics, the clinical significance of these results should be carefully interpreted. High accuracy does not necessarily guarantee improved patient outcomes, and the model's predictions should be validated in real-world clinical settings.

5.5 Motivation and hypothesis

The main focus of this proposed work is to discuss the previous work done in the area of prediction of diabetics, including Type-1, Type-2, and Gestational diabetes based on different machine and deep learning approach. Although the prediction model of diabetic disease based on the machine and deep learning approach has greatly improved over the last decade, different prediction algorithms and its approach of implementation has been improved in terms of Classification accuracy, enhance the Sensitivity rate, enhance Specificity rate and provide overall true diabetic prediction model. , a lot of work still needs to be done to make the prediction process more practicable so that it can be easily applied on different application of health care system. In addition, this review also compared the different research work based on different machine and deep learning based algorithms and discuss the different performance parameter like TP (number of diabetes patients detected as a patient), FP (number of healthy persons detected as a patient), TN (number of healthy persons detected as healthy), and FN (number of patients detected as healthy), different algorithm are used in prediction model and the different classifier are used in the identification process. It also discusses explicitly the factor that may affect the prediction model of the system.

Various researchers have used different algorithm in their research work for data collection, feature extraction, prediction methodology, classification and performance evaluation to measure the accuracy and strength of the system. It is very difficult to directly compare and contrast many of these studies in terms of their accuracy system and performance, as their method for evaluating the performance differs depending upon the aim of the study. The success of any predictive model for diabetic system is mainly depends on the algorithm used, which is compulsory to combine the information presented by multiple domain experts. The main purpose of any predictive model is to determine the best set of experts in a given problem domain and implement an appropriate function that can optimally combine the decision produced by individual experts.

Interest in machine learning for healthcare has grown immensely, including work in diagnosing diabetic retinopathy, cancer detection, heart failure, and hypertensions. Despite these advances, the direct application of machine learning to healthcare remains fraught with pitfalls. Many of these challenges stem from the nominal goal in healthcare to make personalized predictions using data generated and managed via the medical system, where data collection's primary purpose is to support care, rather than facilitate subsequent analysis. In tackling healthcare tasks, there are factors that should be considered carefully in the design and evaluation of machine learning projects: causality, missingness, and outcome definition. These considerations are important across both modeling frameworks (e.g., supervised vs. unsupervised), and learning targets (e.g., classification vs. regression). Even if all important variables are included in a health care dataset, it is likely that many observations will be missing. Truly complete data are often impractical due to cost and volume. Learning from incomplete, or missing, data has received little attention in the machine learning community. Obtaining reliable outcomes for learning is an important step in defining tasks. Outcomes are often used to create the gold-standard labels needed for supervised prediction tasks, but are crucial in other settings as well, e.g., to ensure well-defined cohorts in a clustering task. There are three key factors to consider with outcome definitions: creating reliable outcomes, understanding the relevance of an outcome clinically, and the subtlety of label leakage.

Machine learning is a general-purpose method of artificial intelligence that can learn relationships from the data without the need to define them a priori. The major appeal is the ability to derive predictive models without a need for strong assumptions about the underlying mechanisms, which are usually unknown or insufficiently defined. The typical machine learning workflow involves four steps: data harmonization, representation learning, model fitting and evaluation. For decades, constructing a machine learning system required careful engineering and domain expertise to transform the raw data into a suitable internal representation from which the learning subsystem, often a classifier, could detect patterns in the data set.

Deep learning is different from traditional machine learning in how representations learn from the raw data. In fact, deep learning allows computational models that are composed of multiple processing layers based on neural networks to learn representations of data with multiple levels of abstraction. The major differences between deep learning and traditional artificial neural networks (ANNs) are the number of hidden layers, their connections and the capability learn meaningful abstractions of the inputs. In fact, traditional ANNs are usually limited to three layers and are trained to obtain supervised representations that are optimized only for the specific task and are usually not generalizable. Differently, every layer of a deep learning system produces a representation of the observed patterns based on the data it receives as inputs from the layer below, by optimizing a local unsupervised criterion. The key aspect of deep learning is that these layers of features are not designed by human engineers, but they are learned from data using a general purpose learning procedure.

More recently deep learning has been applied to process, aggregated EHRs, including both structured (e.g. diagnosis, medications, laboratory tests) and unstructured (e.g. free-text clinical notes) data. In particular, a common approach is to show that deep learning obtains better results than conventional machine learning models with respect to certain metrics, such as Area under the Receiver Operating Characteristic Curve, accuracy and F-score.

Several works applied deep learning to predict diseases from the patient clinical status. Cheng et al. [302] used a four-layer CNN to predict congestive heart failure and chronic obstructive pulmonary disease and showed significant advantages over the baselines. RNNs with long short-term memory (LSTM) hidden units, pooling and word embedding were used in DeepCare [387], an end-to-end deep dynamic network that infers current illness states and predicts future medical outcomes.

The authors also proposed to moderate the LSTM unit with a decay effect to handle irregular timed events (which are typically in longitudinal EHRs). Moreover, they incorporated medical interventions in the model to dynamically shape the predictions. DeepCare was evaluated for disease progression modeling, intervention recommendation and future risk prediction of diabetes and mental health patient cohorts. RNNs with gated recurrent unit (GRU) were used by Choi et al. [388] to develop Doctor AI, an end-to-end model that uses patient history to predict diagnoses and medications for subsequent encounters. The evaluation showed significantly higher recall than shallow baselines and good generalizability by adapting the resulting model from one institution to another without losing substantial accuracy. Differently, Miotto et al. [389] proposed to learn deep patient representations from the EHRs using a three-layer Stacked Denoising Autoencoder (SDA). They applied this novel representation on disease risk prediction using random forest as classifiers. The evaluation was performed on 76 214 patients comprising 78 diseases from diverse clinical domains and temporal windows (up to a 1 year). The results showed that the deep representation leads to significantly better predictions than using raw EHRs or conventional representation learning algorithms (e.g. Principal Component Analysis (PCA), k-means). Moreover, they also showed that results significantly improve when adding a logistic regression layer on top of the last AE to fine-tune the entire supervised network [390]. Similarly, Liang et al. [391] used RBMs to learn representations from EHRs that revealed novel concepts and demonstrated better prediction accuracy on a number of diseases.

Deep learning was also applied to model continuous time signals, such as laboratory results, toward the automatic identification of specific phenotypes. For example, Lipton et al. [392] used RNNs with LSTM to recognize patterns in multivariate time series of clinical measurements. Specifically, they trained a model to classify 128 diagnoses from 13 frequently, but irregularly sampled clinical measurements from patients in pediatric intensive unit care. The results showed significant improvements with respect to several strong baselines, including multilayer perceptron trained in hand-engineered features. Che et al. [393] used SDAs regularized with a prior knowledge based on ICD-9s for detecting characteristic patterns of physiology in clinical time series. Lasko et al. [394] used a two-layer stacked AE (without regularization) to model longitudinal sequences of serum uric acid measurements to distinguish the uric-acid signatures of gout and acute leukemia. Razavian et al. [395] evaluated CNNs and RNNs with LSTM units to predict disease onset from laboratory test measures alone, showing better performances than logistic regression with hand-engineered, clinically relevant features. Neural language deep models were also applied to EHRs, in particular to learn embedded representations of medical concepts, such as diseases, medications and laboratory tests that could be used for analysis and prediction [396]. As an example, Tran et al. [397] used RBMs to learn abstractions about ICD-10 codes on a cohort of 7578 mental health patients to predict suicide risk. A deep architecture based on RNNs also obtained promising results in removing protected health information from clinical notes to leverage the automatic de-identification of free-text patient summaries [398]. The prediction of unplanned patient readmissions after discharge recently received attention as well. In this domain, Nguyen et al. [399] proposed Deepr, an end-to-end architecture based on CNNs, which detects and combines clinical motifs in the longitudinal patient EHRs to stratify medical risks. Deepr performed well in predicting readmission within 6 months and was able to detect meaningful and interpretable clinical patterns.

5.6 Challenges and opportunities

Even though the promising outcome obtained using deep architectures, there stay a few strange difficulties confronting the clinical use of deep learning to health care. Specifically, the following main issues should be considered:

- Volume of data: Deep learning consists of a set of highly comprehensive computational models. One typical example is fully connected multi-layer neural networks, where tons of network parameters need to be estimated properly. The basis to achieve this goal is the availability of huge amounts of data. In fact, while there are no hard guidelines about the minimum number of training documents, a general rule of thumb is to have at least about 10 × the number of samples as parameters in the network. This is also one of the reasons why deep learning is so successful in domains where huge amount of data can be easily collected (e.g. computer vision, speech, natural language). However, health care is a different domain; in fact, we only have approximately 7.5 billion people all over the world (as per September 2016), with a great part not having access to primary health care. Consequently, from a big data perspective, the amount of medical data that is needed to train an effective and robust, deep learning model would be much more comparable with other media.
- Data quality: Unlike other domains where the data are clean and well-structured, health care data are highly heterogeneous, ambiguous, noisy and incomplete. Training a good

deep learning model with such massive and variegate data sets is challenging and needs to consider several issues, such as data sparsity, redundancy and missing values.

- Temporality: The diseases are always progressing and changing over time in a nondeterministic way. However, many existing deep learning models, including those already proposed in the medical domain, assume static vector-based inputs, which cannot handle the time factor in a natural way. Designing deep learning approaches that can handle temporal health care data is an important aspect that will require the development of novel solutions.
- Domain complexity: Different from other application domains (e.g. image and speech analysis), the problems in biomedicine and health care are more complicated. The diseases are highly heterogeneous and for most of the diseases there is still no complete knowledge of their causes and how they progress. Moreover, the number of patients is usually limited in a practical clinical scenario and we cannot ask for as many patients as we want.
- Interpretability: Although deep learning models have been successful in quite a few application domains, they are often treated as black boxes. While this might not be a problem in other more deterministic domains such as image annotation (because the end user can objectively validate the tags assigned to the images), in health care, not only the quantitative algorithmic performance is important, but also the reason why the algorithm works is relevant.

All these challenges introduce several opportunities and future research possibilities to improve the field. Therefore, with all of them in mind, we point out the following directions, which we believe would be promising for the future of deep learning in health care.

- Feature enrichment: Because of the limited amount of patients in the world, we should capture as many features as possible to characterize each patient and find novel methods to jointly process them. The data sources for generating those features need to include, but not to be limited to, EHRs, social media (e.g. there is prior research leveraging patient-reported information on social media for pharmacovigilance), wearable devices, environments, surveys, online communities, genome profiles, and omics data such as proteome and so on. The effective integration of such highly heterogeneous data and how to use them in a deep learning model would be an important and challenging research topic.
- Federated inference: Each clinical institution possesses its own patient population. Building a deep learning model by leveraging the patients from different sites without leaking their sensitive information becomes a crucial problem in this setting. Consequently, learning deep model in this federated setting in a secure way will be another important research topic, which will interface with other mathematical domains, such as cryptography (e.g. homomorphic encryption and secure multiparty computation).
- Model privacy: Privacy is an important concern in scaling up deep learning (e.g. through cloud computing services). Machine Learning (ML)-as-a-service (i.e. 'predictive analytics') on a set of common models including deep neural networks. The deployment of intelligent tools for next-generation health care needs to consider these risks and attempt to implement a differential privacy standard.
- Incorporating expert knowledge: The existing expert knowledge of medical problems is invaluable for health care problems. Because of the limited amount of medical data and their various quality problems, incorporating the expert knowledge into the deep learning process to guide it toward the right direction is an important research topic.

For example, the online medical encyclopedia and PubMed abstracts should be mined to extract reliable content that can be included in the deep architecture to leverage the overall performances of the systems. Also semi-supervised learning, an effective scheme to learn from the large amount of unlabeled samples with only a few labeled samples, would be of great potential because of its capability of leveraging both labeled (which encodes the knowledge) and unlabeled samples .

- Temporal modeling: Considering that the time factor is important in all kinds of health care-related problems, in particular in those involving EHRs and monitoring devices, training a time-sensitive deep learning model is critical for a better understanding of the patient condition and for providing timely clinical decision support. Thus, temporal deep learning is crucial for solving health care problems. It expects that RNNs as well as architectures coupled with memory and attention mechanisms will play a more significant role toward better clinical deep architectures.
- Interpretable modeling: Model performance and interpretability are equally important for health care problems. Clinicians are unlikely to adopt a system they cannot understand. Deep learning models are popular because of their superior performance. Yet, how to explain the results obtained from these models and how to make them more understandable is of key importance toward the development of trustable and reliable systems. Deep learning methods are powerful tools that allow computers to learn from the data, so that they can come up with ways to create smarter applications. These approaches have already been used in a number of applications, especially for computer vision and natural language processing. In fact, processing medical data with multilayer neural networks increased the predictive power for several specific applications in different clinical domains.

