A review on prediction of diabetes type 2 by machine learning techniques

Alka Singh *

*Professor Anjna Jayant Deen, Professor Sanjay Silakari, Department of Computer Science and Engineering, University Institute of Technology, RGPV, Bhopal, India
Email: alkasingh929@yahoo.com , anjnadeen@rgtu.net , ssilakari@yahoo.com

Article History: Received: 10 November 2020; Revised 12 January 2021 Accepted: 27 January 2021; Published online: 5 April 2021

Abstract: Machine learning is considered to be one of the most promising tools when it comes to working with heterogeneous data. It provides a new dimension which enables one to extract relevant data and take decision for the effective functioning of the network, making use of generated data. Every sphere of our life is now dependent on machine learning. It has flourished in every dimension. Making it versatile and ever demanding. Department of healthcare contains very abundant and sensitive information which is needed to be carefully handled. Diabetes mellitus is increasing exponentially and is spreading like anything in the world. A reliable prediction system should be present for diagnosing diabetes. Variety of machine learning techniques find their use in the examination of data from variant perspectives and summarizing it into effective information. Usage of new patterns is done to elucidate these patterns in order to deliver relevant information for their users. By making use of techniques such as SVM, random forest, logistic regression, naive bayes etc the prediction of diabetes can be done easily and accurately. In this study we will make use of different machine learning techniques and try to find accurate prediction regarding the same.

Keywords: Machine learning, diabetes type 2, supervised, unsupervised, reinforcement, training and algorithm

1. Introduction

Machine learning has the potential which enables it to learn from previous data to generate futuristic trends in behavior. It has the capability to learn by its own. Machine learning can be applied on numerous data making it very integral to the telecommunication world today (Hang Lai et al 2019) [1]. Machine learning methods detect linearity/non linearity in the relationship between dependent and independent variables (Geoffrey et al 2019) [2]. They can be used for making predictions in case of continuous outcomes, known as regression type problems or can be used for making predictions in case of levels of categorical variable, which is known as classification problems. It gives solution from the problems and learns how to tackle with the problem that may or may not be same by making use of training dataset provided to the algorithm earlier.

Diabetes is such a prolonged disease that can happen when body cannot efficiently make use of the insulin it generates. As a result, diabetes affects organs which include heart diseases which could be heart stroke, high blood pressure and atherosclerosis, nerve damage that could lead to numbness, gradually losing all sense of feeling especially in the limbs, kidney failure is very common in diabetic patients, and hearing impairment is also seen in diabetic patients, the risk of Alzheimer’s disease increases with type 2 diabetes.

Diabetes can be categorized into three types:-

(a) Childhood or Juvenile diabetes
(b) Adult or Type 2 diabetes
(c) Type 3 or Gestational diabetes

Generally, type 1 diabetes occur because of the deficiency in insulin production and is commonly found in children. Diabetes type 2 is a chronic disease which affects how the human body metabolizes glucose. In case of diabetes type 2, the human body behaves in either of the 2 ways; firstly it resists the effect of insulin which is a hormone responsible for regulating the movement of sugar in the cells. Secondly it doesn’t produce ample insulin for the maintenance of normal glucose level.

Diabetes type 2 was known to be adult onset disease but nowadays much of the younger age group is being diagnosed with the same, because of the rise in obesity in children. There is no cure.
available for the same but person can switch from sedentary life style, follow balanced diet and can exercise well to manage the disease, as depicted in figure 1. If this would not suffice then the person should go for medications and insulin therapy. This insulin is secreted into the bloodstream by the pancreas. This insulin then circulates, enabling the sugar to enter the body cells. The amount of glucose in the bloodstream is lowered by the insulin. Glucose i.e. sugar, is a major source of energy for cells that make up muscles and other tissues and it comes from food and liver. In case of low glucose level the liver breaks down glycogen into glucose in order to keep the glucose level normal. When it comes to type 2 diabetes, the sugar starts to build up in the bloodstream instead of moving into the cells which lead to more release of insulin by beta cells in the pancreas. Gradually these cells become impaired and become incapable of releasing more insulin to fulfill the requirements of the body whereas in case of type 1 diabetes the immune system by mistake destroys beta cells which leave the body with little or no insulin.

Gestational diabetes is hyperglycemia which happens due to the change in hormones during pregnancy.

