Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Review Paper: Predicting Diabetes by Machine learning Algorithm

Suhail S.M.Alqrinawi, M.A.Burhanuddin, Lizawati Salahuddin

Universiti Teknikal Malaysia Melaka, Fakulti Teknologi Maklumat dan Komunikasi (FTMK), Melaka, Malaysia.
Email: suhailalqrinawi@gmail.com

One element of artificial intelligence called machine learning enables the creation of computer systems with the capacity to learn from past experiences without the need for extensive programming. In the current environment, machine learning is absolutely necessary to minimize human work and create higher levels of automation with reduced errors and it will give accuracy result for predictive of status. Arthritis, Asthma, Heart disease, chronic kidney disease, and Diabetes, in this paper we will focus in Diabetes illness, whereas the Diabetes defined as the disability of body to produce the Harmon insulin, abnormal metabolism of carbohydrates and glucose level on the blood, the seven classification machine learning techniques are examined to predict the diabetes in early stage, on other hand the accuracy result will affect in patient life.

Keywords: Diabetes, Artificial Intelligence, Machine learning Techniques.

1. Introduction

The huge advancement and rapid progress in technology techniques tied up with both of healthcare department and patients' satisfaction in health care area, The end-user opinion plays a significant responsibility in the measurement of performance satisfaction in any sector, So the company or organization looking for the achievement of key performance indicators with a high percentage for end-user. In the healthcare department, Patient satisfaction is measured by the quality and speed of service which the patients received like (speed of treatment, decision making from the doctor, delay time, data set which evaluated and categorized in which department), moreover, the speed of treatment in healthcare depends on data or patient health records which play a major role in the speeding of treatment (Alelaiwi, 2019).

2. Machine Learning

Machine learning algorithms has been recently playing a more significant part in scientific research, also it is a particular method of artificial intelligence it collects information from training data, also we can define Machine learning as a modern and highly sophisticated

technological application become a huge trend in the industry (Chauhan et al., 2021). moreover, machine learning plays a crucial role in many fields like finance, healthcare, and insecurity. Machine learning has four categories of algorithms:

A. Supervised learning

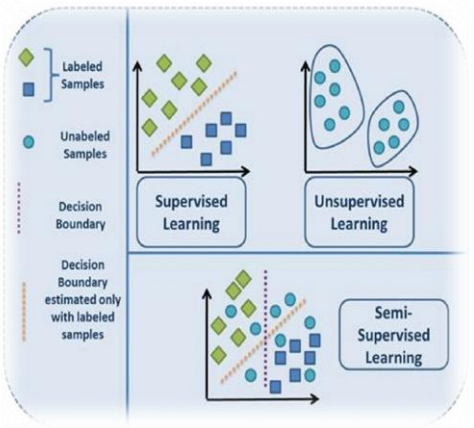
Several machine learning algorithms use the effective method of supervised learning. Major components of the machine learning method are regression and classification, which are based on labelled data obtained from diverse sources and formats. The regression technique makes use of the idea of statistical prediction of data to look for the most effective correlation between two variables. Regression and classification differ significantly in this regard (Kushwaha and Kumaresan, 2021), Classification is the process of breaking up datasets into smaller pieces based on their categories and characteristics. we can set the target data and compare the results to the target output. If everything checks out, we don't make any changes; if not, we repeat the process until the results are unsatisfactory(Rathore and Mannepalli, 2021).

B. Unsupervised learning

Unsupervised learning involves training the algorithm with unlabeled data. In the end, it aids in pattern detection while describing a model. In this scenario, input data aids the computer in identifying pertinent patterns and mining out important laws and regulations. Important points are eventually merged and further summarized to give meaning and better comprehension (Dalal, 2020).

C. Semi-Supervised learning

Semi-Supervised Learning tackles this issue by learning from both Labeled and Unlabeled data which come out from supervised and unsupervised type, the next figure illustrate the meaning of it.



Semi-Supervised learning(Chebli, Djebbar and Marouani, 2019) .

D. Reinforcement learning

Reinforcement learning carries out a specific task to reach the goal of having the model automatically learn from a dynamic environment in addition to the provided dataset. These

algorithms offer the optimal answer, and computer programmers are used to gain access to the dynamic environment in order to accomplish a certain task. Feedback is also given in the form of rewards and restrictions to help the model(Dalal, 2020).