5.7 Research question or hypothesis

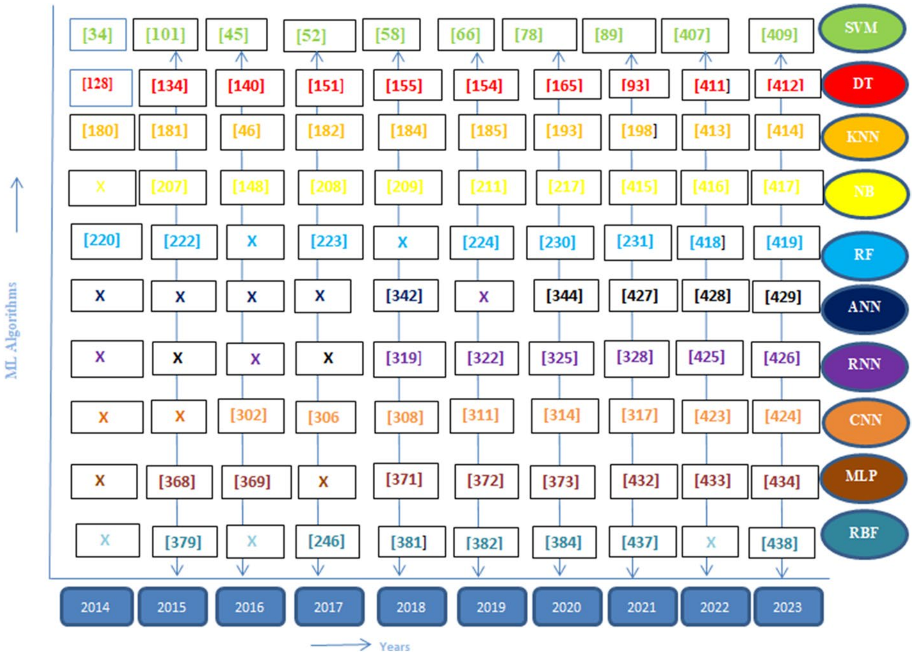
- What extend the research work has to be done in the diabetic prediction model based on different machine and deep learning algorithms from the last two decades?
- What different types of algorithms are used by the researcher in their work in the prediction model?
- What are the different performance parameters that can influence the performance of a prediction model?
- How to enhance the accuracy rate of the prediction model even in the presence of not enough medical data?
- What are the different optimization technique are used in a prediction model to make the system more robust?
- What are the different classification algorithm are used in the prediction model to distinguish between diabetic and non-diabetic patients?
- There is a very limited work in the Autoencoder (AE) based deep learning approach for diabetic prediction model from the last decade, How can improved the performance of the prediction model using AE?
- How we can make the diabetic prediction model is more reliable and secure?
- How can we improve the accuracy rate of diabetic prediction model based on different clustering approach?
- How can we enhance the accuracy rate in a Smartphone enabled diabetic prediction model by using different machines and deep learning approach?

- How the diabetic prediction model based on different association rule approach can be improved the accuracy rate?
- How cloud based diabetic prediction model make more stable in terms of classification accuracy, Sensitivity rate, Specificity rate and provides overall true diabetic prediction model.?
- In future work, what classification algorithms and technique, it should apply to make the prediction model more secure and reliable.
- How can we make deep learning and machine learning based diabetic prediction model more secured using different cryptography technique? (Fig. 5).

6 Conclusion

Machine and deep learning based algorithms play an important role to enhance the overall performance of the diabetic prediction system, in which different classification algorithms are utilized efficiently to form a better prediction system. The proper use of machine or deep learning based algorithms is very important in the diabetic prediction system because it can affect the overall performance and accuracy level of the systems. In designing a diabetic-prediction based system, it is very important, how it can design the classifier for the detection of Diabetes disease with optimal cost and better performance.

This paper is an in-depth study on diabetic prediction strategy, including machine learning (Supervise learning like SVM, KNN, RF, DT, NB and Regression, Unsupervised learning like clustering and association rule) and deep learning (CNN, RNN, AE, RBF and ANN) approaches and its applications in the areas of user diabetic based health prediction



	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
SVM	[34]	[101]	[45]	[52]	[58]	[66]	[78]	[89]	[407]	[409]
DT	[128]	[134]	[140]	[151]	[155]	[154]	[165]	[93]	[411]	[412]
KNN	[180]	[181]	[46]	[182]	[184]	[185]	[193]	[198]	[413]	[414]
NB	X	[207]	[148]	[208]	[209]	[211]	[217]	[415]	[416]	[417]
RF	[220]	[222]	X	[223]	X	[224]	[230]	[231]	[418]	[419]
ANN	X	X	X	X	[342]	X	[344]	[427]	[428]	[429]
RNN	X	X	X	X	[319]	[322]	[325]	[328]	[425]	[426]
CNN	X	X	[302]	[306	[308]	[311]	[314]	[317]	[423]	[424]
MLP	X	[368]	[369]	X	[371]	[372]	[373]	[432]	[433]	[434]
RBF	X	[379]	X	[246]	[381]	[382]	[384]	[437]	X	[438]

Fig. 5 Comparative Graph

model. The main reason behind the success of any health related prediction system are totally depends on the machine and deep learning based algorithm methodology. The main focus of this study is to discuss the methodology and approaches or algorithms used in different diabetic prediction model to enhance the performance of the system and compare the results of related works based on the different diabetic prediction system. Although the prediction model of diabetic disease based on the machine and deep learning approach has greatly improved over the last decade, different prediction algorithms and its approach of implementation has been improved in terms of Classification accuracy, enhance the Sensitivity rate, enhance Specificity rate and provide overall true diabetic prediction model., a lot of work still needs to be done to make the prediction process more practicable so that it can be easily applied on different application of health care system.

It is also discussing the strength and weaknesses of various research works in the area of the diabetic prediction system and providing a comprehensive review of the system based on machine and deep learning. Furthermore, this review paper provides a comparative discussion which represents the different algorithms used in the previous research works since from the year 2010 to 2021 in the area of the diabetic prediction model which can help to find out how proper used a machine and deep learning based algorithm in the diabetic prediction model improved from year to year and in the future how it can make the model more secure by using the proper technique. It discusses the classification of machine learning based algorithms utilized in the area of diabetic prediction systems, namely supervised and unsupervised. The main problem associated with mostly diabetic-prediction system is that the optimization process which required extra time for computation and this will affect the prediction model, choosing an inappropriate optimization technique may result in a very low accuracy rate and affect the overall performance of the model. Hence this review paper is an in depth study of various machine and affect the overall performance of the model. Hence this review paper is an in depth study of various machines and deep learning based algorithms used in the field of the diabetic prediction system and has made clear why more research needs to be done to find a solution to the stated problems found in various diabetic prediction system, also the shortcoming of the various prediction techniques.

Data availability Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Conflict of interest There is no conflict of interest in the current research.

References

1. Bloomgarden Z (2016) Questioning glucose measurements used in the International Diabetes Federation (IDF) Atlas. J Diab 8(6):746–747. https://doi.org/10.1111/1753-0407.12453
2. Ming Z, Wang X, Zhu X (2014) Understanding diabetes from the diagnosis of diabetes mellitus. J Diagn Concept Pract 2:226–228
3. Rajesh K, Sangeetha V (2012) Application of Data Mining Methods and Techniques for Diabetes Diagnosis. Int J Eng Innov Technol 2(3):224–229
4. Kaveeshwar SA, Cornwall J (2014) The current state of diabetes mellitus in India. Australas Med J 7(1):45

5. Wild S, Roglic G, Green A, Sicree R, King H (2004) Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. Diabetes Care *27*(5):1047–1053
6. Shaw JE, Simpson RW (2009) Prevention of type 2 diabetes. Diabetes and Exercise. Springer, pp 55–62
7. IDF Diabetes Atlas - 8th Edition. Available from: http://www.diabetesatlas.org/across-the-globe.html. Accessed 31 Dec. 2017
8. Kaur P, Sharma M (2018) Analysis of data mining and soft computing techniques in prospecting diabetes disorder in human beings: a review. Int J Pharm Sci Res 9:2700–2719
9. Sun YL, Zhang DL (2019) Machine learning techniques for screening and diagnosis of diabetes: a survey. Tehnički Vjesnik 26(3):872–880
10. Yoo I, Alafaireet P, Marinov M, Pena-Hernandez K, Gopidi R, Chang JF, Hua L (2012) Data mining in healthcare and biomedicine: a survey of the literature. J Med Syst 36(4):2431–2448
11. Fatima M, Pasha M (2017) Survey of machine learning algorithms for disease diagnostic. J Intell Learn Syst Appl 9(01):1
12. Kaur H, Kumari V (2020) Predictive modelling and analytics for diabetes using machine learning approach. Appl Comput Inform 18(1/2):90–100
13. Javitt JC, Aiello LP, Chiang Y, Ferris FL, Canner JK, Greenfield S (1994) Preventive eye care in people with diabetes is cost-saving to the federal government: implications for health-care reform. Diabetes Care 17(8):909–917
14. Mendonca AM, Campilho AJ, Nunes JM (1999) Automatic segmentation of microaneurysms in retinal angiograms of diabetic patients. In Proceedings 10th International Conference on Image Analysis and Processing. IEEE. pp 728-733
15. Cree MJ, Olson JA, McHardy KC, Forrester JV, Sharp PF (1996) Automated microaneurysm detection. In Proceedings of 3rd IEEE International Conference on Image Processing. vol. 3. IEEE. pp 699-702
16. Zhang X, Chutatape O (2005) A SVM approach for detection of hemorrhages in background diabetic retinopathy. In Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005. vol. 4. IEEE. pp 2435-2440
17. Stoean R, Stoean C, Preuss M, El-Darzi E, Dumitrescu D (2006) Evolutionary support vector machines for diabetes mellitus diagnosis. In 2006 3rd International IEEE Conference Intelligent Systems. IEEE. pp 182-187
18. Balakrishnan S, Narayanaswamy R, Savarimuthu N, Samikannu R (2008) SVM ranking with backward search for feature selection in type II diabetes databases. In 2008 IEEE International Conference on Systems, Man and Cybernetics. IEEE. pp 2628-2633
19. Polat K, Güneş S, Arslan A (2008) A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine. Expert Syst Appl 34(1):482–487
20. Wu J, Diao YB, Li ML, Fang YP, Ma DC (2009) A semi-supervised learning based method: Laplacian support vector machine used in diabetes disease diagnosis. Interdiscip Sci: Comput Life Sci 1(2):151–155
21. Yu W, Liu T, Valdez R, Gwinn M, Khoury MJ (2010) Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC Med Inform Decis Mak 10(1):1–7
22. Barakat N, Bradley AP, Barakat MNH (2010) Intelligible support vector machines for diagnosis of diabetes mellitus. IEEE Trans Inf Technol Biomed 14(4):1114–1120
23. Çalişir D, Doğantekin E (2011) An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine Classifier. Expert Syst Appl 38(7):8311–8315
24. Gupta S, Kumar D, Sharma A (2011) Performance analysis of various data mining classification techniques on healthcare data. Int J Comput Sci Inform Technol 3(4):155–169
25. Marling C, Wiley M, Cooper T, Bunescu R, Shubrook J, Schwartz F (2011) The 4 diabetes support system: a case study in CBR research and development. In International Conference on Case-Based Reasoning. Springer, Berlin, Heidelberg. pp 137-150
26. Zolfaghari R (2012) Diagnosis of diabetes in female population of pima indian heritage with ensemble of bp neural network and svm. Int J Comput Eng Manag 15:2230–7893
27. Giveki D, Salimi H, Bahmanyar G, Khademian Y (2012) Automatic detection of diabetes diagnosis using feature weighted support vector machines based on mutual information and modified cuckoo search. arXiv preprint arXiv:1201.2173
28. Hashim MF, Hashim SZM (2012) Comparison of clinical and textural approach for Diabetic Retinopathy grading. In 2012 IEEE International Conference on Control System, Computing and Engineering. IEEE. pp 290-295

29. Karatsiolis S, Schizas CN (2012) Region based Support Vector Machine algorithm for medical diagnosis on Pima Indian Diabetes dataset. In 2012 IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE). IEEE. pp 139-144
30. Kumari VA, Chitra R (2013) Classification of diabetes disease using support vector machine. Int J Eng Res Appl 3(2):1797–1801
31. Farran B, Channanath AM, Behbehani K, Thanaraj TA (2013) 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
32. Mansour RF, Abdelrahim EM and Al-Johani AS (2013) Identification of diabetic retinal exudates in digital color images using support vector machine
33. Tapak L, Mahjub H, Hamidi O, Poorolajal J (2013) Real-data comparison of data mining methods in prediction of diabetes in Iran. Healthc Inform Res 19(3):177
34. Anthimopoulos MM, Gianola L, Scarnato L, Diem P, Mougiakakou SG (2014) A food recognition system for diabetic patients based on an optimized bag-of-features model. IEEE J Biomed Health Inform 18(4):1261–1271
35. Choi SB, Kim WJ, Yoo TK, Park JS, Chung JW, Lee YH, Kang EU, Kim DW (2014) Screening for prediabetes using machine learning models. Comput Math Methods Med 2014(1):618976
36. Roychowdhury S, Koozekanani DD, Parhi KK (2014) DREAM: diabetic retinopathy analysis using machine learning. IEEE J Biomed Health Inform 18(5):1717–1728
37. Cai L, Wu H, Li D, Zhou K, Zou F (2015) Type 2 diabetes biomarkers of human gut microbiota selected via iterative sure independent screening method. PLoS One 10(10):e0140827
38. Jaya T, Dheeba J, Singh NA (2015) Detection of hard exudates in colour fundus images using fuzzy support vector machine-based expert system. J Digit Imaging 28(6):761–768
39. Arjun C, Anto M (2015) Diagnosis of diabetes using support vector machine and ensemble learning approach. Int J Eng Appl Sci 2(11):257790
40. Kang S, Kang P, Ko T, Cho S, Rhee SJ, Yu KS (2015) An efficient and effective ensemble of support vector machines for anti-diabetic drug failure prediction. Expert Syst Appl 42(9):4265–4273
41. Ramanathan TT, Sharma D (2015) An SVM-Fuzzy Expert System design for diabetes risk classification. Int J Comput Sci Inform Technol 6(3):2221–2226
42. Santhanam T, Padmavathi MS (2015) Application of K-means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis. Procedia Comput Sci 47:76–83
43. Sowjanya K, Singhal A, Choudhary C (2015) MobDBTest: A machine learning based system for predicting diabetes risk using mobile devices. In 2015 IEEE International Advance Computing Conference (IACC). IEEE. pp 397-402
44. Tafa Z, Pervetica N, Karahoda B (2015) An intelligent system for diabetes prediction. In 2015 4th Mediterranean Conference on Embedded Computing (MECO). IEEE. pp 378-382
45. Abdillah AA, Suwarno S (2016) Diagnosis of diabetes using support vector machines with radial basis function kernels. Int J Technol 7(5)
46. Bano S, Khan MNA (2016) A Framework to Improve Diabetes Prediction using k-NN and SVM. Int J Comput Sci Inform Sec 14(11):450
47. Gill NS, Mittal P (2016) A computational hybrid model with two level classification using SVM and neural network for predicting the diabetes disease. J Theor Appl Inf Technol 87(1):1–10
48. Huang YP, Nashrullah M (2016) SVM-based Decision Tree for medical knowledge representation. In 2016 International Conference on Fuzzy Theory and Its Applications (iFuzzy). IEEE. pp 1-6
49. Kose U, Guraksin GE, Deperlioglu O (2016) Cognitive development optimization algorithm based support vector machines for determining diabetes. Broad Res Artif Intell Neurosci 7(1):80–90
50. Malik S, Khadgawat R, Anand S, Gupta S (2016) Non-invasive detection of fasting blood glucose level via electrochemical measurement of saliva. Springerplus 5(1):1–12
51. Negi A, Jaiswal V (2016) A first attempt to develop a diabetes prediction method based on different global datasets. In 2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC). IEEE. pp 237-241
52. Osman AH, Aljahdali HM (2017) Diabetes disease diagnosis method based on feature extraction using K-SVM. Int J Adv Comput Sci Appl 8(1)
53. Carrera EV, González A, Carrera R (2017) Automated detection of diabetic retinopathy using SVM. In 2017 IEEE XXIV international conference on electronics, electrical engineering and computing (INTERCON). IEEE. pp 1-4
54. Khalil RM, Al-Jumaily A (2017) Machine learning based prediction of depression among type 2 diabetic patients. In 2017 12th international conference on intelligent systems and knowledge engineering (ISKE). IEEE. pp 1-5