For the past few decades we have seen that the machine learning discipline is assisting us to solve different relevant biomedical problems. The machine learning techniques are found to operate in both real-life and scientific problems. In this study, we will be evaluating the performance of various machine learning techniques for the classification of people whether they are diabetic or not.

1.2. Generalized Architecture

![Figure 1](image)

2. Literature survey

This section reviews various research works that are related to our proposed work.

Arianna Dagliati et al [3] designed a machine learning prediction model for type 2 diabetes where demographic data (i.e. age, gender, time to diagnosis), clinical data from the EHR (BMI, Hba1c, lipid...
profile, smoking habit), Administrative data (anti-hypertensive therapy) are collected based on that predictive model. Form microvascular complications in the population was designed which focused on issues such as nephropathy, neuropathy and retinopathy. The model showed higher AUC values in case of SVM and Random Forest. The final model is based on regression with rebalanced classes which supports nomograms.

D. Asir Antony Gnana Singh et al [4] designed a machine learning model for prediction of diabetes. Three different types of supervised learning algorithms namely probabilistic based Naïve Bayes (NB), function based multilayer perceptron (MLP), decision tree based random forests (RF) were used. Test methods such as 10 fold cross validation (FCV), makes use of percentage split (PS) with 66% and training dataset (UTD). The preprocessing technique is used to increase the accuracy of the model. In case of pre-processing technique average accuracy for NB is increased as compared to machine learning algorithm.

K. Srinivas et al [5] developed data mining application techniques that can be used in case of health care and prediction of heart attacks. In their research, they made use of medical profiles such as blood pressure, age, blood sugar and sex and used this to predict the likelihood of getting kidney problems and heart attack.

Idemudia Christian Uwa et al [6] designed a machine learning prediction model for prediction of diabetes. On applying univariate selection method with chi squared statistical test in case of non negative feature we obtain following attributes like plasma, blood pressure, age, pedigree function.

B.M.I. Here the algorithms that were been applied are Naïve Bayes, Logistic regression, SVM, XG Boost, KNN. The dataset were of 2 types one from pima Indian diabetes dataset and the other was a Schorling dataset derived from hospital. We found that on both the dataset, naïve bayes model showed consistency and after naïve bayes logistic regression proved to be better with accuracies of 83% and 81% respectively.

V. Ranjani et al [7] emphasized on the potential use of classification based data mining techniques that includes artificial neural network (ANN), rule-based methods, Naïve Bayes and decision tree algorithm to huge volume of data of health care. In their research, medical problems have been analysed and evaluated which include blood pressure and heart disease.

M. Durairaj et al [8] demonstrates a hybrid prediction system consisting of Rough Set Theory and Artificial Neural Network for depicting medical data. This process of development of a new data mining technique and a software to help competent answers in case of analysis of medical data is explained. A hybrid tool is been proposed that incorporates RST and ANN to make efficient data analysis and indicative predictions. The experiments on spermatological data set that is been used for the prediction of excellence of animal semen. The hybrid prediction system is been applied in case of pre-processing medical database and for the purpose of training the ANN for the prediction of production. The accuracy in case of prediction is obtained in case of comparison that is been made between the observed and predicted cleavage rate.

S.M. Hasan Mahmud et al [30] designed a machine learning model for prediction of diabetes where the comparison is been based on the performance evaluation by 10-fold validation technique. A framework is also been generated for diabetes prediction, monitoring and application (DPMA). Here the basic concept is that multiple machine learning classifiers are supposed to perform better than a single machine learning classifier.

Akm Ashiquzzaman, Abdul Kawser Tushar et al
designed a diabetes prediction model by making use of the application of the dropout method. Novel form of deep neural network for the prognosis of diabetes with increased accuracy is been discussed.

Ioannis Kavakiotis et al [32] designed machine learning and data mining approaches that were applied on all the aspects of DM research and that were applied on biomarker identification and prediction diagnosis.

Muhammad Azeem Sarwar, Nasir Kamal et al [33] designed a model for the prediction of diabetes by making use of various machine learning techniques where the data is divided into training data and testing data. Enthought Canopy method is used and the result is been obtained. SVM and KNN have shown higher accuracy in the model.

