EJ Publications
International Journal Research in Applied Engineering, Science and Technology (IJRAEST) Impact Factor: 4.537(SJIF)
An International Peer-Reviewed Journal Vol-2, Issue-3, 2020
www.ijraest.com Indexed in: Google Scholar, Academia, Cite Factor ISSN: 2582-029X

RESEARCH ARTICLE

METHODS FOR DETECTION OF DIABETES MELLITUS USING MACHINE LEARNING TECHNIQUES

Dr Jyoti[1]
Assistant Professor (CE)
YMCA University of Science & Technology
Sector 6, Mathura Road, Faridabad, India

Peri Arjun[2]
JC Bose YMCA University of Science and Technology,
Faridabad, India

Abstract:
Diabetes mellitus is a metabolic condition that is more and more severe as the body becomes unable to metabolize glucose. This study aimed to evaluate different models for sensitivity and selectivity to better identify patients at risk of diabetes mellitus based on demographic data of the patients and laboratory results during visits to medical centers. One of the norm and developing strategies of good Classification and reorganization approaches focused on recursive learning is machine learning. Machine learning allows a classification method for artificial intelligence to be educated and evaluated. Machine learning assisted the detection of diseases with the correct preparation and test case. We study various methods used by researchers to detect Diabetes using machine learning principles like SVM, KNN, Decision tree and Ensemble approach.

Keywords: Diabetes, Machine learning, SVM, Decision tree.

Introduction

Machine learning techniques offer an excellent way to train a certain system for prediction and classification of anything through training. Machine learning involves learning structures from the supplied data. In recent years machine learning has become quite effective in design, development and training of models for prediction of diseases. Machine learning has gained more attention in medical fields because of reduced processing time and less user contact, reducing resources for health treatment. Diabetes is a persistent condition (life-long). Diabetes is caused by human body's inability to produce insulin to regulate glucose upon ingestion of food. Lack to ability to produce sufficient

EJ Publications
International Journal Research in Applied Engineering, Science and Technology (IJRAEST) Impact Factor: 4.537(SJIF)
An International Peer-Reviewed Journal Vol-2, Issue-3, 2020
www.ijraest.com Indexed in: Google Scholar, Academia, Cite Factor ISSN: 2582-029X

RESEARCH ARTICLE

amount of insulin and insulin tolerance are the primary cause of Diabetes. Chronic Diabetes induces many health problems.

Machine learning is one of the main ways to predict or find an underlying mechanism. The process or algorithm which provides intelligent results by recognizing intricate patterns is, therefore, the main focus of machine training. Models or predictions thought to be characteristics of the underlying data mechanism. A learner can use examples (data) to capture interest characteristics of their unknown underlying distribution of probability. Data can be seen as examples of possible relationships between the variables observed. Machine learning has designed different patterns to make intelligent decisions on the input data. The main challenge for machine-learning is the behavior of inputs that should be trained in the examples observed. Therefore, they are trained for efficient and sensitive output with all possible inputs.

The next chapters include an introduction to various methods used for machine learning techniques for diabetes detection and their advantages and disadvantages with a comprehensive literature survey of multiple researchers with accuracy achieved by them.

Machine learning techniques for Diabetes detection:

It should be mentioned that SVM rises as the most successful algorithm in both biological and clinical datasets in Data Mining. Many articles (~ 85%) used the supervised learning approaches, i.e. in classification and regression tasks. In the remaining 15%, association rules were employed mainly to study associations between biomarkers. More specifically, concerning the part dealing with the evaluation task, in all reported research reports, the identified subsets of biomarkers (features) were evaluated through appropriate procedures, such as splitting the dataset into train and test set or via cross-validation. By analogy, the same approaches have been followed in Diabetes prediction. Worth emphasizes that in many studies, after the feature/biomarker selection, researchers have performed comparative analysis on different machine learning algorithms to assess their predictive performance and finally choose the most efficient one(s). To this end, this should be the baseline of practice in every study to be carried out, considering that several characteristics of the dataset, such as dimensionality, low number of instances compared to number of features or even the type of the dataset itself (genetic or clinical), can affect significantly the performance of the algorithm. Hence, an algorithm with the best performance in one dataset could easily have lower prediction accuracy compared to other algorithms in different datasets.