55. Rathore A, Chauhan S, Gujral S (2017) Detecting and predicting diabetes using supervised learning: an approach towards better healthcare for women. Int J Adv Res Comput Sci 8(5)
56. Wang Y, Liu ZP (2017) Identifying biomarkers of diabetes with gene co expression networks. In 2017 Chinese Automation Congress (CAC). IEEE. pp 5283-5286
57. Zhang J, Xu J, Hu X, Chen Q, Tu L, Huang J, Cui J (2017) Diagnostic method of diabetes based on support vector machine and tongue images. BioMed Res Int 2017(1):7961494
58. Cui S, Wang D, Wang Y, Yu PW, Jin Y (2018) An improved support vector machine-based diabetic readmission prediction. Comput Methods Prog Biomed 166:123–135
59. Dagliati A, Marini S, Sacchi L, Cogni G, Teliti M, Tibollo V, … Bellazzi R (2018) Machine learning methods to predict diabetes complications. J Diabetes Sci Technol 12(2):295–302
60. Joshi TN, Chawan PM (2018) Logistic regression and svm based diabetes prediction system. Int J Technol Res Eng 5:4347–4350
61. Rao NM, Kannan K, Gao XZ, Roy DS (2018) Novel classifiers for intelligent disease diagnosis with multi-objective parameter evolution. Comput Electr Eng 67:483–496
62. Mule DB, Chowhan SS, Somwanshi DR (2018) Detection and classfication of non-proliferative diabetic retinopathy using retinal images. In International Conference on Recent Trends in Image Processing and Pattern Recognition. Springer, Singapore. pp 312-320
63. Abdullah AS, Gayathri N, Selvakumar S, Kumar SR (2018) Identification of the Risk Factors of Type II Diabetic Data Based Support Vector Machine Classifiers upon Varied Kernel Functions. In Computational Vision and Bio Inspired Computing. Springer, Cham. pp 496-505
64. Sisodia D, Sisodia DS (2018) Prediction of diabetes using classification algorithms. Procedia Comput Sci 132:1578–1585
65. Tsao HY, Chan PY, Su ECY (2018) Predicting diabetic retinopathy and identifying interpretable biomedical features using machine learning algorithms. BMC Bioinform 19(9):111–121
66. Alirezaei M, Niaki STA, Niaki SAA (2019) A bi-objective hybrid optimization algorithm to reduce noise and data dimension in diabetes diagnosis using support vector machines. Expert Syst Appl 127:47–57
67. Bernardini M, Romeo L, Misericordia P, Frontoni E (2019) Discovering the type 2 diabetes in electronic health records using the sparse balanced support vector machine. IEEE J Biomed Health Inform 24(1):235–246
68. Raj RS, Sanjay DS, Kusuma M, Sampath S (2019) Comparison of support vector machine and Naive Bayes classifiers for predicting diabetes. In 2019 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE). IEEE. pp. 41-45
69. He K, Huang S, Qian X (2019) Early detection and risk assessment for chronic disease with irregular longitudinal data analysis. J Biomed Inform 96:103231
70. Karkuzhali S, Manimegalai D (2019) Distinguising Proof of Diabetic Retinopathy Detection by Hybrid Approaches in Two Dimensional Retinal Fundus Images. J Med Syst 43(6):1–12
71. Abbas HT, Alic L, Erraguntla M, Ji JX, Abdul-Ghani M, Abbasi QH, Qaraqe MK (2019) Predicting long-term type 2 diabetes with support vector machine using oral glucose tolerance test. PLoS One 14(12):e0219636
72. Lokuarachchi D, Muthumal L, Gunarathna K, Gamage TD (2019) Detection of red lesions in retinal images using image processing and machine learning techniques. In 2019 Moratuwa Engineering Research Conference (MERCon). IEEE. pp 550-555
73. Aminah R, Saputro AH (2019) Application of machine learning techniques for diagnosis of diabetes based on iridology. In 2019 International Conference on Advanced Computer Science and information Systems (ICACSIS). IEEE. pp 133-138
74. Qomariah DUN, Tjandrasa H, Fatichah C (2019) Classification of diabetic retinopathy and normal retinal images using CNN and SVM. In 2019 12th International Conference on Information & Communication Technology and System (ICTS). IEEE. pp 152-157
75. Selvathi D, Suganya K (2019) Support vector machine based method for automatic detection of diabetic eye disease using thermal images. In 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT). IEEE. pp 1-6
76. Sneha N, Gangil T (2019) Analysis of diabetes mellitus for early prediction using optimal features selection. J Big Data 6(1):1–19
77. Hao Y, Cheng F, Pham M, Rein H, Patel D, Fang Y, … Wang Y (2019) A Noninvasive, Economical, and Instant-Result Method to Diagnose and Monitor Type 2 Diabetes Using Pulse Wave: Case-Control Study. JMIR MHealth UHealth 7(4):e11959
78. Azad C, Mehta AK, Mahto D, Yadav DK (2020) Support Vector Machine based eHealth Cloud System for Diabetes Classification. EAI Endorsed Trans Pervasive Health Technol 6(22):e3

79. Harimoorthy K, Thangavelu M (2020) Multi-disease prediction model using improved SVM-radial bias technique in healthcare monitoring system. J Ambient Intell Humaniz Comput 12(3):3715–3723
80. Jayabalan S, Pratheeksha PS, Bolar NS, Malavika NL (2020) Prediction of diabetic retinopathy using svm algorithm. J Crit Rev 7(14):1702–1711
81. Kazerouni F, Bayani A, Asadi F, Saeidi L, Parvizi N, Mansoori Z (2020) Type2 diabetes mellitus prediction using data mining algorithms based on the long-noncoding RNAs expression: a comparison of four data mining approaches. BMC Bioinformatics 21(1):1–13
82. Nnamoko N, Korkontzelos I (2020) Efficient treatment of outliers and class imbalance for diabetes prediction. Artif Intell Med 104:101815
83. Shuja M, Mittal S, Zaman M (2020) Effective prediction of type ii diabetes mellitus using data mining classifiers and SMOTE. In Advances in computing and intelligent systems. Springer, Singapore. pp 195-211
84. Mishra SK, Tiwari AK (2020) An Ensemble Approach for the Prediction of Diabetes. SAMRIDDHI 12(02):122–129
85. Viloria A, Herazo-Beltran Y, Cabrera D, Pineda OB (2020) Diabetes diagnostic prediction using vector support machines. Procedia Comput Sci 170:376–381
86. Wang X, Yang Y, Xu Y, Chen Q, Wang H, Gao H (2020) Predicting hypoglycemic drugs of type 2 diabetes based on weighted rank support vector machine. Knowl-Based Syst 197:105868
87. Srivastava AK, Kumar Y, Singh PK (2020) Computer aided diagnostic system based on SVM and K harmonic mean based attribute weighting method. Obes Med 19:100270
88. Xue J, Min F, Ma F (2020) Research on Diabetes Prediction Method Based on Machine Learning. In Journal of Physics: Conference Series. vol. 1684, no. 1. IOP Publishing. p 012062
89. Ahmad HF, Mukhtar H, Alaqail H, Seliaman M, Alhumam A (2021) Investigating Health-Related Features and Their Impact on the Prediction of Diabetes Using Machine Learning. Appl Sci 11(3):1173
90. Alabdulwahhab KM, Sami W, Mehmood T, Meo SA, Alasbali TA, Alwadani FA (2021) Automated detection of diabetic retinopathy using machine learning classifiers. Eur Rev Med Pharmacol Sci 25(2):583–590
91. Chaves L, Marques G (2021) Data Mining Techniques for Early Diagnosis of Diabetes: A Comparative Study. Appl Sci 11(5):2218
92. Dinesh MG, Prabha D (2021) Diabetes Mellitus Prediction System Using Hybrid KPCA-GA-SVM Feature Selection Techniques. J Phys Conf Ser. 1767(1):012001
93. Khanam JJ, Foo SY (2021) A comparison of machine learning algorithms for diabetes prediction. ICT Express
94. Reddy SS, Sethi N, Rajender R (2021) Discovering Optimal Algorithm to Predict Diabetic Retinopathy using Novel Assessment Methods. EAI Endorsed Trans Scalable Inf Syst 8(29):e1
95. Rodríguez-Rodríguez I, Rodríguez JV, Woo WL, Wei B, Pardo-Quiles DJ (2021) A Comparison of Feature Selection and Forecasting Machine Learning Algorithms for Predicting Glycaemia in Type 1 Diabetes Mellitus. Appl Sci 11(4):1742
96. Tang H, Zhang Y, Xiang B, Liu M, Hu J, Liu C (2021) Risk prediction of early diabetes mellitus based on combination model. In MATEC Web of Conferences. vol. 336, EDP Sciences. p 07018
97. Hossain ME, Uddin S, Khan A (2021) Network analytics and machine learning for predictive risk modelling of cardiovascular disease in patients with type 2 diabetes. Expert Syst Appl 164:113918
98. Senthil Velmurugan N, Viveka T (2021) Performance analysis of ML algorithms on diabetes data. Int Adv Res J Sci Eng Technol 8(2):72–79
99. Suresh K, Obulesu O, Ramudu BV (2020) Diabetes Prediction using Machine Learning Techniques. Helix 10(02):136–142
100. Brisimi TS, Xu T, Wang T, Dai W, Adams WG, Paschalidis IC (2018) Predicting chronic disease hospitalizations from electronic health records: an interpretable classification approach. Proc IEEE 106(4):690–707
101. Baitharu TR, Pani SK, Dhal S (2015) Comparison of Kernel selection for support vector machines using diabetes dataset. J Comput Sci Appl 3(6):181–184
102. Breault JL, Goodall CR, Fos PJ (2002) Data mining a diabetic data warehouse. Artif Intell Med 26(1-2):37–54
103. Miyaki K, Takei I, Watanabe K, Nakashima H, Watanabe K, Omae K (2002) Novel statistical classification model of type 2 diabetes mellitus patients for tailormade prevention using data mining algorithm. J Epidemiol 12(3):243–248
104. Duhamel A, Nuttens MC, Devos P, Picavet M, Beuscart R (2003) A preprocessing method for improving data mining techniques. Appl Large Med Diab Database Stud Health Technol Inform 95:269–274