2.1. Literature review:-
<table>
<thead>
<tr>
<th>S.No.</th>
<th>AUTHOR</th>
<th>DESCRIPTION/WORK</th>
<th>TITLE</th>
<th>METHODS/TOOLS</th>
<th>RESEARCH GAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Arianna Dagliati, Simone Marini, Lucia Sacchi, Giulia Cogni</td>
<td>Data mining and computational methods are adopted to derive patient specific information to predict outcome of interest.</td>
<td>Machine learning methods to predict diabetes complications</td>
<td>SVM, Random Forest, Logistic Regression</td>
<td>Development of predictive model for the onset of microvascular complications in case of T2DM.</td>
</tr>
<tr>
<td>2.</td>
<td>Dr. D. Asir Antony Gnana Singh, Dr. E. Jebamalarleavline, B. Shahawaz Baig</td>
<td>Supervised learning algorithm is used for the diagnosis and prediction of diabetes and the accuracy is increased by pre-processing technique.</td>
<td>Diabetes prediction using medical data</td>
<td>Probabilistic based naïve Bayes (NB), function based multilayer perceptron (MLP), decision tree based random forest (RF)</td>
<td>The pre-processing increased the accuracy of all the models that include naïve Bayes, MLP, RF except the 10-fold cross validation method.</td>
</tr>
<tr>
<td>3.</td>
<td>K. Srinivas, B. Kavihta Rani, Dr. A. Govrdhan</td>
<td>An effective approach for the extraction of significant patterns has been established, on the basis of calculated weight, the value greater than the threshold is chosen.</td>
<td>Applications of data mining techniques in healthcare and prediction of heart attacks.</td>
<td>Naïve Bayes, Artificial Neural Network, Decision Tree</td>
<td>The role of text mining can be expanded to widen its role in case of unstructured data.</td>
</tr>
<tr>
<td>4.</td>
<td>Idemudia Christian Uwa, Nehikhare Efehi</td>
<td>Data mining processes that are been used in case of medical diagnosis and the usage of various machine learning techniques for predicting diabetes.</td>
<td>Evaluating the performance of machine learning algorithms for diagnosing diabetes in individuals</td>
<td>Logistic regression, naïve bayes, support vector machine, XGBoost, kNN</td>
<td>To gather new data and fine-tuning techniques to be used, means of handling imbalanced class data can be explored.</td>
</tr>
<tr>
<td>5.</td>
<td>M. Durairaj, V. Ranjani</td>
<td>Combination of more than one data mining technique for diabetes prediction yielding better accuracy comparatively.</td>
<td>Data mining applications in healthcare sector: A Study</td>
<td>Rough Set, Artificial neural network, ANN and Hybrid Technique</td>
<td>Hybrid techniques when applied for various diseases can yield better accuracy.</td>
</tr>
<tr>
<td></td>
<td>Author(s)</td>
<td>Title</td>
<td>Methodology</td>
<td>Citation</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>-----------------------------------</td>
<td>----------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>6.</td>
<td>M. Durairaj, K. Meena</td>
<td>Hybridization of two ML techniques such as ANN and RST is used as an alternative to the conventional methods for the prediction.</td>
<td>A hybrid prediction system using rough sets and artificial neural networks.</td>
<td>Rough Set Theory (RST), Artificial Neural Network (ANN)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Incorporation of biological information, systematic comparison of different machine learning algorithms, hybridization of rough sets and neural network ensembles to build predictors for improving performance.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.</td>
<td>SM Hasan Mahmud, Md Altab Hossin, Md. Razu Ahmed</td>
<td>A framework is designed for real-time diabetes prediction, monitoring, and application. An optimized and efficient machine learning application is developed which could predict diabetes. Different classification criteria are used for investigating the performance of different classification techniques.</td>
<td>Machine learning based unified framework for diabetes prediction.</td>
<td>10-fold validation technique along with Naïve Bayes, ANN, Logistic Regression, Decision Tree, Random Forest, and SVM</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mobile based application for prediction of diabetes based on multiple machine learning classifiers that would perform better than a single learning classifier.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Performance increase of predictive model's accuracy can have better predictive scores which will pave the way for breakthrough in health prognostication.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
A review on prediction of diabetes type 2 by machine learning techniques


A systematic review of the applications of machine learning, datamining techniques and tools with respect to diagnosis, prediction, genetic background, and environment, and healthcare and management has been done based on different machine learning techniques.


With the advent of bio-technology, and with the huge amount of data produced, and with the ever-increasing amount of EHRs, the diagnosis, and treatment of diseases can be enriched.