EJ Publications
International Journal Research in Applied Engineering, Science and Technology (IJRAEST) Impact Factor: 4.537(SJIF)
An International Peer-Reviewed Journal Vol-2, Issue-3, 2020
www.ijraest.com Indexed in: Google Scholar, Academia, Cite Factor ISSN: 2582-029X

RESEARCH ARTICLE

Support vector machine (SVM)

This is a controlled learning technique which means that the data set is trained to achieve the predetermined output. It displays the data collection as cloud points in space. The goal is to create a hyper plane separating data sets into different categories. The hyper plane splits the data collection into groups such that data analysis and Classification can be easily carried out. This hyper plane will be as long as the different divisions are concerned. Nevertheless, if the groups in which the data collection is categorized are broad, advanced kernel configuration techniques are used. In addition, [1] used Decision Tree, SVM, and Naive Bayes classifiers to detect Diabetes. The aim was to identify the classifier with the highest accuracy. The Pima Indian dataset was used for this study. The partition of the dataset is done by means of 10-folds cross-validation. The authors didn't discuss the data preprocessing. The performance was evaluated using the measures of accuracy, precision, recall, and F-measure. The highest accuracy was obtained by the Naïve Bayes, which reached 76.30%. In addition to the other studies, Negi and Jaiswal [2] aimed to apply the SVM to predict Diabetes. The Pima Indians and Diabetes 130-US datasets were used as a combined dataset. By using a combined dataset, the diabetes prediction might be more reliable, with an accuracy of 72%.

Advantages of SVM

1. Works well with unstructured and semi-structured datasets such as images and text.
2. Can attain accurate and robust results.
3. Is successfully used in medical applications.

Disadvantages of SVM

1. It requires long training time when it is used with large datasets.
2. It is hard sometimes to select the right kernel function.
3. The weights of the variables are difficult to interpret in the final model.

Related work	Advantages	Disadvantages
FCM and SVM and testing it on a set of PIDD.[20]	FCM and SVM gives good Classification	Better machine learning algorithm should be employed along with them.
Combination	Fuzzy C-	Real-time data

EJ Publications
International Journal Research in Applied Engineering, Science and Technology (IJRAEST) Impact Factor: 4.537(SJIF)
An International Peer-Reviewed Journal Vol-2, Issue-3, 2020
www.ijraest.com Indexed in: Google Scholar, Academia, Cite Factor ISSN: 2582-029X

RESEARCH ARTICLE

of fuzzy c-means and SVM is used for diabetes prediction on dataset[21]	means classify data set in better way as it involves membership function.	is noisy, so the effort is required to make it useable for processing
LDA–MWSVM[22]	The system performs feature extraction and reduction using the Linear Discriminant Analysis (LDA) method, followed by Classification using the Morlet Wavelet Support Vector Machine (MWSVM) classifier.	Accuracy can be improved further.

Decision tree

Decision Tree is a supervised method used to solve classification problems. The key purpose of using the Decision Tree is used to estimate the goal class using previously applied decisions. It uses prediction and classification nodes and internodes. Root nodes identify instances with different characteristics. Root nodes may have two or three divisions, and the leaf nodes are graded.

In every stage, the Decision tree chooses each node by evaluating the highest information gain among all the attributes [3]. Decision trees build a classification and regression tree model in the

EJ Publications
International Journal Research in Applied Engineering, Science and Technology (IJRAEST) Impact Factor: 4.537(SJIF)
An International Peer-Reviewed Journal Vol-2, Issue-3, 2020
www.ijraest.com Indexed in: Google Scholar, Academia, Cite Factor ISSN: 2582-029X

RESEARCH ARTICLE

form of tree structure by breaking data set into smaller subsets and simultaneously developing the associated decision tree. The decision tree is a top-down structure with one root

node, and it is splitting its branches, which have a parent-child relationship. The tree includes a root node, some leaf nodes representing any classes, and internal nodes representing test condition.[4][5] predicted a modified J48 Classification Algorithm for the Prediction of Diabetes.

In the case of nephropathy, Huang et al. employed a Decision Tree-based prediction tool that combines genetic and clinical features to identify diabetic nephropathy in patients with T2D [25]. Leung et al. compared several machine learning methods: partial least square regression, Classification and regression tree, the C5.0 Decision Tree, Random Forest, naïve Bayes, neural networks, and support vector machines [26]. The dataset used consists of both genetic (Single Nucleotide Polymorphisms — SNPs) and clinical data. Age, age of diagnosis, systolic blood pressure, and genetic polymorphisms of uteroglobin and lipid metabolism arose as the most efficient predictors.