We use the first type of machine learning (Supervised learning) the output data must relate to input by training and developing an exercise (Winter, 2019).

The huge information about the patients founded, So it is impossible processing by the human without any external intervention, that why the machine learning provide algorithms to predict the future outcome, moreover there are many types in the healthcare sector like(clinical, sensor data, genomic data omics data, Transcriptomic data and proteomic) (Islam et al., 2020).

Depending on the data and the information to be extracted from the data, various machine learning algorithms for analysis and preprocessing are used. Support vector machines are used for analyzing the ECG and EEG data for diabetes classification (Site, Nurmi and Lohan, 2021).

3. Chronic disease

Continues sharply spreading of chronic disease be one of the leading ones that caused death and danger for humans and health, The umbrella team of chronic diseases includes many diseases like Cancer, Alzheimer's, Arthritis, Asthma, Heart disease, chronic kidney disease, liver disease, and Diabetes. as a result of these diseases, the lifestyle will change to an abnormality new style with more restrictions, As the permanent diabetes patients number increased rapidly, Diabetes is defined as the lack of response to treatment and making insulin inside the human body, also when the sugar in the blood is too high diabetes will be found in the patient life (Mahedy Hasan et al., 2020).

A. Chronic Diabetes Disease

One of the most hazardous chronic illnesses that might cause other major, complicated diseases is diabetes. Diabetic mellitus, which refers to a group of metabolic illnesses, is another name for diabetes diseases. Heart attack, blindness, kidney disease, and many other illnesses can develop as a result of diabetes. One of the deadliest diseases, diabetes, commonly known as diabetes mellitus, is a chronic condition (Chauhan et al., 2021). Unstoppable increase of Diabetes makes it one of the more important diseases which need to predict and to know the ways to try to avoid it, Diabetes has three types: the first type's defined as when the pancreas can't produce enough amount of insulin, the second definition comes when the body can't respond to dealing with insulin effectively, and the last one is pregnancy diabetes is due to insulin-blocking hormones produced during pregnancy. This type of diabetes only occurs during pregnancy, highlighted prominently in the information from the World Health Organization (WHO), In 2019 the main reason for 1.5 million death occurred by diabetes whereas the percent of 4.8% for those with age is less than 70 years (Mahedy Hasan et al., 2020).

The pancreases operation focus on producing the main hormone which is called insulin, this hormone will affect the cell absorption blood sugar operation from foods, lack of insulin in the body will be called Diabetes, in these days Diabetes is still one of the most diseases that lead to death and will change the life to uncomfortable life style, increasing or decreasing the

sugar in the blood will play an opposite role in the parts operation in the human body, However, the existence of too much sugar level in the blood is known as Hyperglycemia, The main reason for accruing this type of situation is one of both, the first reason has come when the pancreas does not correctly produce insulin and the second main one is the response operation of the body to the insulin correctly, many problems will be accrued by this type of diabetes for example(hyperglycemia can cause vomiting, excessive hunger and thirst, rapid heartbeat, vision problems, and other symptoms), according to International Diabetes Federation (IDA) in 2035 number of diabetes patients will arrive at 592 million all over the world, the machine learning techniques play a critical impact on predicting diabetes, Undoubtedly, the advance and update of progress technology connect to the healthcare, for this reason, the doctors said the early stage of prediction will impact recovery with a high percent of achievement of treatment with new techniques will be supported by Artificial Intelligence (Yahyaoui et al., 2019).

Diabetes is the main reason causes of kidney disease and heart attack and blindness, poor knowledge on management the treatment and early prediction of diabetes will lead to blindness by Clogged arteries and branches of the ophthalmic artery, on other hand, Diabetes will lead to Diabetic Retinopathy(DR) and this is the main reason of blindness, according to WHO the 17% percent of blindness in the United State of America is come out from DR where the percent on China is less than USA percent of blindness causing with 7% percent, in addition, the etiology percent of blindness on Brazil is 7.6% of populations whom suffering of this disease(Alves et al., 2020).