Dataset of patient’s record is obtained and various machine learning algorithms are applied and based on that accuracy and prediction is done.


The advancement can be made in terms of applying various techniques such as big data, cloud computing with machine learning tools.

3 Machine Learning Algorithms:-

The significance of machine learning algorithms depends in the development of models that is based on the existing data and consequently, classification or prediction by making use of novel data. Machine learning methods have been widely used in various applications in diversified domains like system biology, genomics. Specifically speaking, supervised machine learning techniques have been finding immense importance in a number of bioinformatics prediction techniques. The aim here is to showcase an overview of the machine learning algorithms as well as application methods based on the same.

<table>
<thead>
<tr>
<th>Machine learning</th>
<th>Supervised learning</th>
<th>Reinforcement learning</th>
<th>Unsupervised learning</th>
<th>Classification</th>
<th>Regression</th>
<th>Clustering</th>
<th>Dimensionality reduction</th>
</tr>
</thead>
</table>

3.3 Supervised learning:

Supervised learning has the involvement of supervisor which works in the same way as a teacher in real life. It

Figure 2: Types of machine learning algorithms
is such a type of learning in which we teach or train a machine by making use of data which has already been tagged with the correct answer (Paul Akangah et al., 2018) [9].

Further, we experiment the machine with new sets of data so that the supervised learning algorithm can analyze the training data and generate a correct outcome on the basis of the previous labeled data (R. Sathya et al., 2013) [10].

Supervised learning is classified into two categories:

- **Regression**: If the output variable includes a real value, for example “dollars” or “weight” then we call it a regression problem. We see the use of regression algorithm while dealing with decision trees, linear regression, logistic regression etc (Giovanni Grano et al., 2018) [11].
- **Classification**: If the output variable includes a category, for example “Red” or “blue” or “disease” and “no disease” then we call it a classification. We see the use of classification algorithm while dealing with naive bayes classifier, support vector machine, K- Nearest Neighbour (Kondi Srujan Kumar et al., 2019) [12].

3.1.1. **Artificial Neural Network**:

This algorithm is conceptualized on the basis of biological neurons. We can see that in case of biological learning process the process of learning is thought to be based on minor adjustments to the synaptic connections between neurons whereas in ANN the learning process is totally based on the interconnections between the processing elements which combine into a network topology.

Basicallly, ANN consists of 3 layers i.e. input layer, hidden layer, and the output layer. We see that in case of ANN, the training of hidden layer containing network and makes use of its connected structures for the purpose of pattern recognition and classification. In case of bioinformatics applications of ANN, we employ different types of architectures with perceptron and multiple layered perceptron being the simplest in this category.

Radial basis function networks and Kohonen self-organizing maps are also found useful. The major steps which are involved in ANN algorithm are as follows:

- By processing available data, training and test datasets are generated.
- Data is encoded into digital format by making use of encoding systems, such as binary systems.
- ANN architecture is designed and developed by making use of 3 layers for the purpose of prediction.
- ANN is trained by making use of appropriate input data and parameters.
- ANN model is such selected which gives the valid output.
- ANN model is thus validated by using test datasets for the purpose of estimation of efficacy for prediction.

The biggest advantage we observe in case of ANNs is its ability to analyze and process large complex datasets, having non-linear relationships. This model includes more benefits like having the ability to handle noisy data and the caliber of generalization. The limitation of the method observed is in the amount of time that would be taken in case of processing complex datasets. ANN has extensively been used in case of gene prediction, sequence feature analysis etc.

3.1.2. **Support Vector Machine**

Support Vector Machine is a supervised learning method that is based on statistical learning theory. For linearly separable illustrations, SVM creates a maximum margin hyper-plane that separates the data points into 2 different classes. The hyper-plane works as a decision surface between two classes (Affsan Abbrar et al., 2018) [22]. In case of non-linearly separable data, firstly SVM changes data into higher dimensional feature space and consequently makes use of a linear maximum margin hyper-plane. This leads to the introduction of computational intractability that requires transformation into a higher dimensional space (An T. Nguyen et al., 2018) [23]. SVM resolves this by defining most appropriate kernel functions by the help of which the computations can be taken into consideration in the original space itself. The three popular kernel functions that are used generally are linear, polynomial and radial basis function (Sandra Vieira et al., 2019) [24]. In case of bioinformatics, we see many domain specific kernel functions such as graph.
A review on prediction of diabetes type 2 by machine learning techniques

kernel,string kernel (Jesse H. Krijthe et al, 2017) [25]. This concept can also be used in case of multiclass classification. The two most common multiclass classification methods that find their use here are the one against all and one against one (Konstantinos Sechidis et al, 2017) [26]. The steps that are employed in SVM algorithm are given below:

- Feature vector is constructed in order to represent positive and negative dataset; this feature vector contains properties of the input data that could be amino acid, physiochemical properties etc.
- Appropriate kernel function is chosen so as to fit for the prediction task by making use of classifier training.
- The model is selected with best performance to make predictions.
- The application of chosen model for doing predictions on the unknown input data set, the most robust classifier is SVM, it has the best generalization ability in case of unseen data in comparison to other methods.

SVM is the most commonly used machine learning method that is used in case of computational biology and bioinformatics. It is also been used for secondary structure prediction, gene finding, fold recognition as well as binding site prediction.

Support vector machine is a distinguishing classifier which is previously defined by excluding hyper-plane which means, on the given labeled training data, here supervised learning, the algorithm gives output in the form of a hyper-plane which will categorize new examples. The hyper-plane is a line which divides a plane into two parts, in case of the two dimensional space where each of the class lie on the either side.

Figure 3

Figure 2, in the above example diagram “b” shows that a line in this case separates the two different classes as depicted in example “a”. Here we use the equation of line as \( y = ax \). We may also use the following \( y = mx + c \).

Figure 4

Figure 3, in the above example we see different property of SVM where we are making use of \( z^2 = x^2 + y^2 \), here by making use of square values we can separate “c” as “d.

3.1.3. K-Nearest Neighbor

KNN classifier finds the k nearest examples in the reference set, and considering majority vote
Assignment of classes in decision boundaries are implicitly derived in case of KNN. Below are the relevant steps that are involved in the development of KNN classifier:

- Featureset is constructed and a distance metric is used to compute distances between features.
- Number of nearest neighbors is determined for the training set.
- Euclidean distances or any other distance measure such as Mahalanobis distance are calculated between the query instance and the training samples.
- Distances are sorted and nearest neighbors are determined on the basis of the k-th minimum distance.
- Class label is predicted in case of new or unknown instance by making use of the class label of nearest neighbors.

The most significant advantage of KNN method is that it has higher efficiency on large datasets and robustness while processing noisy data (Gopi Battineni et al, 2019) [19]. The drawback of KNN is its high computation cost, which deduces its speed. In case of bioinformatics, we observe that KNN model is been employed successfully (Yun-lei Cai et al, 2010) [20].

### 3.1.4 Decision Tree

Decision tree are considered to be a branch test based classifier. The construction of the same involves the analysis of the set of training examples, class labels are known for them. New and unseen examples are classified by this information. A leaf node symbolizes a specific class and every branch represents a group of classes (Mikolas Janota et al, 2018) [21]. A test on a single attribute value is been represented by the decision node, with its one main branch and the subsequent classes are represented as possible outcomes (Sullivan hue et al, 2018) [29]. The major steps that are to be considered in case of decision tree algorithm is given below:

- Training dataset is prepared in such an appropriate form in case of the classifier by the method of feature extraction from input data.
- Decision tree is constructed by putting the instances in training set at the initial node.
- The instances are divided into two distinguishable classes i.e. child nodes based on their chosen test value.
- By the recursive application of the last step it is checked that the fulfillment of termination or pre-pruning condition is met.
- The resultant tree is pruned with its applications for performing predictions. Decision trees are simple classifiers and hence have better interpretability as compared to other machine learning methods. They are widely used in bioinformatics for predicting genetic interactions and related applications.

### 3.1.5 Random Forests

Random Forests is a group of randomly created independent classifiers and decision trees (Paul Akangah et al, 2018) [9]. It generally depicts substantial performance improvements over single tree classifiers such as C4.5, CART. Randomness or feasibility can be introduced in the RF algorithm in two ways:

1. **Bootstrapping**. A bootstraps set is created from the original training data set by making use of random sampling by doing replacement to generate each tree (Marcus Muller et al, 2018) [28].