Table 1. Related work in Decision tree algorithm:

Ref	Technique	Result	Dataset
Ref[6]	J48	Sensitivity: 0.89, Specificity: 0.91	Private Dataset (Collected Manually)
Ref[7]	Standalone J48	Accuracy: 81%	CPCSSN Database
Ref[8]	Classification and Regression Trees	Accuracy: 92%	Pima Indian Dataset

EJ Publications
International Journal Research in Applied Engineering, Science and Technology (IJRAEST) Impact Factor: 4.537(SJIF)
An International Peer-Reviewed Journal Vol-2, Issue-3, 2020
www.ijraest.com Indexed in: Google Scholar, Academia, Cite Factor ISSN: 2582-029X

RESEARCH ARTICLE

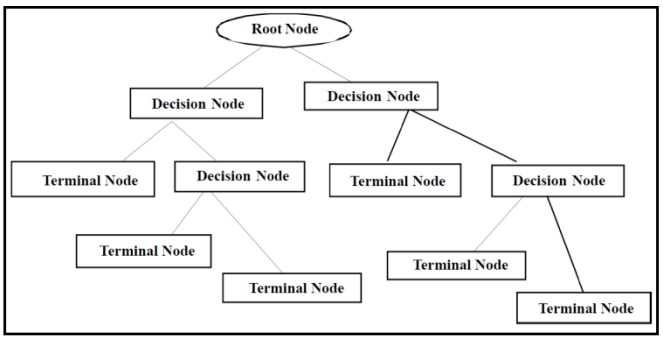


Figure 1. Sample Decision tree

Types of Decision tree classifiers

AD Tree generates an alternating decision tree for two-class problems using an optimized induction and heuristic search methods to speed up learning, [34].

- J48 uses a pruned or unpruned C4.5 decision tree, [35].
- NBTree generates a decision tree with naive Bayes classifiers at the leaves, [36].
- RandomTree constructs a tree with randomly chosen attributes at each node without pruning, [37].
- REPTree builds a decision tree using information gain and prunes it using reduced-error pruning with backfitting, [37].
- SimpleCart creates a tree and implements minimal cost complexity pruning, [38].

Advantages of Decision tree

1. Ability to handle attribute with different costs.
2. Ability to handle missing values in attributes.
3. Ability to handle both continuous and discrete attributes.
4. Ability to prune trees after creation in an attempt to remove branches that are not helpful and replacing them with leaf nodes

Disadvantages of Decision tree

1. Decision trees are also prone to errors in Classification, owing to differences in perceptions and the limitations of applying statistical tools.
2. Complexity
3. Too much information

EJ Publications
International Journal Research in Applied Engineering, Science and Technology (IJRAEST) Impact Factor: 4.537(SJIF)
An International Peer-Reviewed Journal Vol-2, Issue-3, 2020
www.ijraest.com Indexed in: Google Scholar, Academia, Cite Factor ISSN: 2582-029X

RESEARCH ARTICLE

Related work	Advantages	Disadvantages
Use of a rule extraction algorithm, ReRX with J48 graft, combined with sampling selection techniques (sampling Re-RX with J48 graft) is done.[23]	High accuracy in terms of rule extraction.	The diagnosis of T2DM remains a complex problem; diagnosis
It presents an approach using principal component analysis and modified Gini index based fuzzy SLIQ decision tree algorithm. [24]	Sharp decision boundary can be overcome by fuzzy SLIQ.	Accuracy can be improved further by better fuzzy membership.
Decision Tree-based prediction tool that combines[25] both genetic and clinical features	To identify diabetic nephropathy in patients with T2D	It does not perform well with large datasets.

EJ Publications
International Journal Research in Applied Engineering, Science and Technology (IJRAEST) Impact Factor: 4.537(SJIF)
An International Peer-Reviewed Journal Vol-2, Issue-3, 2020
www.ijraest.com Indexed in: Google Scholar, Academia, Cite Factor ISSN: 2582-029X

RESEARCH ARTICLE

K- Nearest Neighbor Algorithm (KNN)

KNN is a method which is used for classifying objects based on closest training examples in the feature space. KNN is the most basic type of instance-based learning or lazy learning. It assumes all instances are points in n-dimensional space. A distance measure is needed to determine the "closeness" of instances. KNN classifies an instance by finding its nearest neighbors and picking the most popular class among the neighbors.

Features of KNN

a) All instances of the data correspond to the points in an n-dimensional Euclidean space
b) Classification is delayed till a new instance arrives
c) In KNN, the Classification is done by comparing feature vectors of the different points in a space region.
d) The target function may be discrete or real valued.
In KNN, the training samples are mainly described by n-dimensional numeric attributes. The training samples are stored in an n dimensional space. When a test sample (unknown class label) is given, k-nearest neighbor classifier starts searching the 'k' training samples which are closest to the unknown sample or test sample. Closeness is mainly defined in terms of Euclidean distance.

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \quad \text{.....(1)}$$

Advantages of KNN

1. It is very simple algorithm to understand and interpret.
2. It is very useful for nonlinear data because there is no assumption about data in this algorithm.
3. It is a versatile algorithm as we can use it for Classification as well as regression.
4. It has relatively high accuracy but there are much better supervised learning models than KNN.