105. Huang Y, McCullagh P, Black N and Harper R (2004) Evaluation of outcome prediction for a clinical diabetes database. In International Symposium on Knowledge Exploration in Life Science Informatics (pp. 181-190). Springer, Berlin, Heidelberg
106. Harper PR (2005) A review and comparison of classification algorithms for medical decision making. Health Policy 71(3):315–331
107. Sigurdardottir AK, Jonsdottir H, Benediktsson R (2007) Outcomes of educational interventions in type 2 diabetes: WEKA data-mining analysis. Patient Educ Couns 67(1-2):21–31
108. Huang Y, McCullagh P, Black N, Harper R (2007) Feature selection and classification model construction on type 2 diabetic patients' data. Artif Intell Med 41(3):251–262
109. Liou FM, Tang YC, Chen JY (2008) Detecting hospital fraud and claim abuse through diabetic outpatient services. Health Care Manag Sci 11(4):353–358
110. Toussi M, Lamy JB, Le Toumelin P, Venot A (2009) Using data mining techniques to explore physicians' therapeutic decisions when clinical guidelines do not provide recommendations: methods and example for type 2 diabetes. BMC Med Inform Decis Mak 9(1):1–12
111. Hische M, Luis-Dominguez O, Pfeiffer AF, Schwarz PE, Selbig J, Spranger J (2010) Decision trees as a simple-to-use and reliable tool to identify individuals with impaired glucose metabolism or type 2 diabetes mellitus. Eur J Endocrinol 163(4):565
112. Patil BM, Joshi RC, Toshniwal D (2010) Hybrid prediction model for type-2 diabetic patients. Expert Syst Appl 37(12):8102–8108
113. Ahmad A, Mustapha A, Zahadi ED, Masah N, Yahaya NY (2011) Comparison between neural networks against decision tree in improving prediction accuracy for diabetes mellitus. In International conference on digital information processing and communications. Springer, Berlin, Heidelberg. pp 537-545
114. Al Jarullah AA (2011) Decision tree discovery for the diagnosis of type II diabetes. In 2011 International conference on innovations in information technology. IEEE 303-307
115. Karegowda AG, Manjunath AS, Jayaram MA (2011) Application of genetic algorithm optimized neural network connection weights for medical diagnosis of pima Indians diabetes. Int J Soft Comput 2(2):15–23
116. Kelarev AV, Stranieri A, Yearwood JL, Jelinek HF (2012) Empirical study of decision trees and ensemble classifiers for monitoring of diabetes patients in pervasive healthcare. In 2012 15th International Conference on Network-Based Information Systems. IEEE. pp 441-446
117. Hemant P, Pushpavathi T (2012) A novel approach to predict diabetes by Cascading Clustering and Classification. In 2012 Third International Conference on Computing, Communication and Networking Technologies (ICCCNT'12). IEEE. pp 1-7
118. Hussein AS, Omar WM, Li X, Ati M (2012) Efficient chronic disease diagnosis prediction and recommendation system. In 2012 IEEE-EMBS Conference on Biomedical Engineering and Sciences. IEEE. pp 209-214
119. Rajesh K, Sangeetha V (2012) Application of data mining methods and techniques for diabetes diagnosis. Int J Eng Innov Technol 2(3):224–229
120. Li CP, Zhi XY, Jun MA, Zhuang CUI, Zhu ZL, Zhang C, Hu LP (2012) Performance comparison between logistic regression, decision trees, and multilayer perceptron in predicting peripheral neuropathy in type 2 diabetes mellitus. Chin Med J 125(5):851–857
121. Chen H, Tan C (2012) Prediction of type-2 diabetes based on several element levels in blood and chemometrics. Biol Trace Elem Res 147(1):67–74
122. Karegowda AG, Punya V, Jayaram MA, Manjunath AS (2012) Rule based classification for diabetic patients using cascaded k-means and decision tree C4. 5. Int J Comput Appl 45(12):45–50
123. Karthikeyani V, Begum IP, Tajudin K, Begam IS (2012) Comparative of data mining classification algorithm (CDMCA) in diabetes disease prediction. Int J Comput Appl 60(12)
124. Ameri H, Alizadeh S, Barzegari A (2013) Knowledge extraction of diabetics' data by decision tree method. J Healthc Adm 16(53):58–72
125. Meng XH, Huang YX, Rao DP, Zhang Q, Liu Q (2013) Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. Kaohsiung J Med Sci 29(2):93–99
126. Karthikeyani V, Begum IP (2013) Comparison a performance of data mining algorithms (CPDMA) in prediction of diabetes disease. Int J Comput Sci Eng 5(3):205
127. Rahman RM, Afroz F (2013) Comparison of various classification techniques using different data mining tools for diabetes diagnosis. J Softw Eng Appl 6(03):85
128. Varma KV, Rao AA, Lakshmi TSM, Rao PN (2014) A computational intelligence approach for a better diagnosis of diabetic patients. Comput Electr Eng 40(5):1758–1765
129. Kaur G, Chhabra A (2014) Improved J48 classification algorithm for the prediction of diabetes. Int J Comput Appl 98(22):13–17

130. Seera M, Lim CP (2014) A hybrid intelligent system for medical data classification. Expert Syst Appl 41(5):2239–2249
131. Uppin S, Anusuya MA (2014) Expert system design to predict heart and diabetes diseases. Int J Sci EngTechnol 3(8):1054–1059
132. Ramezankhani A, Pournik O, Shahrabi J, Khalili D, Azizi F, Hadaegh F (2014) Applying decision tree for identification of a low risk population for type 2 diabetes. Tehran Lipid and Glucose Study. Diabetes Res Clin Pract 105(3):391–398
133. Bashir S, Qamar U, Khan FH, Javed MY (2014) An efficient rule-based classification of Diabetes using ID3, C4. 5, & CART ensembles. In 2014 12th International Conference on Frontiers of Information Technology. IEEE. pp 226-231
134. Habibi S, Ahmadi M, Alizadeh S (2015) Type 2 diabetes mellitus screening and risk factors using decision tree: results of data mining. Global J Health Sci 7(5):304
135. Kandhasamy JP, Balamurali SJPCS (2015) Performance analysis of classifier models to predict diabetes mellitus. Procedia Comput Sci 47:45–51
136. Iyer A, Jeyalatha S, Sumbaly R (2015) Diagnosis of diabetes using classification mining techniques. arXiv preprint arXiv:1502.03774
137. Vijayan VV, Anjali C (2015) Prediction and diagnosis of diabetes mellitus—A machine learning approach. In 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS). IEEE. pp 122-127
138. Thirumal PC, Nagarajan N (2015) Utilization of data mining techniques for diagnosis of diabetes mellitus-a case study. ARPN J Eng Appl Sci 10(1):8–13
139. Nai-arun N, Moungmai R (2015) Comparison of classifiers for the risk of diabetes prediction. Procedia Comput Sci 69:132–142
140. Heydari M, Teimouri M, Heshmati Z, Alavinia SM (2016) Comparison of various classification algorithms in the diagnosis of type 2 diabetes in Iran. Int J Diabetes Dev Ctries 36(2):167–173
141. Ahmed TM (2016) Developing a predicted model for diabetes type 2 treatment plans by using data mining. J Theor Appl Inf Technol 90(2):181
142. Ahmed TM (2016) Using data mining to develop model for classifying diabetic patient control level based on historical medical records. J Theor Appl Inf Technol 87(2):316
143. Daghistani T, Alshammari R (2016) Diagnosis of diabetes by applying data mining classification techniques. Int J Adv Comput Sci Appl 7(7):329–332
144. Orabi KM, Kamal YM, Rabah TM (2016) Early predictive system for diabetes mellitus disease. In Industrial Conference on Data Mining. Springer, Cham. pp 420-427
145. Perveen S, Shahbaz M, Guergachi A, Keshavjee K (2016) Performance analysis of data mining classification techniques to predict diabetes. Procedia Comput Sci 82:115–121
146. Pradeep KR, Naveen NC (2016) Predictive analysis of diabetes using J48 algorithm of classification techniques. In 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I). IEEE. pp 347-352
147. Shetty SP, Joshi S (2016) A tool for diabetes prediction and monitoring using data mining technique. Int J Inform TechnolComput Sci 8(11):26–32
148. Songthung P, Sripanidkulchai K (2016) Improving type 2 diabetes mellitus risk prediction using classification. In 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE). IEEE. pp 1-6
149. Srikanth P, Deverapalli D (2016) A critical study of classification algorithms using diabetes diagnosis. In 2016 IEEE 6th International Conference on Advanced Computing (IACC). IEEE. pp 245-249
150. Teimouri M, Farzadfar F, Alamdari MS, Hashemi-Meshkini A, Alamdari PA, Rezaei-Darzi E, … Zeynalabedini A (2016) Detecting diseases in medical prescriptions using data mining tools and combining techniques. Iranian J Pharmaceut Res 15(Suppl):113
151. Chen W, Chen S, Zhang H, Wu T (2017) A hybrid prediction model for type 2 diabetes using K-means and decision tree. In 2017 8th IEEE International conference on software engineering and service science (ICSESS). IEEE. pp 386-390
152. Kasbekar PU, Goel P, Jadhav SP (2017) A decision tree analysis of diabetic foot amputation risk in indian patients. Front Endocrinol 8:25
153. Sayadi M, Zibaeenezhad M, Taghi Ayatollahi SM (2017) Simple prediction of type 2 diabetes mellitus via decision tree modeling. Int Cardiovasc Res J 11(2):71–76
154. Yuvaraj N, SriPreethaa KR (2019) Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster. Clust Comput 22(1):1–9
155. Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H (2018) Predicting diabetes mellitus with machine learning techniques. Front Genet 9:515

156. Kadhm MS, Ghindawi IW, Mhawi DE (2018) An accurate diabetes prediction system based on K-means clustering and proposed classification approach. Int J Appl Eng Res 13(6):4038–4041
157. Esmaily H, Tayefi M, Doosti H, Ghayour-Mobarhan M, Nezami H, Amirabadizadeh A (2018) A comparison between decision tree and random forest in determining the risk factors associated with type 2 diabetes. J Res Health Sci 18(2):412
158. Barhate R, Kulkarni P (2018) Analysis of classifiers for prediction of type ii diabetes mellitus. In 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA). IEEE. pp 1-6
159. Mahmud SH, Hossin MA, Ahmed MR, Noori SRH, Sarkar MNI (2018) Machine learning based unified framework for diabetes prediction. In Proceedings of the 2018 International Conference on Big Data Engineering and Technology. pp 46-50
160. Fiarni C, Sipayung EM, Maemunah S (2019) Analysis and prediction of diabetes complication disease using data mining algorithm. Procedia Comput Sci 161:449–457
161. Hebbar A, Kumar M, Sanjay HA (2019) DRAP: Decision Tree and Random Forest Based Classification Model to Predict Diabetes. In 2019 1st International Conference on Advances in Information Technology (ICAIT). IEEE. pp 271-276
162. Pei D, Zhang C, Quan Y, Guo Q (2019) Identification of potential type II diabetes in a Chinese population with a sensitive decision tree approach. J Diabetes Res 2019(1):4248218
163. Choudhury A, Gupta D (2019) A survey on medical diagnosis of diabetes using machine learning techniques. In Recent developments in machine learning and data analytics. Springer, Singapore. pp 67-78
164. Sun Y, Zhang D (2019) Diagnosis and analysis of diabetic retinopathy based on electronic health records. Ieee Access 7:86115–86120
165. Choubey DK, Kumar P, Tripathi S, Kumar S (2020) Performance evaluation of classification methods with PCA and PSO for diabetes. Netw Model Anal Health Inform Bioinform 9(1):1–30
166. Al-Zebari A, Sengur A (2019) Performance Comparison of Machine Learning Techniques on Diabetes Disease Detection. In 2019 1st International Informatics and Software Engineering Conference (UBMYK). IEEE. pp 1-4
167. Pei D, Yang T, Zhang C (2020) Estimation of Diabetes in a High-Risk Adult Chinese Population Using J48 Decision Tree Model. Diabetes Metab Syndr Obes 13:4621
168. Maniruzzaman M, Rahman MJ, Ahammed B, Abedin MM (2020) Classification and prediction of diabetes disease using machine learning paradigm. Health inform Sci Syst 8(1):1–14
169. Pranto B, Mehnaz S, Mahid EB, Sadman IM, Rahman A, Momen S (2020) Evaluating machine learning methods for predicting diabetes among female patients in bangladesh. Information 11(8):374
170. Tigga NP, Garg S (2020) Prediction of type 2 diabetes using machine learning classification methods. Procedia Comput Sci 167:706–716
171. Haq AU, Li JP, Khan J, Memon MH, Nazir S, Ahmad S, … Ali A (2020) Intelligent Machine Learning Approach for Effective Recognition of Diabetes in E-Healthcare Using Clinical Data. Sensors 20(9):2649
172. Taser PY (2021) Application of Bagging and Boosting Approaches Using Decision Tree-Based Algorithms in Diabetes Risk Prediction. In Multidisciplinary Digital Publishing Institute Proceedings. vol. 74, no. 1. p. 6
173. Chen T, Shang C, Su P, Keravnou-Papailiou E, Zhao Y, Antoniou G, Shen Q (2021) A Decision Tree-Initialised Neuro-fuzzy Approach for Clinical Decision Support. Artif Intell Med 111:101986
174. Emon MU, Zannat R, Khatun T, Rahman M, Keya MS (2021) Performance Analysis of Diabetic Retinopathy Prediction using Machine Learning Models. In 2021 6th International Conference on Inventive Computation Technologies (ICICT). IEEE. pp 1048-1052
175. Lee M, Gatton TM, Lee KK (2010) A monitoring and advisory system for diabetes patient management using a rule-based method and KNN. Sensors 10(4):3934–3953
176. Chikh MA, Saidi M, Settouti N (2012) Diagnosis of diabetes diseases using an artificial immune recognition system2 (AIRS2) with fuzzy k-nearest neighbor. J Med Syst 36(5):2721–2729
177. Aslam MW, Zhu Z, Nandi AK (2013) Feature generation using genetic programming with comparative partner selection for diabetes classification. Expert Syst Appl 40(13):5402–5412
178. Christobel YA, Sivaprakasam P (2013) A new classwise k nearest neighbor (CKNN) method for the classification of diabetes dataset. Int J Eng Adv Technol 2(3):396–200
179. NirmalaDevi M, Alias Balamurugan SA, Swathi UV (2013) An amalgam KNN to predict diabetes mellitus. In 2013 IEEE International Conference ON Emerging Trends in Computing, Communication and Nanotechnology (ICECCN). IEEE. pp 691-695
180. Sarwar A, Sharma V (2014) Comparative analysis of machine learning techniques in prognosis of type II diabetes. AI & Soc 29(1):123–129