2. **Node Splitting**: Here the selection of subset of attributes is carried out. On splitting a node, where there are M input attributes, then the number ‘m’, where m<<M and is been specified in such a way that at each node, m attributes are randomly selected and the best split is considered on them. A value that is good of ‘m’ is by default selected by making use of various implementations, considering ‘m’ as sqrt (M) for the very purpose of classification. On the basis of the CART algorithm the classification tree is induced by making use of ‘inbag’ data. After that an out of bag data, that is been formed after leaving out the in-bag samples from those of the original data is used in cross validation work. The steps involved in case of random forest algorithm are given below:

CART algorithm is been employed on data for the growth of random classification trees.
• Bootstrap data is been used which is known as in-bag set that is used to train the CART algorithm.
• On the basis of the best condition on a random subset of ‘m’ attributes node splitting is done.
• By making use of majority vote strategy in order to decide class affiliation in case of each OOB sample.
• Variable importance (VI) ranking, that can be used later to retrain random forest by using a smaller subset of the most relevant variables.
• Resistance to over fitting of data random forest and its variants are been applied to solve a huge amount of bioinformatics problems which includes classification of gene expression, analysis of mass spectroscopy data for diabetes prediction, sequence annotation and prediction of diabetes 2 mellitus.

3.1.6. Ensemble Classifiers

In case of ensemble classifiers, the individual decisions in case of a set of classifiers are rejoined with weighted or un-weighted voting for the purpose of classification of new instances.

Ensemble classifiers are also called as multi-classifier systems. These classifiers are found to be inefficient in prediction tasks because of the fact that they find use of a combined classifier and cannot even be captured by making use of any single model alone. These methods are been applied in different bioinformatics problems because of their high prediction accuracy.

3.1.7. Unsupervised learning:

Unsupervised learning is that type of training in machine where we make use of information that is neither labeled nor is classified and so it lets the algorithm to work on this information without any prior guidance as in case of supervised learning [Nagdev Amruthnath et al, 2018] [13]. The task of the machine here is to group unsorted information into patterns or on the basis of differences and similarities without the prior training being done on the data [Memoona Khanam et al, 2015] [14].

Unsupervised learning is classified into two categories:

• Association: Dimensionality reduction is the other name of association rule learning problem. An association learning problem is one where one needs to find rules that could be applied to large data sets that may include for example people who wish to buy A and also intend to buy B.

• Clustering: A clustering problem is one where we want to find the inherent groupings within the data, which includes grouping various customers by their purchasing behavior [Oyelade et al 2010] [15].

We find use of clustering algorithm while dealing with K-means, meanshift, K-medoids.

3.1.8. Artificial Hidden Markov Models (HMM)

Hidden Markov Models have found their use in every popular machine learning approaches such as in case of bioinformatics. They are probabilistic models that are generally implied in time series and linear sequences. It can be used to describe the evolution of those events which are observable and these depend on internal factors, which themselves are not observable. Here, we see the observed events are recalled as a symbol and the invisible factors that are underlying the observations that are referred to as a state. An HMM comprises of several states, that are connected by means of transition probabilities, which leads to the formation of a Markov process. Every state here has an observable symbol that is been attached to it. An HMM comprises of visible process with observable events and a hidden process which includes internal states with their movement in tandem. The goal here is to find the optimal path from the states, which leads to maximization of the occurrence of observed sequence of symbols. Therelevant steps that associated in the algorithm for the generation of HMM are given below:

• HMM architecture is been developed by making use of various states which ultimately represent the given set of features.
• Assignment is been done of the hidden states to the features and so is the construction of the HMM is done.
• The HMM is thus trained using supervised technique or unsupervised technique inorder to let the model sufficiently fit the problem that is under study.
Emission probabilities are derived that influence the distribution of observed symbols, which implies that the probability of a symbol being observed provided that HMM is in a specific state.

HMM is decoded for the prediction of hidden states from the data.

The benefits associated with HMMs are the ease of their use, need of smaller datasets and precise comprehension of the process.

Among the major drawbacks associated with HMMs is their higher computational cost. HMMs are found to be most effective in case of biological sequence analysis and so they are periodically applied for multiple sequence alignments, gene finding, etc.