Disadvantages of KNN

1. It is computationally a bit expensive algorithm because it stores all the training data.
2. High memory storage required as compared to other supervised learning algorithms.
3. Prediction is slow in case of big N.
4. It is very sensitive to the scale of data as well as irrelevant features.

EJ Publications
International Journal Research in Applied Engineering, Science and Technology (IJRAEST) Impact Factor: 4.537(SJIF)
An International Peer-Reviewed Journal Vol-2, Issue-3, 2020
www.ijraest.com Indexed in: Google Scholar, Academia, Cite Factor ISSN: 2582-029X

RESEARCH ARTICLE

Related work	Advantages	Disadvantages
In this study C4.5, Neural Network, Kmeans, Visualization is used to detect Diabetes.[25]	It is good approach as hybrid method is used.	prediction, Classification, visualisation requires tremendous effort
Artificial neural network combined with fuzzy logic is used to detect diabetes[25]	It allows better result as fuzzy accounts for uncertainties also.	Extracting rules from existing methods is not very efficient as it takes times.

For noisy training data and complex goal functions, KNN is a highly powerful inductive inference tool. The target function can be described as a combination of less complex local approximations for a whole space. KNN Learning is very simple and it takes time to classify.

Table 2. Related work in KNN algorithm

Ref	Technique	Result	Dataset
Ref[18]	K-means and KNN	Accuracy: 97.0%	Pima Indians Diabetes Dataset
Ref[18]	Simple KNN	Accuracy: 73.17%	Pima Indians Diabetes

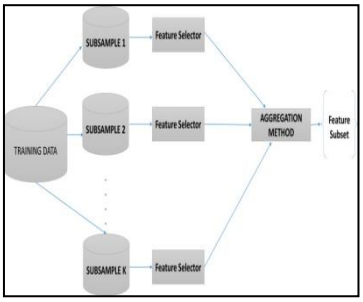
EJ Publications
International Journal Research in Applied Engineering, Science and Technology (IJRAEST) Impact Factor: 4.537(SJIF)
An International Peer-Reviewed Journal Vol-2, Issue-3, 2020
www.ijraest.com Indexed in: Google Scholar, Academia, Cite Factor ISSN: 2582-029X

RESEARCH ARTICLE

			Dataset
Ref[11]	Kernel-Based Adaptive Filtering Algorithm	The CGM signals of a random subject are used to assess the prediction accuracy.	Private Dataset

Ensemble method

Ensemble is a Machine Learning technique whose methods are meta-algorithms that combine several machine learning techniques into one optimal predictive model in order to reduce variance, bias or improve predictions. This approach enables improved predictive performance when compared to that of a single model. There are various assembling methods such as bagging, boosting, ads-boosting, stacking, voting, averaging, etc. When construct the classification model, the data used to construct model may have noise or imbalanced information. To improve classification accuracy, the ensemble methods were introduced. We can combine multiple models that lead to bias and variance reductions. The ensemble methods such as bagging and boosting have been presented [29]. They can comply with individual base classifier. Researchers have accomplished the use of ensemble methods. For instance, [30] proposed the effectiveness of the bagging predictor by comparing statistical tests of 12 bagging classifiers for each medical dataset. The results revealed that bagging with a decision tree performs well on the extremely imbalance and high dimensional large datasets. [31] Examined ensemble methods with decision tree classifier based on imprecision probabilities and uncertainty measures. The results show that boosting is an excellent method to combine with a decision tree.



EJ Publications
International Journal Research in Applied Engineering, Science and Technology (IJRAEST) Impact Factor: 4.537(SJIF)
An International Peer-Reviewed Journal Vol-2, Issue-3, 2020
www.ijraest.com Indexed in: Google Scholar, Academia, Cite Factor ISSN: 2582-029X

RESEARCH ARTICLE

Figure 2. Feature selection in an ensemble approach:

The impetus behind the entire approach to learning was recently applied to other computer learning areas, such as the collection of apps. The aim is then to produce more reliable performance than a single function selection approach by integrating the outputs of different feature selection models. However, are not only many versions usable, as is the case for classification ensembles, but also the various subsets of features obtained. Function collection sets can be categorized according to a number of parameters regarding one or more of the above, but the easiest distinction applies to the form of selectors used. The Ensemble is known as homogenous if the basis selectors are all of the same nature; otherwise, the Ensemble is heterogeneous. Experiments on a large and diverse collection of UCI data sets have demonstrated that MultiBoost achieves higher accuracy significantly more often than Bagging or AdaBoost[39]. A comprehensive collection of tests have established that Decorate consistently creates ensembles more accurate than the base classifier, Bagging, Random Forests, which are also more accurate than Boosting on small training sets and are comparable to Boosting on larger training sets. [40].