181. Farahmandian M, Lotfi Y, Maleki I (2015) Data mining algorithms application in diabetes diseases diagnosis: A case study. Magnt Res Tech Rep 3(1):989–997
182. Hidalgo JI, Colmenar JM, Kronberger G, Winkler SM, Garnica O, Lanchares J (2017) Data based prediction of blood glucose concentrations using evolutionary methods. J Med Syst 41(9):1–20
183. Kumar PS, Pranavi S (2017) Performance analysis of machine learning algorithms on diabetes dataset using big data analytics. In 2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions)(ICTUS). IEEE. pp 508-513
184. Aiello EM, Toffanin C, Messori M, Cobelli C, Magni L (2018) Postprandial glucose regulation via KNN meal classification in type 1 diabetes. IEEE Control Syst Lett 3(2):230–235
185. Mittal K, Aggarwal G, Mahajan P (2019) Performance study of K-nearest neighbor classifier and K-means clustering for predicting the diagnostic accuracy. Int J Inf Technol 11(3):535–540
186. Dey SK, Hossain A, Rahman MM (2018) Implementation of a web application to predict diabetes disease: an approach using machine learning algorithm. In 2018 21st international conference of computer and information technology (ICCIT). IEEE. pp 1-5
187. Azrar A, Ali Y, Awais M, Zaheer K (2018) Data mining models comparison for diabetes prediction. Int J Adv Comput Sci Appl 9(8):320–323
188. Alehegn M, Joshi RR, Mulay P (2019) Diabetes analysis and prediction using random forest KNN Naïve Bayes and J48: An ensemble approach. Int J Sci Technol Res 8(9):1346–1354
189. Aminah R, Saputro AH (2019) Diabetes prediction system based on iridology using machine learning. In 2019 6th International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE). IEEE. pp 1-6
190. Faruque MF, Sarker IH (2019) Performance analysis of machine learning techniques to predict diabetes mellitus. In 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE). IEEE. pp 1-4
191. Dahiwade D, Patle G, Meshram E (2019) Designing disease prediction model using machine learning approach. In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC). IEEE. pp 1211-1215
192. El-Sappagh S, Elmogy M, Ali F, Abuhmed T, Islam SM, Kwak KS (2019) A comprehensive medical decision–support framework based on a heterogeneous ensemble classifier for diabetes prediction. Electronics 8(6):635
193. Ali AMEER, Alrubei MA, Hassan LFM, Al-Ja'afari MA, Abdulwahed SH (2020) Diabetes classification based on KNN. IIUM Eng J 21(1):175–181
194. Garcia-Carretero R, Vigil-Medina L, Mora-Jimenez I, Soguero-Ruiz C, Barquero-Perez O, Ramos-Lopez J (2020) Use of a K-nearest neighbors model to predict the development of type 2 diabetes within 2 years in an obese, hypertensive population. Med Biol Eng Comput 58(5):991–1002
195. Gupta SC, Goel N (2020) Performance enhancement of diabetes prediction by finding optimum K for KNN classifier with feature selection method. In 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT). IEEE. pp 980-986
196. Hassan AS, Malaserene I, Leema AA (2020) Diabetes Mellitus Prediction using Classification Techniques. Int J Innov Technol Explor Eng 9(5):2080–2084
197. Sarker IH, Faruque F, Alqahtani H, Kalim A (2018) K-nearest neighbor learning based diabetes mellitus prediction and analysis for eHealth services. EAI Endorsed Trans Scalable Inf Syst 7(26):e4–e4
198. Bhardwaj C, Jain S, Sood M (2021) Hierarchical severity grade classification of non-proliferative diabetic retinopathy. J Ambient Intell Humaniz Comput 12(2):2649–2670
199. Mohanty S, Mishra A, Saxena A (2021) Medical Data Analysis Using Machine Learning with KNN. In International Conference on Innovative Computing and Communications. Springer, Singapore. pp 473-485
200. Patra R (2021) Analysis and Prediction Of Pima Indian Diabetes Dataset Using SDKNN Classifier Technique. In IOP Conference Series: Materials Science and Engineering. vol. 1070, no. 1. IOP Publishing. p 012059
201. Shinde VD, Raut JR, Sharma Y (2021) Performance evaluation of various supervised machine learning algorithms for diabetes prediction. Eur J Mol Clin Med 7(8):4921–4925
202. Sopharak A, Dailey MN, Uyyanonvara B, Barman S, Williamson T, Nwe KT, Moe YA (2010) Machine learning approach to automatic exudate detection in retinal images from diabetic patients. J Mod Opt 57(2):124–135
203. Tama BA (2011) An early detection method of type-2 diabetes mellitus in public hospital. Telkomnika 9(2):287–294
204. Guo Y, Bai G, Hu Y (2012) Using bayes network for prediction of type-2 diabetes. In 2012 International Conference for Internet Technology and Secured Transactions. IEEE. pp 471-472

205. Leung RK, Wang Y, Ma RC, Luk AO, Lam V, Ng M, … Chan JC (2013) Using a multi-staged strategy based on machine learning and mathematical modeling to predict genotype-phenotype risk patterns in diabetic kidney disease: a prospective case–control cohort analysis. BMC Nephrol 14(1):1–9
206. Lee BJ, Ku B, Nam J, Pham DD, Kim JY (2013) Prediction of fasting plasma glucose status using anthropometric measures for diagnosing type 2 diabetes. IEEE J Biomed Health Inform 18(2):555–561
207. Huang GM, Huang KY, Lee TY, Weng JTY (2015) An interpretable rule-based diagnostic classification of diabetic nephropathy among type 2 diabetes patients. BMC Bioinform 16(1):1–10
208. Singh DAAG, Leavline EJ, Baig BS (2017) Diabetes prediction using medical data. J Comput Intell Bioinforma 10(1):1–8
209. Das H, Naik B, Behera HS (2018) Classification of diabetes mellitus disease (DMD): a data mining (DM) approach. In Progress in computing, analytics and networking. Springer, Singapore. pp 539-549
210. Insani MI, Alamsyah A, Putra AT (2018) Implementation of Expert System for Diabetes Diseases using Naïve Bayes and Certainty Factor Methods. Sci J Inform 5(2):185–193
211. Uddin S, Khan A, Hossain ME, Moni MA (2019) Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak 19(1):1–16
212. Birjais R, Mourya AK, Chauhan R, Kaur H (2019) Prediction and diagnosis of future diabetes risk: a machine learning approach. SN Appl Sci 1(9):1–8
213. Khan NS, Muaz MH, Kabir A, Islam MN (2019) A Machine Learning-Based Intelligent System for Predicting Diabetes. Int J Big Data Analytics Healthcare 4(2):1–20
214. Sonar P, JayaMalini K (2019) Diabetes prediction using different machine learning approaches. In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC). IEEE. pp 367-371
215. Nakra A, Duhan M (2019) Comparative Analysis of Bayes Net Classifier, Naive Bayes Classifier and Combination of both Classifiers using WEKA. IJ Inf Technol Comput Sci 11:38–45
216. Jackins V, Vimal S, Kaliappan M, Lee MY (2021) AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. J Supercomput 77(5):5198–5219
217. Priya KL, Kypa MSCR, Reddy MMS, Reddy GRM (2020) A Novel Approach to Predict Diabetes by Using Naive Bayes Classifier. In 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184). IEEE. pp 603-607
218. Rghioui A, Lloret J, Harane M, Oumnad A (2020) A Smart Glucose Monitoring System for Diabetic Patient. Electronics 9(4):678
219. Khalilia M, Chakraborty S, Popescu M (2011) Predicting disease risks from highly imbalanced data using random forest. BMC Med Inform Decis Mak 11(1):1–13
220. Casanova R, Saldana S, Chew EY, Danis RP, Greven CM, Ambrosius WT (2014) Application of random forests methods to diabetic retinopathy classification analyses. PLoS One 9(6):e98587
221. Sabariah MMK, Hanifa SA and Sa'adah MS (2014) Early detection of type II Diabetes Mellitus with random forest and classification and regression tree (CART). In 2014 International Conference of Advanced Informatics: Concept, Theory and Application (ICAICTA). IEEE 238-242
222. Butwall M, Kumar S (2015) A data mining approach for the diagnosis of diabetes mellitus using random forest classifier. Int J Comput Appl 120(8)
223. Xu W, Zhang J, Zhang Q, Wei X (2017) Risk prediction of type II diabetes based on random forest model. In 2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB). IEEE. pp 382-386
224. Kumar NK, Vigneswari D, Krishna MV, Reddy GP (2019) An optimized random forest classifier for diabetes mellitus. In Emerging Technologies in Data Mining and Information Security. Springer, Singapore. pp 765-773
225. VijiyaKumar K, Lavanya B, Nirmala I, Caroline SS (2019) Random Forest Algorithm for the Prediction of Diabetes. In 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN). IEEE. pp 1-5
226. Kaur M, Gianey HK, Singh D, Sabharwal M (2019) Multi-objective differential evolution based random forest for e-health applications. Modern Phys Lett B 33(05):1950022
227. Alam MZ, Rahman MS, Rahman MS (2019) A Random Forest based predictor for medical data classification using feature ranking. Inform Med Unlocked 15:100180
228. Kaur P, Kumar R, Kumar M (2019) A healthcare monitoring system using random forest and internet of things (IoT). Multimed Tools Appl 78(14):19905–19916
229. Benbelkacem S, Atmani B (2019) Random forests for diabetes diagnosis. In 2019 International Conference on Computer and Information Sciences (ICCIS). IEEE. pp 1-4

230. Wang J, Shi L (2020) Prediction of medical expenditures of diagnosed diabetics and the assessment of its related factors using a random forest model, MEPS 2000–2015. Int J Qual Health Care 32(2):99–112
231. Wang X, Zhai M, Ren Z, Ren H, Li M, Quan D, … Qiu L (2021) Exploratory study on classification of diabetes mellitus through a combined Random Forest Classifier. BMC Med Inform Decis Mak 21(1):1–14
232. Ooka T, Johno H, Nakamoto K, Yoda Y, Yokomichi H, Yamagata Z (2021) Random forest approach for determining risk prediction and predictive factors of type 2 diabetes: large-scale health check-up data in Japan. BMJ Nutrition, Prevention & Health 4(1):140
233. Padmaja P, Vikkurty S, Siddiqui NI, Dasari P, Ambica B, Rao VV, … Rudraraju VR (2008) Characteristic evaluation of diabetes data using clustering techniques. IJCSNS 8(11):244
234. Khanna S, Agarwal S (2013) An Integrated Approach towards the prediction of Likelihood of Diabetes. In 2013 International Conference on Machine Intelligence and Research Advancement. IEEE. pp 294-298
235. Paul R, Hoque ASML (2010) Clustering medical data to predict the likelihood of diseases. In 2010 fifth international conference on digital information management (ICDIM). IEEE. pp 44-49
236. Al Hazemi F, Youn CH, Al-Rubeaan KA (2011) Grid-based interactive diabetes system. In 2011 IEEE First International Conference on Healthcare Informatics, Imaging and Systems Biology. IEEE. pp 258-263
237. Antonelli D, Baralis E, Bruno G, Cerquitelli T, Chiusano S, Mahoto N (2013) Analysis of diabetic patients through their examination history. Expert Syst Appl 40(11):4672–4678
238. Al-Hazemi F (2014) Grid-based Workflow System for Chronic Disease Study. Life Sci J 11(7):1–3
239. Jeong S, Youn CH, Kim YW, Shim SO (2014) Temporal progress model of metabolic syndrome for clinical decision support system. IRBM 35(6):310–320
240. Kim E, Oh W, Pieczkiewicz DS, Castro MR, Caraballo PJ, Simon GJ (2014) Divisive hierarchical clustering towards identifying clinically significant pre-diabetes subpopulations. In AMIA Annual Symposium Proceedings, vol. 2014. American Medical Informatics Association. p 1815
241. Vijayarani DS, Jothi MP (2014) Hierarchical and partitioning clustering algorithms for detecting outliers in data streams. International Journal of Advanced Research in Computer and Communication Engineering, ISSN, pp 2278–1021
242. Sanakal R, Jayakumari T (2014) Prognosis of diabetes using data mining approach-fuzzy C means clustering and support vector machine. Int J Comput Trends Technol 11(2):94–98
243. Flynt A, Daepp MI (2015) Diet-related chronic disease in the northeastern United States: a model-based clustering approach. Int J Health Geogr 14(1):1–14
244. Barale MS, Shirke DT (2016) Cascaded modeling for PIMA Indian diabetes data. Int J Comput Appl 139(11):1–4
245. Bhatia K, Syal R (2017) Predictive analysis using hybrid clustering in diabetes diagnosis. In 2017 Recent Developments in Control, Automation & Power Engineering (RDCAPE). IEEE. pp 447-452
246. Cheruku R, Edla DR, Kuppili V (2017) Diabetes classification using radial basis function network by combining cluster validity index and bat optimization with novel fitness function. Int J Comput Intell Syst 10(1):247–265
247. Ahlqvist E, Storm P, Käräjämäki A, Martinell M, Dorkhan M, Carlsson A, … Groop L (2018) Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. Lancet Diabetes Endocrinol 6(5):361–369
248. Rani S, Kautish S (2018) Association Clustering and Time Series Based Data Mining in Continuous Data for Diabetes Prediction. In 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS). IEEE. pp 1209-1214
249. Derevitskii IV, Kovalchuk SV (2019) Analysis course of the disease of type 2 diabetes patients using Markov chains and clustering methods. Procedia Comput Sci 156:114–122
250. Lasek P, Mei Z (2019) Clustering and visualization of a high-dimensional diabetes dataset. Procedia Comput Sci 159:2179–2188
251. Raihan M, Islam MT, Farzana F, Raju MGM, Mondal HS (2019) An Empirical Study to Predict Diabetes Mellitus using K-Means and Hierarchical Clustering Techniques. In 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT). IEEE. pp 1-6
252. Nguyen HT, Phan NYK, Luong HH, Cao NH, Huynh HX (2020) Binning approach based on classical clustering for type 2 diabetes diagnosis. Int J Adv Comput Sci Appl 11(3)
253. Syafaah L, Azizah DF, Sofiani IR, Lestandy M, Faruq A (2020) Self-Monitoring and Detection of Diabetes with art Toilet based on Image Processing and K-Means Technique. In 2020 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS). IEEE. pp 87-91