### 3.1.9 K-Means clustering

The k mean clustering algorithm provides a generalized method to implement approximate solution. The reason why k mean clustering algorithm is very popular is because of the ease and simplicity. Kmean can be considered to be a gradient descent procedure, where the initiation in the algorithm is done at starting cluster centroids and it iteratively decreases the objective function. The convergence of the k mean generally takes place at the local minimum. It basically performs the updation work unless the local minimum is found. The problem to find the global minimum is NP-complete. The time complexity of the k-means clustering algorithm is O(nkl) where, the required number of clusters is denoted by “k”, the total number of objects in the dataset is denoted by “n” and the number of iterations is denoted by “I”, k≤n, I≤n.

![Figure 5: Diagram depicting K-means clustering](image)

**Reinforcement learning**

Reinforcement learning belongs to that area of Machine Learning where the actions are taken purely to achieve maximize rewards in a specific situation. It can be used on different types of machines and even on software for finding the best path possible or behavior it is supposed to take in any specific situation (Jiachi Xie et al 15) [16]. It distinguishes itself from supervised learning in a way that in case of supervised learning the training data has the answer key with it and the model is trained with the correct answer by its own on the other hand, in case of reinforcement learning, answer key is not available but here we can see that the reinforcement agent decides what is to be done in order to perform the given task (Nicolas Bougie et al, 2019) [17]. In the absence of training data set, it is bound to learn from its own previous experiences.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Advantages</th>
<th>Drawbacks</th>
</tr>
</thead>
</table>
| ARTIFICIAL NEURALNETWORK      | • Good performance in case of nonlinear relationships.  
• Capacity to handle noisy data.            | • Computational burden is greater.  
• Over fitting is a problem.                      |
| HIDDEN MARKOV MODEL           | • Precise comprehension of background process.  
• Easy to use and is powerful.                   | • Intensive computation  
• Slower than other methods.  
• Pronetoover-fitting.                             |
A review on prediction of diabetes type 2 by machine learning techniques

| SUPPORTVECTORMACHINE | • Best generalization ability is provided.  
• Robust to noisy database.  
• Susceptibility is less too fitting. | • Computation is expensive in some cases such as incase of non-linearly separable problems. |
|-----------------------|-------------------------------------------------|-----------------------------------------------------------------|
| K-NEARESTNEIGHBOR     | • Simple and easy to learn.  
• It is found efficient when the training data is large. | • Complexity in computation  
• As the number of attributes gets increased. |
|                       | • Fast training. | performance becomes inconsistent. |
| DECISIONTREE          | • Capability to handle both continuous and discrete attributes.  
• Interpretability is better.  
• Results are better in cases of redundant attributes. | • In the presence of large number of classes, the data prone to errors.  
• Sensitivity towards small variations in data. |
| RANDOM FOREST         | • High speed and accuracy.  
• Less prone towards over-fitting.  
• Able to evaluate every attribute for prediction. | • Tendency of over-fitting in case of noisy data. |
| ENSEMBLECLASSIFIERS   | • Incase of prediction, greater efficiency is obtained.  
• Utilization of data is more. | • Computational complexity is more. |

4. Machine Learning Advancements in diabetes prediction:

Machine learning can be used in case of digital diagnosis of any disease. It can detect patterns of certain diseases and help in providing a broader perspective.

4.1. Diabots:

It is found that this chatbot is capable of interacting with patients seamlessly based on the symptoms. There are many generic text-to-text diabot i.e., diagnostic chatbot which makes use of Natural Language Understanding (NLU) for the providing personalized prediction by making use of a generalized health dataset and also on the basis of various symptoms sought from the patient.

4.2. Oncology:

Here the researchers are making use of deep learning techniques for the purpose of training the algorithm and to make it recognize carcinogenic tissue (but at the same time it is taken into consideration that the blood sugar level is normal) at such a level that is comparable to even physicians.

4.3. Better Radiotherapy:

As the machine learning algorithms have the potential to learn from the multitude of various samples that are available, finding an effective to diagnose and find the variables if any. The example includes Google’s DeepMind Health which is assisting the healthcare professionals to distinguish between the healthy and unhealthy people. Here the advancement is been made in terms of diagnosing eye damage done by various diseases which includes diabetes too.

4.4. Outbreak Prediction:

Machine learning is used in monitoring and predicting epidemics around the globe. ANN can be used to collect information from different websites and predict information from dengue outbreak to severe chronic infectious diseases. This can also assist in knowing the world wide increase in diabetes patients round the globe which led to the conclusion that India is the diabetic capital of the world.
4.5. Crowd sourced Data Collection:

Crowd sourcing has helped researchers and practitioners to get access to huge amount of information that are been uploaded by people based on their consent. This helps in collected data that is been collected by the consent of the patients and is assisting in the research.