Nowadays Many people claim that Diabetes is very dangerous and destroy the internal part of the human body, but in fact, the two most popular reasons for diabetes is the pancreas product level of insulin and the other is sugar in the blood, however, diabetes is a non-communicable disease and the people who have suffered from it can live with stable life just if they follow the instruction of chronic disease doctors on diabetes control management (Tanvir Islam et al., 2019).

4. The selected Classification Algorithm in Chronic disease predictions

A. Decision Tree(DT)

DT is a supervised algorithm that uses a tree-like model to assess decisions, potential implications, and outcomes. Each branch represents a conclusion, and each node is a query. E leaf nodes define class labels. When a sample data reaches a leaf node, the matching node's title will be assigned to the sample. This strategy works well when the problem is a simple and small dataset. Even though the technique is simple to grasp, it has flaws like overfitting and biased results when working with unbalanced datasets. However, DT can map both linear and nonlinear relationships (Jayatilake & Ganegoda, 2021).

B. K-Nearest Neighbor (KNN)

KNN is a well-known supervised classification technique that is utilized in a variety of applications, including pattern recognition and intrusion detection.

KNN is a simple and straightforward algorithm. Even though KNN has a high accuracy, it is computationally expensive and requires a lot of memory because both testing and training data must be saved (Jayatilake & Ganegoda, 2021).

C. Linear Discriminant Analysis (LDA)

principal component analysis PCA is a linear dimensionality reduction approach that can be used to condense a large number of variables into a smaller number that retains the majority of the original data. It looks for a linear combination of variables from which to extract the most variation. After removing this variance, PCA looks for a second linear combination that iteratively explains the most significant proportion of the remaining variation. The principal axis approach produces orthogonal (uncorrelated) factors in this case. It also entails computing the eigenvalues and eigenvectors of covariance matrices, sorting these eigenvectors in descending order of their eigenvalues, and then projecting the actual data into the directions of the sorted eigenvectors (Ricciardi et al., 2020).

D. Naïve Bayes (NB)

Naïve Bayes is similar to SVM, but it uses statistical approaches in the process. The NB method is a classification technique for binary and multiclass situations. e NB classifiers are a collection of Bayes theorem-based classification algorithms. However, they all follow the same rule: every pair of traits being classed must be independent of one another. When a new input is added, the probabilistic value for that input is calculated among the classes. The data is labelled with the type with the highest probabilistic value(Jayatilake & Ganegoda, 2021).

E. Fuzzy Logic(FL)

Fuzzy logic techniques were employed to make the predictions. Fuzzy set theory is the basis for these strategies. Between the numbers o and 1, fuzzy values are utilized. Neural networks, language processing, expert systems, and other engineering applications(Jayatilake & Ganegoda, 2021).

F. Rule Induction(RI)

Rule induction is one of the most effective machine learning approaches.

Rule induction is one of the essential tools of Data Mining since regularities concealed in data are commonly described in terms of rules, some rule induction system more complex rules, in which value of attributes may be expressed by negation of some value subset of the attribute domain. Data from which rules are induced usually presented in a form similar to a table in which cases (or example) labels (or names) for rows and variables are labeled as attributes and decision. This research restricts our attention to rule induction that belongs to supervised learning: all cases are pre-classified by an expert. In other words, the decision value is assigned by an expert to each instance. Attributes are independent variables, and the decision is a dependent variable(Grzymala-Busse, J. W. (2005). Rule induction. In Data mining and knowledge discovery handbook (pp. 277-294). Springer, Boston, n.d.).

G. Support Vector Machine (SVM)

SVM is asupervised machine learning technique mainly used to solve classification problems, but it may also be used to solve regression problems. The data items are first shown as points

in an n-dimensional space, with the feature value representing the specific coordinate. It determines the hyperplane that divides the data points into two classes in English. The minimal distance between the decision hyperplane and instances at the boundary can be maximized (Björklund, Björklund and Martens, 2021).

5. Related Works with Machine Learning

A. Waiting Phlebotomy Unit

Developed an artificial neural network as a classification software tool to predict the patients waiting time at Phlebotomy Unit, and this tool helps the patients to expect the waiting time while it needs some necessary data like pregnant, polyclinics, and elderly, all of these statuses significant impact in the waiting time (Javadifard et al., 2019).