254. Anwar S, Alqarni A, Alafnan A, Alamri A, Mathew S, Ricciardi E, Mathew S, Alamri A, Alafnan A, Alqarni A, Anwar S (2021) Cluster identification of diabetic risk factors among Saudi population. J Pharma Res Int 3(8)45–58

255. Takahashi K, Uchiyama H, Yanagisawa S, Kamae I (2006) The logistic regression and ROC analysis of group-based screening for predicting diabetes incidence in four years. Kobe J Med Sci 52(6):171

256. Sparacino G, Zanderigo F, Corazza S, Maran A, Facchinetti A, Cobelli C (2007) Glucose concentration can be predicted ahead in time from continuous glucose monitoring sensor time-series. IEEE Trans Biomed Eng 54(5):931–937

257. Eren-Oruklu M, Cinar A, Quinn L, Smith D (2009) Estimation of future glucose concentrations with subject-specific recursive linear models. Diabetes Technol Ther 11(4):243–253

258. Eren-Oruklu M, Cinar A, Quinn L, Smith D (2009) Adaptive control strategy for regulation of blood glucose levels in patients with type 1 diabetes. J Process Control 19(8):1333–1346

259. Gani A, Gribok AV, Rajaraman S, Ward WK, Reifman J (2008) Predicting subcutaneous glucose concentration in humans: data-driven glucose modeling. IEEE Trans Biomed Eng 56(2):246–254

260. Estrada GC, Kirchsteiger H, del Re L, Renard E (2010) Innovative approach for online prediction of blood glucose profile in type 1 diabetes patients. In Proceedings of the 2010 American Control Conference. IEEE. pp 2015-2020

261. Gani A, Gribok AV, Lu Y, Ward WK, Vigersky RA, Reifman J (2009) Universal glucose models for predicting subcutaneous glucose concentration in humans. IEEE Trans Inf Technol Biomed 14(1):157–165

262. Lu Y, Rajaraman S, Ward WK, Vigersky RA, Reifman J (2011) Predicting human subcutaneous glucose concentration in real time: a universal data-driven approach. In 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE. pp 7945-7948

263. Zhao C, Dassau E, Zisser HC, Jovanovič L, Doyle FJ III, Seborg DE (2014) Online prediction of subcutaneous glucose concentration for type 1 diabetes using empirical models and frequency-band separation. AICHE J 60(2):574–584

264. Georga EI, Protopappas VC, Ardigo D, Marina M, Zavaroni I, Polyzos D, Fotiadis DI (2012) Multivariate prediction of subcutaneous glucose concentration in type 1 diabetes patients based on support vector regression. IEEE J Biomed Health Inform 17(1):71–81

265. Bayrak ES, Turksoy K, Cinar A, Quinn L, Littlejohn E, Rollins D (2013) Hypoglycemia early alarm systems based on recursive autoregressive partial least squares models. J Diabetes Sci Technol 7(1):206–214

266. Yu C, Zhao C (2014) Rapid model identification for online glucose prediction of new subjects with type 1 diabetes using model migration method. IFAC Proc Volumes 47(3):2094–2099

267. Zhao C, Yu C (2015) Rapid model identification for online subcutaneous glucose concentration prediction for new subjects with type I diabetes. IEEE Trans Biomed Eng 62(5):1333–1344

268. Paul SK, Samanta M (2015) Predicting upcoming glucose levels in patients with type 1 diabetes using a generalized autoregressive conditional heteroscedasticity modelling approach. Int J Stat Med Res 4(2):188–198

269. Bagherzadeh-Khiabani F, Ramezankhani A, Azizi F, Hadaegh F, Steyerberg EW, Khalili D (2016) A tutorial on variable selection for clinical prediction models: feature selection methods in data mining could improve the results. J Clin Epidemiol 71:76–85

270. Agarwal V, Podchiyska T, Banda JM, Goel V, Leung TI, Minty EP, … Shah NH (2016) Learning statistical models of phenotypes using noisy labeled training data. J Am Med Inform Assoc 23(6):1166–1173

271. Lee BJ, Kim JY (2016) Identification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on machine learning. IEEE J Biomed Health Inform 20(1):39–46

272. Rahimloo P, Jafarian A (2016) Prediction of diabetes by using artificial neural network, logistic regression statistical model and combination of them. Bull Soc R Sci Liège 85:1148–1164

273. Rau HH, Hsu CY, Lin YA, Atique S, Fuad A, Wei LM, Hsu MH (2016) Development of a web-based liver cancer prediction model for type II diabetes patients by using an artificial neural network. Comput Methods Prog Biomed 125:58–65

274. Usman S, Reaz MBI, Ali MAM (2016) Risk prediction of having increased arterial stiffness among diabetic patients using logistic regression. In 2016 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES). IEEE. pp 699-701

275. Zhao LP, Bolouri H, Zhao M, Geraghty DE, Lernmark Å, Better Diabetes Diagnosis Study Group (2016) An object-oriented regression for building disease predictive models with multiallelic HLA genes. Genet Epidemiol 40(4):315–332

276. Bajestani NS, Kamyad AV, Esfahani EN, Zare A (2018) Prediction of retinopathy in diabetic patients using type-2 fuzzy regression model. Eur J Oper Res 264(3):859–869

277. Hassan M, Butt MA, Baba MZ (2017) Logistic regression versus neural networks: the best accuracy in prediction of diabetes disease. Asi J Comp Sci Tech 6:33–42
278. Zheng T, Xie W, Xu L, He X, Zhang Y, You M, … Chen Y (2017) A machine learning-based framework to identify type 2 diabetes through electronic health records. Int J Med Inform 97:120–127
279. Wu H, Yang S, Huang Z, He J, Wang X (2018) Type 2 diabetes mellitus prediction model based on data mining. Inform Med Unlocked 10:100–107
280. Qiu S, Li J, Chen B, Wang P, Gao X (2019) An improved prediction method for diabetes based on a feature-based least angle regression algorithm. In Proceedings of the 3rd International Conference on Machine Learning and Soft Computing. pp 232-238
281. Yao L, Zhong Y, Wu J, Zhang G, Chen L, Guan P, … Liu L (2019) Multivariable logistic regression and back propagation artificial neural network to predict diabetic retinopathy. Diabetes Metab Syndr Obes 12:1943
282. Alshamlan H, Taleb HB, Al Sahow A (2020) A Gene Prediction Function for Type 2 Diabetes Mellitus using Logistic Regression. In 2020 11th International Conference on Information and Communication Systems (ICICS). IEEE. pp 1-4
283. Kopitar L, Kocbek P, Cilar L, Sheikh A, Stiglic G (2020) Early detection of type 2 diabetes mellitus using machine learning-based prediction models. Sci Rep 10(1):1–12
284. Hsu W, Lee ML, Liu B, Ling TW (2000) Exploration mining in diabetic patients databases: findings and conclusions. In Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. pp 430-436
285. Stilou S, Bamidis PD, Maglaveras N, Pappas C (2001) Mining association rules from clinical databases: an intelligent diagnostic process in healthcare. Stud Health Technol Inform 2:1399–1403
286. Zorman M, Masuda G, Kokol P, Yamamoto R, Stiglic B (2002) Mining diabetes database with decision trees and association rules. In Proceedings of 15th IEEE Symposium on Computer-Based Medical Systems (CBMS 2002). IEEE. pp 134-139
287. Duru N (2005) An application of apriori algorithm on a diabetic database. In International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (pp. 398-404). Springer, Berlin, Heidelberg
288. Mao W and Mao J (2009) The application of apriori-gen algorithm in the association study in type 2 diabetes. In 2009 3rd International Conference on Bioinformatics and Biomedical Engineering. IEEE 1-4
289. Patil B, Joshi R, Toshniwal D (2010) Association rule for classification of type -2 diabetic patients. In 2010 Second International Conference on Machine Learning and Computing. IEEE 330-334
290. Patil BM, Joshi RC, Toshniwal D (2011) Classification of type-2 diabetic patients by using Apriori and predictive Apriori. Int J Comput Vis Robotics 2(3):254–265
291. Kasemthaweesab P and Kurutach W (2012) Association analysis of diabetes mellitus (DM) with complication states based on association rules. In 2012 7th IEEE Conference on Industrial Electronics and Applications (ICIEA). IEEE 1453-1457
292. Kim HS, Shin AM, Kim MK, Kim YN (2012) Comorbidity study on type 2 diabetes mellitus using data mining. Korean J Int Med 27(2):197
293. Simon GJ, Schrom J, Castro MR, Li PW and Caraballo P J (2013) Survival association rule mining towards type 2 diabetes risk assessment. In AMIA annual symposium proceedings. Am Med Inform Assoc 2013:1293
294. Schrom JR, Caraballo PJ, Castro MR and Simon GJ (2013) Quantifying the effect of statin use in pre-diabetic phenotypes discovered through association rule mining. In AMIA Annual Symposium Proceedings. Am Med Inform Assoc 2013:1249
295. Lakshmi KS and Kumar GS (2014) Association rule extraction from medical transcripts of diabetic patients. In The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014). IEEE 201-206
296. Karthikeyan T, Vembandasamy K (2015) A novel algorithm to diagnosis type II diabetes mellitus based on association rule mining using MPSO-LSSVM with outlier detection method. Indian J Sci Technol 8(S8):310–320
297. Ramezankhani A, Pournik O, Shahrabi J, Azizi F, Hadaegh F (2015) An application of association rule mining to extract risk pattern for type 2 diabetes using tehran lipid and glucose study database. Int J Endocrinol Metabol 13(2)
298. Simon GJ, Caraballo PJ, Therneau TM, Cha SS, Castro MR, Li PW (2013) Extending association rule summarization techniques to assess risk of diabetes mellitus. IEEE Trans Knowl Data Eng 27(1):130–141

299. Kamalesh MD, Prasanna KH, Bharathi B, Dhanalakshmi R and Canessane RA (2016) Predicting the risk of diabetes mellitus to subpopulations using association rule mining. In proceedings of the international conference on soft computing systems (pp. 59-65). Springer, New Delhi
300. Alam TM, Iqbal MA, Ali Y, Wahab A, Ijaz S, Baig TI, … Abbas Z (2019) A model for early prediction of diabetes. Inform Med Unlocked 16:100204
301. Lu PH, Keng JL, Tsai FM, Lu PH, Kuo CY (2021) An apriori algorithm-based association rule analysis to identify acupoint combinations for treating diabetic gastroparesis. Evid-Based Complement Altern Med 2021(1):6649331
302. Cheng Y, Wang F, Zhang P and Hu J (2016). Risk prediction with electronic health records: A deep learning approach. In Proceedings of the 2016 SIAM International Conference on Data Mining. Soc Industrial App Mathematics 432-440
303. Pratt H, Coenen F, Broadbent DM, Harding SP, Zheng Y (2016) Convolutional neural networks for diabetic retinopathy. Procedia Comput Sci 90:200–205
304. Shi X, Hu Y, Zhang Y, Li W, Hao Y, Alelaiwi A, … Hossain MS (2016) Multiple disease risk assessment with uniform model based on medical clinical notes. IEEE Access 4:7074–7083
305. Zhu Z, Yin C, Qian B, Cheng Y, Wei J and Wang F (2016) Measuring patient similarities via a deep architecture with medical concept embedding. In 2016 IEEE 16th International Conference on Data Mining (ICDM). IEEE 749-758
306. Lekha S, Suchetha M (2017) Real-time non-invasive detection and classification of diabetes using modified convolution neural network. IEEE J Biomed Health Inform 22(5):1630–1636
307. Mohebbi A, Aradóttir TB, Johansen AR, Bengtsson H, Fraccaro M and Mørup M (2017) A deep learning approach to adherence detection for type 2 diabetics. In 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE 2896-2899
308. Kwasigroch A, Jarzembinski B and Grochowski M (2018) Deep CNN based decision support system for detection and assessing the stage of diabetic retinopathy. In 2018 International Interdisciplinary PhD Workshop (IIPhDW). IEEE 111-116
309. Swapna G, Kp S, Vinayakumar R (2018) Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals. Procedia Comput Sci 132:1253–1262
310. Swapna G, Vinayakumar R, Soman KP (2018) Diabetes detection using deep learning algorithms. ICT Express 4(4):243–246
311. Butt MM, Latif G, Iskandar DA, Alghazo J, Khan AH (2019) Multi-channel Convolutions Neural Network Based Diabetic Retinopathy Detection from Fundus Images. Procedia Comput Sci 163:283–291
312. Khan SH, Abbas Z and Rizvi SD (2019) Classification of diabetic retinopathy images based on customised CNN architecture. In 2019 Amity International Conference on Artificial Intelligence (AICAI). IEEE 244-248
313. Sun Y (2019) The neural network of one-dimensional convolution-an example of the diagnosis of diabetic retinopathy. IEEE Access 7:69657–69666
314. Raj MAH, Al Mamun M and Faruk MF (2020) CNN Based Diabetic Retinopathy Status Prediction Using Fundus Images. In 2020 IEEE Region 10 Symposium (TENSYMP). IEEE 190-193
315. Rahman M, Islam D, Mukti RJ, Saha I (2020) A deep learning approach based on convolutional LSTM for detecting diabetes. Comput Biol Chem 88:107329
316. Ismail WN, Hassan MM, Alsalamah HA, Fortino G (2020) CNN-based health model for regular health factors analysis in Internet-of-medical things environment. IEEE Access 8:52541–52549
317. Islam MT, Al-Absi HR, Ruagh EA, Alam T (2021) DiaNet: A deep learning based architecture to diagnose diabetes using retinal images only. IEEE Access 9:15686–15695
318. Allam F, Nossai Z, Gomma H, Ibrahim I, Abdelsalam M (2011) A recurrent neural network approach for predicting glucose concentration in type-1 diabetic patients, In Engineering Applications of Neural Networks (pp. 254-259). Springer, Berlin, Heidelberg
319. Chu J, Dong W, He K, Duan H, Huang Z (2018) Using neural attention networks to detect adverse medical events from electronic health records. J Biomed Inform 87:118–130
320. Wang WW, Li H, Cui L, Hong X and Yan Z (2018) Predicting clinical visits using recurrent neural networks and demographic information. In 2018 IEEE 22nd International Conference on Computer Supported Cooperative Work in Design ((CSCWD)). IEEE 353-358
321. Wu S, Liu S, Sohn S, Moon S, Wi CI, Juhn Y, Liu H (2018) Modeling asynchronous event sequences with RNNs. J Biomed Inform 83:167–177
322. Dong Y, Wen R, Zhang K and Zhang L (2019) A Novel RNN-Based Blood Glucose Prediction Approach Using Population and Individual Characteristics. In 2019 IEEE 7th International Conference on Bioinformatics and Computational Biology (ICBCB). IEEE 145-149