The various applications available for the prediction of diabetes includes Diabetik by UglyApps, DiabetesinCheckby Everyday Health, iCookbook Diabetic by Publications International, mySugr Junior by mySugr GmbH, HealthyOut by Healthy Out and many more.

5. Conclusion/ Future work:--

The applications of machine learning could be applied for the diagnosis of various diseases, their symptoms, their cause, their treatment. The sudden deaths occurring due to kidney failure, heart attack, strokes etc. accompanied with diabetes can be prevented through early treatment and diagnosis. In the study we saw various algorithms such as SVM, decision tree, KNN, naïve bayes, etc. making their use in the prediction of incidence of diabetes. The classification techniques gave different results when applied to different datasets. We found that various classification techniques are useful for different data sets. The variation in the model performance can be noticed for different datasets and the cause could be predicted accordingly.

Future study can be focused on acquiring new dataset that would lead to new insight and knowledge to improving the prediction of diabetes using machine learning techniques. Based on the parameters like age, body mass index, obesity level, history of chronic disease, etc. when accompanied by various machine learning techniques will lead to better prediction levels. The new dimension which is extending is using deep learning which when assisted with machine learning can give tremendous results in terms of pattern recognition and better predicted values.

References

Hang Lai, Huaxiong Huang, “Predicting models for diabetes mellites using machine learning techniques, BMC publication, 2019


Arianna Dagliati,” Machine Learning methods to predict diabetes complications”, Journal of diabetes Science and technology, 2018


Idemudia Christian Uwa, Nehikhare Efêhî, “Evaluating the performance of machine learning algorithms for diagnosing diabetes in individuals” International Journal of Science and Research (IJSR), 2018


Paul Akangah, Leotis Parrish, Andrea Nana Ofori-Boadu, Francis Davis,”Predicting academic achievement in fundamentals of thermodynamics using supervised learning technique” American Society For Engineering Education (ASEE) Southeastern Section conference, 2018


Giovanni Grano, Timofey V. Titov, Sebastiano Panichella, Harald C. Gall,” How High Will It Be? Using Machine Learning Models to Predict Branch Coverage in Automated Testing” MALTESQUE@ SANER, 2018


Nagdev Amruthnath, Tarun Gupta, “ A research study on Unsupervised Machine Learning Algorithms for Fault detection in Predictive Maintenance” Research Gate, 2018


Oyelade, Oladipupo, Obagbuwa, ”Application of k-means clustering algorithm for prediction of students’ academic performance” IJCSIS 2010

Jiachi Xie, Chelsea M. Myers, Jichen Zhu, “Interactive Visualizer to facilitate Game
A review on prediction of diabetes type 2 by machine learning techniques

Designers in Understanding Machine learning” CHI 2019
Nicolas Bougie, Ryuntaro Ichise,” Skill- based curiosity for intrinsically motivated reinforcement learning,” Springer 2019
Gopi Battineni, Getu Gamo Sagro, Chintalapudi Nalini, Francesco Amenta, Seyed Khosrow Tayebati,”Comparative machine learning approach: follow up study on type 2 diabetes prediction by cross validation methods”, 2019
Yun-lei Cai, Duo Ji .Dong-feng Cai , “A kNN research paper classification method based on shared nearest neighbor” NTCIR-8 Workshop Meeting, 2010
Mikolas Janota, “Towards generalization in QBF Solving via Machine Learning” AAAI-I, 2018
Afsan Abbrar, Dr. Pranam Paul,”Development of Application to Recognise Hand Written Digit at rum time” International Journal of Innovative Research in Computer and Communication Engineering, 2018
An T. Nguyen, Matthew Lease, Byron C. Wallace, ”Mash: software tools for developing interactive and transparent machine learning systems, IUI Workshops 2019
Paul Akangah, Leotis Parrish, Andrea Nana Ofori-Boadu, Francis Davis,”Predicting academic achievement in fundamentals of thermodynamics using supervised learning technique” American Society For Engineering Education (ASEE) Southeastern Section conference,2018
Sullivan hue, Christophe Hurlin, Sessi Tokpavi,” Machine Learning for Credit Scoring: Improving Logistic Regression with Non Linear Decision Tree Effects” Research Gate, 2018