Bagging

Bagging stands for Bootstrap Aggregating, it is a method to diminish the variance of prediction by generating extra data for training from the original dataset [32]. To improve classification accuracy and unstable classification problems [7]. Over fitting is avoided by bagging method. This is a simple method to understand if the quantity is a descriptive statistic such as a mean or a Standard Deviation (SD).

Boosting

Boosting is a two-step approach. The boosting method uses subsets of the original data to generate a series of averagely performing models. As, the name suggests boosting means it "boosts" the performance by combining them using a particular cost function. Boosting method will create a strong classifier from many different weak classifiers. In general, this method works by building a model from the training data and then creating a second model that attempts to correct the errors from the first model[32].

Table 3. Comparison between Bagging and Boosting

EJ Publications
International Journal Research in Applied Engineering, Science and Technology (IJRAEST) Impact Factor: 4.537(SJIF)
An International Peer-Reviewed Journal Vol-2, Issue-3, 2020
www.ijraest.com Indexed in: Google Scholar, Academia, Cite Factor ISSN: 2582-029X

RESEARCH ARTICLE

	Bagging	Boosting
Example	Random Forest	Gradient Boosting Method
Bias/Variance Trade Off	Decreases Variance	Decreases Bias
Training Data	Partition before training	Adaptive data weighting
Base Learner	Complex	Simple
Classifier Strength	Tries to overfit with a flexible model and then average to decrease variance.	Tries to underfit using weak learners and then improve model based on classifier performance.

Ada-boosting:

Ada Boost was the first really successful boosting algorithm developed for binary Classification. It is the best starting point for understanding boosting. Ada Boost is best used to boost the performance of decision trees on binary classification problems. Ada Boost assigns a weight, which determines the probability that this observation will appear in the training set. Observations with higher weights are more likely to be included in the training set. Hence, Ada Boost tends to assign higher weights to those observations which have been misclassified, so that they will represent a larger part of the next classifiers training set, with the aim that, this time, the next classifier trained will perform better on them.

Stacking

Stacking is an ensemble machine learning algorithm that learns how to best combine the predictions from multiple well-performing machine learning models. The benefit of stacking is that it can harness the capabilities of a range of well-performing models on a classification or regression task and make predictions that have better performance than any single model in the Ensemble. It uses a meta-learning algorithm to learn how to best combine the predictions from two or more base machine learning algorithms [32].

Voting

Voting is used for Classification. Every model makes a prediction (votes) for each test instance and the final output prediction is the one that receives more than half of the votes. If none of the predictions get more than half of the votes, we may say that the ensemble method could not make a

EJ Publications
International Journal Research in Applied Engineering, Science and Technology (IJRAEST) Impact Factor: 4.537(SJIF)
An International Peer-Reviewed Journal Vol-2, Issue-3, 2020
www.ijraest.com Indexed in: Google Scholar, Academia, Cite Factor ISSN: 2582-029X

RESEARCH ARTICLE

stable prediction for this instance. Although this is a widely used technique, you may try the most voted prediction (even if that is less than half of the votes) as the final prediction.

Averaging

Averaging is used for regression. In simple averaging method, for every instance of test dataset, the average predictions are calculated. This method often reduces overfit and creates a smoother regression model.

Related wok	Advantages	Disadvantages
Decision support system is proposed that uses AdaBoost algorithm with Decision Stump as base classifier for Classification.[27]	Adaboost gives an edge to yield combined and better results.	Accuracy of classifiers needs to be improved with nn classifiers and other approaches
Homogeneous ensemble method uses the same type of base learner in each iteration.[28]	Adaboost and Stacking Classifier to be the best out of all the five classifiers in the aspects of accuracy, since they give better accuracy	A particular method to identify Diabetes is not very sophisticated way for initial diabetes detection.