323. Dong Y, Wen R, Li Z, Zhang K and Zhang L (2019) Clu-RNN: a new RNN based approach to diabetic blood glucose prediction. In 2019 IEEE 7th International Conference on Bioinformatics and Computational Biology (ICBCB). IEEE 50-55
324. Jang JS, Lee MJ, Lee TR (2019) Development of T2DM Prediction Model Using RNN. J Digit Converg 17(8):249–255
325. Munoz-Organero M (2020) Deep Physiological Model for Blood Glucose Prediction in T1DM Patients. Sensors 20(14):3896
326. Zhou H, Myrzashova R, Zheng R (2020) Diabetes prediction model based on an enhanced deep neural network. EURASIP J Wirel Commun Netw 2020(1):1–13
327. Zhu T, Li K, Chen J, Herrero P, Georgiou P (2020) Dilated recurrent neural networks for glucose forecasting in type 1 diabetes. J Healthc Inform Res 4(3):308–324
328. Rabby MF, Tu Y, Hossen MI, Lee I, Maida AS, Hei X (2021) Stacked LSTM based deep recurrent neural network with kalman smoothing for blood glucose prediction. BMC Med Inform Decis Mak 21(1):1–15
329. Martinsson J, Schliep A, Eliasson B, Meijner C, Persson S and Mogren O (2018) Automatic blood glucose prediction with confidence using recurrent neural networks. In KHD@ IJCAI
330. Chen J, Li K, Herrero P, Zhu T and Georgiou P (2018) Dilated Recurrent Neural Network for Short-time Prediction of Glucose Concentration. In KHD@ IJCAI (pp. 69-73)
331. Jaafar SFB and Ali DM (2005) Diabetes mellitus forecast using artificial neural network (ANN). In 2005 Asian Conference on Sensors and the International Conference on New Techniques in Pharmaceutical and Biomedical Research. IEEE 135-139
332. Mougiakakou SG, Prountzou A, Iliopoulou D, Nikita KS, Vazeou A and Bartsocas CS (2006) Neural network based glucose-insulin metabolism models for children with type 1 diabetes. In 2006 International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE 3545-3548
333. Dey R, Bajpai V, Gandhi G and Dey B (2008) Application of artificial neural network (ANN) technique for diagnosing diabetes mellitus. In 2008 IEEE Region 10 and the Third international Conference on Industrial and Information Systems. IEEE 1-4
334. Pappada SM, Cameron BD, Rosman PM (2008) Development of a neural network for prediction of glucose concentration in type 1 diabetes patients. J Diabetes Sci Technol 2(5):792–801
335. Zainuddin Z, Pauline O, Ardil C (2009) A neural network approach in predicting the blood glucose level for diabetic patients. Int J Comput Intell 5(1):72–79
336. Pérez-Gandía C, Facchinetti A, Sparacino G, Cobelli C, Gómez EJ, Rigla M, … Hernando ME (2010) Artificial neural network algorithm for online glucose prediction from continuous glucose monitoring. Diabetes Technol Ther 12(1):81–88
337. Pappada SM, Borst MJ, Cameron BD, Bourey RE, Lather JD, Shipp D, … Papadimos TJ (2010) Development of a neural network model for predicting glucose levels in a surgical critical care setting. Patient Safety Surg 4(1):1–5
338. Chakraborty M and Tudu B (2010) Comparison of ANN models to predict LDL level in Diabetes Mellitus type 2. In 2010 International Conference on Systems in Medicine and Biology. IEEE 392-396
339. Allam F, Nossair Z, Gomma H, Ibrahim I and Abd-el Salam M (2011) Prediction of subcutaneous glucose concentration for type-1 diabetic patients using a feed forward neural network. In The 2011 International Conference on Computer Engineering and Systems. IEEE 129-133
340. Pappada SM, Cameron BD, Rosman PM, Bourey RE, Papadimos TJ, Olorunto W, Borst MJ (2011) Neural network-based real-time prediction of glucose in patients with insulin-dependent diabetes. Diabetes Technol Ther 13(2):135–141
341. Robertson G, Lehmann ED, Sandham W, Hamilton D (2011) Blood glucose prediction using artificial neural networks trained with the AIDA diabetes simulator: a proof-of-concept pilot study. J Electr Comput Eng 2011(1):681786
342. Ali JB, Hamdi T, Fnaiech N, Di Costanzo V, Fnaiech F, Ginoux JM (2018) Continuous blood glucose level prediction of Type 1 Diabetes based on Artificial Neural Network. Biocybern Biomed Eng 38(4):828–840
343. Kathiroli R, RajaKumari R and Gokulprasanth P (2018) Diagnosis Of Diabetes Using Cascade Correlation And Artificial Neural Network. In 2018 Tenth International Conference on Advanced Computing (ICoAC). IEEE 299-306
344. Senturk Z (2020) Artificial Neural Networks based decision support system for the detection of diabetic retinopathy. Sakarya Univ Fen Bilim Enst Derg 24(2):424–431
345. Sun Q, Jankovic MV, Bally L and Mougiakakou SG (2018) Predicting blood glucose with an LSTM and Bi-LSTM based deep neural network. In 2018 14th Symposium on Neural Networks and Applications (NEUREL). IEEE 1-5

346. Farías AFS, Mendizabal A, González-Garrido AA, Romo-Vázquez R and Morales A (2018) Long Short-Term Memory Neural Networks for Identifying Type 1 Diabetes Patients with Functional Magnetic Resonance Imaging. In 2018 IEEE Latin American Conference on Computational Intelligence (LA-CCI). IEEE 1-4
347. Bahadur EH, Masum AKM, Barua A, Alam MGR, Chowdhury MAUZ and Alam MR (2019) LSTM Based Approach for Diabetic Symptomatic Activity Recognition Using Smartphone Sensors. In 2019 22nd International Conference on Computer and Information Technology (ICCIT). IEEE 1-6
348. De Bois M, El Yacoubi MA and Ammi M (2019) Prediction-coherent LSTM-based recurrent neural network for safer glucose predictions in diabetic people. In International Conference on Neural Information Processing (pp. 510-521). Springer, Cham
349. De Bois M, El Yacoubi MA and Ammi M (2019) Study of short-term personalized glucose predictive models on type-1 diabetic children. In 2019 International Joint Conference on Neural Networks (IJCNN). IEEE 1-8
350. Massaro A, Maritati V, Giannone D, Convertini D, Galiano A (2019) LSTM DSS automatism and dataset optimization for diabetes prediction. Appl Sci 9(17):3532
351. Padmapritha T (2019) Prediction of Blood Glucose Level by using an LSTM based Recurrent Neural networks. In 2019 IEEE International Conference on Clean Energy and Energy Efficient Electronics Circuit for Sustainable Development (INCCES). IEEE 1-4
352. Carrillo-Moreno J, Pérez-Gandía C, Sendra-Arranz R, García-Sáez G, Hernando ME, Gutiérrez A (2020) Long short-term memory neural network for glucose prediction. Neural Comput Applic 33:4191–4203
353. Amalia R, Bustamam A, Sarwinda D (2021) Detection and description generation of diabetic retinopathy using convolutional neural network and long short-term memory. J Phys Conf Ser 1722(1):012010
354. El Idrissi T and Idri A (2020) Deep Learning for Blood Glucose Prediction: CNN vs LSTM. In International Conference on Computational Science and Its Applications (pp. 379-393). Springer, Cham
355. El Idriss T, Idri A, Abnane I and Bakkoury Z (2019) Predicting blood glucose using an LSTM neural network. In 2019 Federated Conference on Computer Science and Information Systems (FedCSIS). IEEE 35-41
356. Wang W, Tong M, Yu M (2020) Blood Glucose Prediction With VMD and LSTM Optimized by Improved Particle Swarm Optimization. IEEE Access 8:217908–217916
357. Beaulieu-Jones BK, Moore JH and POOLED RESOURCE OPEN-ACCESS ALS CLINICAL TRIALS CONSORTIUM (2017) Missing data imputation in the electronic health record using deeply learned autoencoders. In Pacific Symposium on Biocomputing 2017 (pp. 207-218)
358. Hwang U, Choi S, Lee HB and Yoon S (2018) Adversarial training for disease prediction from electronic health records with missing data. arXiv preprint arXiv:1711.04126
359. Babu SB, Suneetha A, Babu GC, Kumar YJN, Karuna G (2018) Medical disease prediction using grey wolf optimization and auto encoder based recurrent neural network. Period Eng Nat Sci 6(1):229–240
360. Kannadasan K, Edla DR, Kuppili V (2019) Type 2 diabetes data classification using stacked autoencoders in deep neural networks. Clin Epidemiol Glob Health 7(4):530–535
361. Kumar VB, Vijayalakshmi K and Padmavathamma M (2019) A hybrid data mining approach for diabetes prediction and classification. In 2019 World Congress on Engineering and Computer Science, WCECS (Vol. 22, pp. 298-303)
362. Sahoo AK, Pradhan C and Das H (2020) Performance evaluation of different machine learning methods and deep-learning based convolutional neural network for health decision making. In Nature inspired computing for data science (pp. 201-212). Springer, Cham
363. Zhang Q, Zhou J and Zhang B (2020) A noninvasive method to detect diabetes mellitus and lung cancer using the stacked sparse autoencoder. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE 1409-1413
364. García-Ordás MT, Benavides C, Benítez-Andrades JA, Alaiz-Moretón H, García-Rodríguez I (2021) Diabetes detection using deep learning techniques with oversampling and feature augmentation. Comput Methods Prog Biomed 202:105968
365. Kayaer K and Yildirim T (2003) 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).
366. Ergün U, Barýþçý N, Ozan AT, Serhatlýoðlu S, Oğur E, Hardalaç F, Güler İ (2004) Classification of MCA stenosis in diabetes by MLP and RBF neural network. J Med Syst 28(5):475–487
367. Quchani SA, Tahami E (2007) Comparison of MLP and Elman neural network for blood glucose level prediction in type 1 diabetics, In 3rd Kuala Lumpur International Conference on Biomedical Engineering 2006 (pp. 54-58). Springer, Berlin, Heidelberg

368. Bhatkar AP and Kharat GU (2015) Detection of diabetic retinopathy in retinal images using MLP classifier. In 2015 IEEE international symposium on nanoelectronic and information systems. IEEE 331-335
369. Ambilwade RP and Manza RR (2016) Prognosis of diabetes using fuzzy inference system and multilayer perceptron. In 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I). IEEE 248-252
370. Choubey DK, Paul S (2016) GA_MLP NN: a hybrid intelligent system for diabetes disease diagnosis. Int J Intell Syst Appl 8(1):49
371. Alfian G, Syafrudin M, Ijaz MF, Syaekhoni MA, Fitriyani NL, Rhee J (2018) A personalized healthcare monitoring system for diabetic patients by utilizing BLE-based sensors and real-time data processing. Sensors 18(7):2183
372. Mohapatra SK, Swain JK, Mohanty MN (2019) Detection of diabetes using multilayer perceptron, In International conference on intelligent computing and applications (pp. 109-116). Springer, Singapore
373. Bani-Salameh H, Alkhatib SM, Abdalla M, Banat R, Zyod H, Alkhatib AJ (2020) Prediction of diabetes and hypertension using multi-layer perceptron neural networks. Int J Model Simul Sci Comput 12(02):2150012
374. Güldoğan E, Zeynep TUNÇ, Ayça ACET, Çolak C (2020) Performance Evaluation of Different Artificial Neural Network Models in the Classification of Type 2 Diabetes Mellitus. J Cogn Syst 5(1):23–32
375. Mishra S, Tripathy HK, Mallick PK, Bhoi AK, Barsocchi P (2020) EAGA-MLP—An Enhanced and Adaptive Hybrid Classification Model for Diabetes Diagnosis. Sensors 20(14):4036
376. Om KS, Kim HC, Min BG, Shin CS, Lee HK (1998) Statistical RBF Network with Applications to an Expert System for Characterizing Diabetes Mellitus. J Electr Eng Inf Sci 3(3):355–365
377. Nabney IT (2004) Efficient training of RBF networks for classification. Int J Neural Syst 14(03):201–208
378. Venkatesan P, Anitha S (2006) Application of a radial basis function neural network for diagnosis of diabetes mellitus. Curr Sci 91(9):1195–1199
379. Sa'di S, Maleki A, Hashemi R, Panbechi Z, Chalabi K (2015) Comparison of data mining algorithms in the diagnosis of type II diabetes. Int J Comput Sci Appl 5(5):1–12
380. Ashiquzzaman A, Tushar AK, Islam MR, Shon D, Im K, Park JH, … and Kim J (2018) Reduction of overfitting in diabetes prediction using deep learning neural network. In IT convergence and security 2017 (pp. 35-43). Springer, Singapore
381. Chetoui M, Akhloufi MA and Kardouchi M (2018) Diabetic retinopathy detection using machine learning and texture features. In 2018 IEEE Canadian Conference on Electrical & Computer Engineering (CCECE). IEEE 1-4
382. Adegoke V, Chen D, Banissi E (2019) Improving prediction accuracy of breast cancer survivability and diabetes diagnosis via RBF networks trained with EKF models. Int J Comput Inf Syst Ind Manag 11:19 –19
383. Hosseini H, Bardsiri AK (2019) Improving Diagnosis Accuracy of Diabetic Disease Using Radial Basis Function Network and Fuzzy Clustering. Front Health Inform 8(1):24
384. Kamble VV, Kokate RD (2020) Automated diabetic retinopathy detection using radial basis function. Procedia Comput Sci 167:799–808
385. Dwivedi AK (2018) Analysis of computational intelligence techniques for diabetes mellitus prediction. Neural Comput Applic 30(12):3837–3845
386. Thyde DN, Mohebbi A, Bengtsson H, Jensen ML, Mørup M (2021) Machine learning-based adherence detection of type 2 diabetes patients on once-daily basal insulin injections. J Diabetes Sci Technol 15(1):98–108
387. Pham T, Tran T, Phung D and Venkatesh S (2016) Deepcare: a deep dynamic memory model for predictive medicine. In Pacific-Asia conference on knowledge discovery and data mining (pp. 30-41). Springer, Cham
388. Choi E, Bahadori MT, Schuetz A, Stewart WF and Sun J (2016) Doctor ai: Predicting clinical events via recurrent neural networks. In Machine learning for healthcare conference. PMLR 301-318
389. Miotto R, Li L, Kidd BA, Dudley JT (2016) Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Sci Rep 6(1):1–10
390. Miotto R, Li L and Dudley JT (2016) Deep learning to predict patient future diseases from the electronic health records. In European Conference on Information Retrieval (pp. 768-774). Springer, Cham
391. Liang Z, Zhang G, Huang JX and Hu QV (2014) Deep learning for healthcare decision making with EMRs. In 2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE 556-559