EJ Publications
International Journal Research in Applied Engineering, Science and Technology (IJRAEST) Impact Factor: 4.537(SJIF)
An International Peer-Reviewed Journal Vol-2, Issue-3, 2020
www.ijraest.com Indexed in: Google Scholar, Academia, Cite Factor ISSN: 2582-029X

RESEARCH ARTICLE

Table 4. Related work in Ensemble methods

Ref	Technique	Result	Dataset
Ref[7]	Adaboost Ensemble using J48	Accuracy: 81%	CPCSSN Database
Ref[12]	Ensemble Perception	Accuracy: 0.75	NHANES0910

Measurement

Evaluation parameters in machine learning for above approaches are

1. Accuracy
2. Precision
3. Recall
4. F-Measure

TABLE I. EVALUATION PARAMETERS

S.No	Metrics	Formula	Evaluation focus
1	Accuracy	(TP+TN) / (TP+TN+FP+FN)	Measures the ratio of correct predictions over the total number of instances evaluated
4	Precision	TP / (TP+FP)	Measure the positive patterns that are correctly predicted from the total predicted patterns in a positive class.
5	Recall	TP / (TP+TN)	Measure the fraction of positive patterns
6	F-Measure	F2 * (precision * recall) / (precision + recall)	Represents the harmonic mean between recall and precision values

Where true positive represents (TP) the number of identified positive samples in the positive set.

EJ Publications
International Journal Research in Applied Engineering, Science and Technology (IJRAEST) Impact Factor: 4.537(SJIF)
An International Peer-Reviewed Journal Vol-2, Issue-3, 2020
www.ijraest.com Indexed in: Google Scholar, Academia, Cite Factor ISSN: 2582-029X

RESEARCH ARTICLE

True negative (TP) means the number of Classification negative samples in the negative set. False positive (FP) is the number of the number of identified positive samples in the negative set. And false negative (FN) represents the number of identified negative samples in the positive set. It is often used to evaluate the quality of classification models. The accuracy is defined as the ratio of the number of samples correctly classified by the classifier to the total number of samples. In medical statistics, there are two basic characteristics, sensitivity (SN) and specificity (SP). Sensitivity is the true positive rate, and specificity is the true negative rate.

Related works

AUTHOR	DATASET	METHOD	RESULTS
Ref[13]	OMAN DIABETES DATASET	SVM	94%
Ref[14]	CHNS DATASET	SVM	94.2%
Ref[15]	EHR REPOSITORY	SVM, KNN, Decision Tree and Random forest.	95%
Ref[16]	Unknown	Ensemble and Random forest.	95.19%
Ref[17]	17 Medical Dataset Including Pima Indian Diabetes Dataset	Sequential Minimal Optimization (SMO), Support Vector Machine (SVM) and	78.21%

EJ Publications
International Journal Research in Applied Engineering, Science and Technology (IJRAEST) Impact Factor: 4.537(SJIF)
An International Peer-Reviewed Journal Vol-2, Issue-3, 2020
www.ijraest.com Indexed in: Google Scholar, Academia, Cite Factor ISSN: 2582-029X

RESEARCH ARTICLE

		Elephant Herding Optimizer	
Ref[18]	Electronic Health Records	Unsupervised Deep Learning Neural Network (Deep Patient) Area Under the ROC Curve (AUC-ROC)	ROC:0.91
Ref[19]	Pima Indians Diabetes Dataset	Support Vector Machine (SVM) and Neural Network (NN)	96.09%

Conclusion:

Although a vast variety of work has accrued in developing strategies for forecasting diabetes, most of these approaches utilize conventional mathematical techniques. Machine learning methods are gaining momentum and the community's attention. Researchers are excited about testing out various styles of classifiers and designing new models to increase the precision of the diagnosis of Diabetes. The same dream has been pursued in this paper to achieve good prediction accuracy. Both classifications of Machine Learning (ML) and Deep Learning (DL) used over the last six years have been investigated about their frequency of usage and accuracy. On the PID data collection, ML

EJ Publications
International Journal Research in Applied Engineering, Science and Technology (IJRAEST) Impact Factor: 4.537(SJIF)
An International Peer-Reviewed Journal Vol-2, Issue-3, 2020
www.ijraest.com Indexed in: Google Scholar, Academia, Cite Factor ISSN: 2582-029X

classifiers of one or zero frequency have been introduced to allow suggestions for their use. The accuracy obtained by these ML techniques was 68%–74%. The average accuracy obtained by researchers for DL algorithms was 95%. In the future, unused classificatory can be applied to other datasets in a combined model to further improve the accuracy of diabetes prediction.

References:

[1] Sisodia, D.; Sisodia, D.S. Prediction of Diabetes using Classification Algorithms. Procedia Comput. Sci. 2018,132, 1578–1585.
[2] Iyer, A., S, J., Sumbaly, R., 2015. Diagnosis of Diabetes Using Classification Mining Techniques. International Journal of Data Mining & Knowledge Management Process 5, 1–14. doi:10.5121/ijdkp.2015.5101, arXiv:1502.03774.
[3] Quinlan J. R. (1996b). Improved use of continuous attributes in C4.5. *J. Artif. Intell. Res.* 4 77–90. 10.1613/jair.279
[4] L. Breiman, Bagging predictors, machine leraning,24(2),123- 140,1996.
[5] Gaganjot Kaur "Diabetes Research" Department of Computer Science and Diabetes Federation
[6] Joshi, S.; Borse, M. Detection and Prediction of Diabetes Mellitus Using Back-Propagation Neural Network. In Proceedings of the 2016 International Conference on Micro-Electronics and Telecommunication Engineering (ICMETE), Uttarpradesh, India, 22–23 September 2016; pp. 110–113.
[7] Perveen, S.; Shahbaz, M.; Guergachi, A.; Keshavjee, K. Performance Analysis of Data Mining Classification Techniques to Predict Diabetes. Procedia Comput. Sci. 2016, 82, 115–121.
[8] Soltani, Z.; Jafarian, A. A New Artificial Neural Networks Approach for Diagnosing Diabetes Disease Type II. Int. J. Adv. Comput. Sci. Appl. 2016, 7, 89–94.
[9] Sun, X.; Yu, X.; Liu, J.;Wang, H. Glucose prediction for type 1 diabetes using KLMS algorithm. In Proceedings of the 2017 36th Chinese Control Conference (CCC), Liaoning, China, 26–28 July 2017; pp. 1124–1128.
[10] Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus Nahla H. Barakat, Andrew P. Bradley, Senior Member, IEEE, and Mohamed Nabil H. Barakat
[11] Han L, Luo S, Yu J, Pan L, Chen S. Rule extraction from support vector machines using ensemble learning approach: an application for diagnosis of Diabetes. IEEE J Biomed Health Inform. 2015;19(2):728-734. doi:10.1109/JBHI.2014.2325615

EJ Publications
International Journal Research in Applied Engineering, Science and Technology (IJRAEST) Impact Factor: 4.537(SJIF)
An International Peer-Reviewed Journal Vol-2, Issue-3, 2020
www.ijraest.com Indexed in: Google Scholar, Academia, Cite Factor ISSN: 2582-029X

RESEARCH ARTICLE

[12] Mirshahvalad, R.; Zanjani, N.A. Diabetes prediction using Ensemble perceptron algorithm. In Proceedings of the 2017 9th International Conference on Computational Intelligence and Communication Networks (CICN),Girne, Cyprus, 16–17 September 2017; pp. 190–194.

[13] Tao Zheng, Wei Xie, Liling Xu, Xiaoying He, Ya Zhang, Mingrong You, Gong Yang, You Chen, A machine learning-based framework to identify type 2 diabetes through electronic health records, International Journal of Medical Informatics, Volume 97,2017,Pages 120-127,ISSN 1386-5056.

[14] Pham, T.; Tran, T.; Phung, D.; Venkatesh, S. Predicting healthcare trajectories from medical records: A deep learning approach. J. Biomed. Inform. 2017, 69, 218–229.

[15] Rao, N.M.; Kannan, K.; Gao, X.Z.; Roy, D.S. Novel classifiers for intelligent disease diagnosis with multi-objective parameter evolution. Comput. Electr. Eng. 2018, 67, 483–496.

[16] Miotto, R.; Li, L.; Kidd, B.A.; Dudley, J.T. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. Sci. Rep. 2016, 6, 26094.

[17] NirmalaDevi, M.; Alias Balamurugan, S.A.; Swathi, U.V. An amalgam KNN to predict diabetes mellitus. In Proceedings of the 2013 IEEE International Conference ON Emerging Trends in Computing, Communication and Nanotechnology (ICECCN), Tirunelveli, India, 25–26 March 2013; pp. 691–695.

[18] Gill, N.S.; Mittal, P. A computational hybrid model with two level classification using SVM and neural network for predicting the diabetes disease. J. Theor. Appl. Inf. Technol. 2016, 87, 1–10.

[19] Ashiquzzaman, A.; Kawsar Tushar, A.; Rashedul Islam, M.D.; Shon, D.; Kichang, L.M.; Jeong-Ho, P.; Dong-Sun, L.; Jongmyon, K. Reduction of overfitting in diabetes prediction using deep learning neural network. In IT Convergence and Security; Lecture Notes in Electrical Engineering; Springer: Singapore, 2017; Volume 449.

[20] Sanakal, Ravi, and T. Jayakumari. "Prognosis of diabetes using data mining approach-fuzzy C means clustering and support vector machine." Int. J. Comput. Trends Technol.(IJCTT) 11.2 (2014): 94-98

[21] Polat, Kemal, and Salih Güneş. "An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease." Digital Signal Processing 17.4 (2007): 702-710.

[22] Çalisir D., Dogantekin E. An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine Classifier. Expert Syst. Appl. 2011;38(7):8311–8315.

[23] Kamadi, VSRP Varma, Appa Rao Allam, and Sita Mahalakshmi Thummala. "A computational intelligence technique for the effective diagnosis of diabetic patients using

principal component analysis (PCA) and modified fuzzy SLIQ decision tree approach." Applied Soft Computing 49 (2016): 137-145.

[24] Hayashi, Yoichi, and Shonosuke Yukita. "Rule extraction using Recursive-Rule extraction algorithm with J48graft combined with sampling selection techniques for the diagnosis of type 2 diabetes mellitus in the Pima Indian dataset." Informatics in Medicine Unlocked 2 (2016): 92-104.

[25] Huang G.-M., Huang K.-Y., Lee T.-Y., Weng J. An interpretable rule-based diagnostic classification of diabetic nephropathy among type 2 diabetes patients. BMC Bioinforma. 2015;16(S-1):S5.

[26] Leung R.K., Wang Y., Ma R.C., Luk A.O., Lam V., Ng M. 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. Jul 23 2013;14:162.

[27] Vijayan, V. Veena, and C. Anjali. "Prediction and diagnosis of diabetes mellitus—A machine learning approach." Intelligent Computational Systems (RAICS), 2015 IEEE Recent Advances in. IEEE, 2015.

[28] Ms. Komal Patil , Dr. S. D. Sawarkar , Mrs. Swati Narwane, 2019, Designing a Model to Detect Diabetes using Machine Learning, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 08, Issue 11 (November 2019),

[29] N. Hosseinpour, S. Setayeshi, K. Ansari-asl and M. Mosleh, Diabetes Diagnosis by Using Computational Intelligence Algorithms, Journal of Advanced Research in Computer Science and Software Engineering, 2(12), 71-77, (2012).

[30] G. Liang, and C. Zhang, Empirical Study of Bagging Predictors on Medical Data, 9th Australasian Data Mining Conference, 121, 31-40, (2011)

[31] J. Abellán, Ensemble of decision tree based on imprecise probabilities an uncertainty measures, Information Fusion, 14,423-430, (2013).

[32] L. Breiman, Bagging Predictors, Machine Learning, 24(2), 123-140, (1996)

[33] I. Syarif, E. Zaluska, A. Prugel-Bennett and G. Wills, Application of Bagging, Boosting and Stacking to Intrusion Detection, MLDM2012, LNAI7376, 513-602, (2012)

[34] Y. Freund and L. Mason, “The alternating decision tree learning algorithm,” in Proc. 16th Internat. Conf. Machine Learning, 1999, pp. 124–133.

[35] R. Quinlan, C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann, 1993.

[36] R. Kohavi, “Scaling up the accuracy of Naive-Bayes classifiers: A Decision-Tree hybrid,” in Second International Conference on Knoledge Discovery and Data Mining, 1996, pp. 202–207.

EJ Publications
International Journal Research in Applied Engineering, Science and Technology (IJRAEST) Impact Factor: 4.537(SJIF)
An International Peer-Reviewed Journal Vol-2, Issue-3, 2020
www.ijraest.com Indexed in: Google Scholar, Academia, Cite Factor ISSN: 2582-029X

[37] I. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques. Amsterdam: Elsevier/Morgan Kaufman, 2005.
[38] L. Breiman, J. Friedman, R. Olshen, and C. Stone, Classification and Regression Trees. Belmont, California: Wadsworth International Group, 1984.
[39] E. Bauer and R. Kohavi, "An empirical comparison of voting classification algorithms: Bagging, boosting, and variants," Machine Learning, vol. 36, pp. 105–139, 1999.
[40] G. Webb, "Multiboosting: A technique for combining boosting and wagging," Machine Learning, vol. 40, pp. 159 – 196, 2000.

AUTHORS PROFILE

Dr. Jyoti Verma is currently working as Assistant Professor in the Department of Computer Science in J.C. Bose University of Science and Technology. She completed her Ph.D in the area of Information Retrieval in 2011. She has over 15 years of teaching experience with almost 10 years of research experience. She has over 30 publications in her name. her current areas of interest are Information Retrieval and Big Data Analytics.

Dr Jyoti
Assistant Professor(CE)
YMCA University of Science & Technology
Sector 6, Mathura Road, Faridabad - 121006
Phone: +91-9910341139

Peri Arjun is student at JC Bose YMCA University of Science and Technology, Faridabad where he is currently pursuing Master of Technology course in Computer Science with specialization in Computer Networking. His research interest lies in the area of Machine Learning and Artificial Intelligence with special focus on healthcare domain. He completed his Bachelor of Technology in Computer Science and Engineering with first division.

Peri Arjun
periarjun@gmail.com
+91-9871303495