392. Lipton ZC, Kale DC, Elkan C and Wetzel R (2015) Learning to diagnose with LSTM recurrent neural networks. arXiv preprint arXiv:1511.03677
393. Che Z, Kale D, Li W, Bahadori MT and Liu Y (2015) Deep computational phenotyping. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 507-516)
394. Lasko TA, Denny JC, Levy MA (2013) Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. PLoS One 8(6):e66341
395. Razavian N, Marcus J and Sontag D (2016) Multi-task prediction of disease onsets from longitudinal laboratory tests. In Machine learning for healthcare conference. PMLR 73-100
396. Choi Y, Chiu CYI, Sontag D (2016) Learning low-dimensional representations of medical concepts. AMIA Summits Transl Sci Proc 2016:41
397. Tran T, Nguyen TD, Phung D, Venkatesh S (2015) Learning vector representation of medical objects via EMR-driven nonnegative restricted Boltzmann machines (eNRBM). J Biomed Inform 54:96–105
398. Dernoncourt F, Lee JY, Uzuner O, Szolovits P (2017) De-identification of patient notes with recurrent neural networks. J Am Med Inform Assoc 24(3):596–606
399. Nguyen P, Tran T, Wickramasinghe N et al (2017) Deepr: a Convolutional Net for Medical Records. IEEE J Biomed Health Inform 21:22–30
400. Kumar Dewangan A, Agrawal P (2015) Classification of diabetes mellitus using machine learning techniques. Int J Eng Appl Sci 2(5):257905
401. Deperlioğlu, O, Köse, U. (2018). Diagnosis of Diabetes by Using Deep Neural Network. 2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT).IEEE.
402. Katsuki T, Ono M, Koseki A, Kudo M, Haida K, Kuroda J, … and Suzuki A (2018) Risk Prediction of Diabetic Nephropathy via Interpretable Feature Extraction from EHR Using Convolutional Autoencoder. In MIE (pp. 106-110)
403. Makino M, Yoshimoto R, Ono M, Itoko T, Katsuki T, Koseki A, … Suzuki A (2019) Artificial intelligence predicts the progression of diabetic kidney disease using big data machine learning. Sci Rep 9(1):1–9
404. Deepthi K, Jereesh AS (2020) An ensemble approach for CircRNA-disease association prediction based on autoencoder and deep neural network. Gene 762:145040
405. Tran D, Nguyen H, Tran B, La Vecchia C, Luu HN, Nguyen T (2021) Fast and precise single-cell data analysis using a hierarchical autoencoder. Nat Commun 12(1):1–10
406. Li K, Daniels J, Liu C, Herrero P, Georgiou P (2019) Convolutional recurrent neural networks for glucose prediction. IEEE J Biomed Health Inform 24(2):603–613
407. Sistla S (2022) Predicting Diabetes using SVM Implemented by Machine Learning. International Journal of Soft Computing and Engineering 12(2):2231–2307
408. Li J, Ding J, Zhi DU, Gu K, Wang H (2022) Identification of type 2 diabetes based on a ten-gene biomarker prediction model constructed using a support vector machine algorithm. BioMed Res Int 2022(1):1230761
409. Rastogi R, Bansal M (2023) Diabetes prediction model using data mining techniques. Meas Sens 25:100605
410. Aslan MF, Sabanci K (2023) A novel proposal for deep learning-based diabetes prediction: Converting clinical data to image data. Diagnostics *13*(4):796
411. Ahamed BS, Arya MS, Nancy VAO (2022) Prediction of type-2 diabetes mellitus disease using machine learning classifiers and techniques. Front Comput Sci 4:835242
412. Özge ŞEN, Keser SB, Keskin K (2023) Early stage diabetes prediction using decision tree-based ensemble learning model. Int Adv Res Eng J 7(1):62–71
413. Suyanto S, Meliana S, Wahyuningrum T, Khomsah S (2022) A new nearest neighbor-based framework for diabetes detection. Expert Syst Appl 199:116857
414. Prasad BS, Gupta S, Borah N, Dineshkumar R, Lautre HK, Mouleswararao B (2023) Predicting diabetes with multivariate analysis an innovative KNN-based classifier approach. Prev Med 174:107619
415. Khanam JJ, Foo SY (2021) A comparison of machine learning algorithms for diabetes prediction. Ict Express 7(4):432–439
416. Hasan MK, Saeed RA, Alsuhibany SA, Abdel-Khalek S (2022) An empirical model to predict the diabetic positive using stacked ensemble approach. Front Public Health 9:792124
417. Okikiola FM, Adewale OS, Obe OO (2023) A diabetes prediction classifier model using naive bayes algorithm. Fudma J Sci *7*(1):253–260
418. Mondal S, Banik A, Roy S, Das J, Banerjee S and Navin H (2022) Random Forest Based Diabetic Prediction Model on Highly Unbalanced Dataset. In 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon). IEEE 1-6

419. Gündoğdu S (2023) Efficient prediction of early-stage diabetes using XGBoost classifier with random forest feature selection technique. Multimed Tools Appl 82(22):34163–34181
420. Hassan MM, Mollick S, Yasmin F (2022) An unsupervised cluster-based feature grouping model for early diabetes detection. Healthc Analyt 2:100112
421. Alghamdi T (2023) Prediction of Diabetes Complications Using Computational Intelligence Techniques. Appl Sci 13(5):3030
422. Khafaga DS, Alharbi AH, Mohamed I, Hosny KM (2022) An integrated classification and association rule technique for early-stage diabetes risk prediction. Healthcare 10(10):2070
423. Madan P, Singh V, Chaudhari V, Albagory Y, Dumka A, Singh R, … AlGhamdi AS (2022) An optimization-based diabetes prediction model using CNN and Bi-directional LSTM in real-time environment. Appl Sci 12(8):3989
424. Aslan MF, Sabanci K (2023) A novel proposal for deep learning-based diabetes prediction: Converting clinical data to image data. Diagnostics *13*(4):796
425. Srinivasu PN, Shafi J, Krishna TB, Sujatha CN, Praveen SP, Ijaz MF (2022) Using recurrent neural networks for predicting type-2 diabetes from genomic and tabular data. Diagnostics 12(12):3067
426. Kiruthiga G, Shakkeera L, Asha A, Dhiyanesh B, Saraswathi P, Murali M (2023) Deep Learning-Based Continuous Glucose Monitoring with Diabetic Prediction Using Deep Spectral Recurrent Neural Network. In: International Conference on Information, Communication and Computing Technology. Springer Nature Singapore, Singapore, pp 485–497
427. Bukhari MM, Alkhamees BF, Hussain S, Gumaei A, Assiri A, Ullah SS (2021) An improved artificial neural network model for effective diabetes prediction. Complexity 2021:1–10
428. Prakash EP, Srihari K, Karthik S, Kamal MV, Dileep P, Bharath Reddy S, Mukunthan MA, Somasundaram K, Jaikumar R, Gayathri N, Sahile K (2022) Implementation of artificial neural network to predict diabetes with high-quality health system. Comput Intell Neurosci 2022(1):1174173
429. Al Sadi K, Balachandran W (2023) Prediction Model of Type 2 Diabetes Mellitus for Oman Prediabetes Patients Using Artificial Neural Network and Six Machine Learning Classifiers. Appl Sci 13(4):2344
430. Alex SA, Jhanjhi NZ, Humayun M, Ibrahim AO, Abulfaraj AW (2022) Deep LSTM Model for Diabetes Prediction with Class Balancing by SMOTE. Electronics 11(17):2737
431. Prendin F, Pavan J, Cappon G, Del Favero S, Sparacino G, Facchinetti A (2023) The importance of interpreting machine learning models for blood glucose prediction in diabetes: an analysis using SHAP. Sci Rep 13(1):16865
432. Bani-Salameh H, Alkhatib SM, Abdalla M, Al-Hami MT, Banat R, Zyod H, Alkhatib AJ (2021) Prediction of diabetes and hypertension using multi-layer perceptron neural networks. Int J Model Simul Sci Comput 12(02):2150012
433. Sivasankari SS, Surendiran J, Yuvaraj N, Ramkumar M, Ravi CN and Vidhya RG (2022)Classification of diabetes using multilayer perceptron. In 2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE). IEEE 1-5
434. Ali R, Hussain J, Lee SW (2023) Multilayer perceptron-based self-care early prediction of children with disabilities. Digital Health 9:20552076231184054
435. Bodapati JD (2022) Stacked convolutional auto-encoder representations with spatial attention for efficient diabetic retinopathy diagnosis. Multimed Tools Appl 81(22):32033–32056
436. Ismael HA, Al-A'araji NH, Shukur BK (2023) Enhanced the prediction approach of diabetes using an autoencoder with regularization and deep neural network. Period Eng Nat Sci 10(6):156–167
437. Rashmi K, Rao NK, Bala MM, Lahari M, Fathima N, Prudhvi V (2021) Prediction of diabetes mellitus using rbf neural model and genetic algorithm. Turkish Journal of Physiotherapy and Rehabilitation 32:3
438. Sivaraman M and Sumitha J (2023) An efficiency of DCKSVM and HRBFNN techniques for diabetic prediction. In AIP Conference Proceedings (Vol. 2831, No. 1). AIP Publishing
439. Zhang C, Hu C, Wu T, Zhu L, Liu X (2022) Achieving efficient and privacy-preserving neural network training and prediction in cloud environments. IEEE Trans Dependable Secure Comput 20(5):4245–4257
440. Zhang C, Zhu L, Xu C, Lu R (2018) PPDP: An efficient and privacy-preserving disease prediction scheme in cloud-based e-Healthcare system. Futur Gener Comput Syst 79:16–25
441. Lei D, Liang J, Zhang C, Liu X, He D, Zhu L, Guo S (2023) Publicly verifiable and secure SVM classification for cloud-based health monitoring services. IEEE Internet Things J
442. Mathew TE (2019) A comparative study of the performance of different Support Vector machine Kernels in Breast Cancer Diagnosis. Int J Inf Comput Sci 6(6):432–441
443. Podgorelec V, Kokol P, Stiglic B, Rozman I (2002) Decision trees: an overview and their use in medicine. J Med Syst 26:445–463

444. Dey L, Chakraborty S, Biswas A, Bose B and Tiwari S (2016) Sentiment analysis of review datasets using naive bayes and k-nn classifier. arXiv preprint arXiv:1610.09982
445. Khamis HS (2014) Application of k-Nearest Neighbour classification in medical data mining in the context of kenya. In Scientific Conference Proceedings 2022(1):5416722
446. Breiman L (2001) Random forests. Mach Learn 45:5–32
447. Cutler DR, Edwards TC Jr, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ (2007) Random forests for classification in ecology. Ecology 88(11):2783–2792
448. Montgomery DC, Peck EA, Vining GG (2021) Introduction to linear regression analysis. John Wiley & Sons
449. Nathiya G, Punitha SC, Punithavalli M (2010) An analytical study on behavior of clusters using k means, em and k* means algorithm. arXiv preprint arXiv:1004.1743.
450. Agrawal R, Imieliński T, Swami A (1993) Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD international conference on Management of data. pp 207-216
451. Zhu X, Ghahramani Z, Lafferty JD (2003) Semi-supervised learning using gaussian fields and harmonic functions. In Proceedings of the 20th International conference on Machine learning (ICML-03). pp 912-919
452. Mnih V, Badia AP, Mirza M, Graves A, Lillicrap T, Harley T, … Kavukcuoglu K (2016) Asynchronous methods for deep reinforcement learning. In International conference on machine learning. PMLR. pp 1928-1937
453. Eiben AE, Smith JE (2015) Introduction to evolutionary computing. Springer-Verlag, Berlin Heidelberg
454. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition. pp 770-778
455. Madhiarasan M, Louzazni M (2022) Analysis of artificial neural network: architecture, types, and forecasting applications. J Electr Comput Eng 2022(1):5416722
456. Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, … Farhan L (2021) Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. J Big Data 8:1–74
457. Sherstinsky A (2020) Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Phys D: Nonlinear Phenom 404:132306

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