B. Heart disease prediction

The early prediction of heart disease plays an essential role in the control and managing the disease and will give the patients a chance to control and deal with suffering efficiently; machine learning is tied up with predicting and discovering heart disease early by using two applicable machine learning algorithm which is known as Hybrid grid search and Random search the patients will get the approximately result to predict the early stage of heart disease (Katarya and Srinivas, 2020).

C. Diabetes prediction

The health record is vital to collect the data about any diabetes patient; moreover, getting the previous result and test status will help to predict the future test status, using six machine learning techniques to predict diabetes and these techniques named (NB, KNN, SVM, LR, DT, and RF), in addition, the researcher used Indian data set and divided these data into two-part, the first 70% for training the machine and the other 30% for a test, the outcome of this research come out with the accuracy of using this machine learning techniques, The result of accuracy for (NB, KNN, SVM, LR, DT, and RF) will be (74%,77%,74%,71%,77%,77%) respectively(Sonar and Jaya Malini, 2019).

D. Kidney prediction

Predicting chronic kidney disease in the early stage plays a vital role in saving the life, keeping the kidney working well, and managing the treatment with new patients' kidney issues status. in addition, chronic disease is tied up with machine learning. All Machine learning techniques play a crucial role in this predicting process, but the linear support vector machine (LSVM) comes out with the highest accuracy of 98.86% (Chittora et al., 2021).

E. Liver disease

Support Vector Machine (SVM)with 99.76% accuracy is the best one on the list of methods (KNN, BP, SVM, and NBC) to predict the liver disease (Shaheamlung, Kaur and Singla, 2019).

6. Conclusion

In this study, we studied and assessed the state-of-the-art for using machine learning approaches to predict and identify diabetes disease, as well as the many types of machine learning and their critical role in forecasting chronic disease. Diabetes is a chronic disease that must be detected early to avoid dangerous phases. We applied a number of supervised learning techniques, such as Naive Bayes, SVM, decision trees, etc.

ACKNOWLEDGEMENT

The authors would like to thank BIOCORE Research Group, Center for Advanced Computing Technology (C-ACT), Fakulti Teknologi Maklumat dan Komunikasi (FTMK), Centre for Research and Innovation Management, Universiti Teknikal Malaysia Melaka for providing all facilities and support for this research. Also, authors would like to thank to Ministry of Higher Education Malaysia for Malaysia International Scholarship(MIS) for contribution in this work.

References

1. Alves, S. S. A. et al. (2020) ‘A New strategy for the detection of diabetic retinopathy using a smartphone app and machine learning methods embedded on cloud computer’, Proceedings - IEEE Symposium on Computer-Based Medical Systems, 2020-July, pp. 542–545. doi: 10.1109/CBMS49503.2020.00108.
2. Björklund, H., Björklund, J. and Martens, W. (2021) ‘Learning algorithms’, Handbook of Automata Theory, pp. 375–409. doi: 10.4171/automata-1/11.
3. Chauhan, T. et al. (2021) ‘Supervised and Unsupervised Machine Learning based Review on Diabetes Care’, 2021 7th International Conference on Advanced Computing and Communication Systems, ICACCS 2021, pp. 581–585. doi: 10.1109/ICACCS51430.2021.9442021.
4. Chebli, A., Djebbar, A. and Marouani, H. F. (2019) ‘Semi-Supervised Learning for Medical Application: A Survey’, Proceedings of the 2018 International Conference on Applied Smart Systems, ICASS 2018, (November), pp. 24–25. doi: 10.1109/ICASS.2018.8651980.
5. Chittora, P. et al. (2021) ‘Prediction of Chronic Kidney Disease - A Machine Learning Perspective’, IEEE Access, 9, pp. 17312–17334. doi: 10.1109/ACCESS.2021.3053763.
6. Dalal, K. R. (2020) ‘Analysing the Role of Supervised and Unsupervised Machine Learning in IoT’, Proceedings of the International Conference on Electronics and Sustainable Communication Systems, ICESC 2020, (Icesc), pp. 75–79. doi: 10.1109/ICESC48915.2020.9155761.
7. Javadifard, H. et al. (2019) ‘Predicting Patient Waiting Time in Phlebotomy Units Using a Deep Learning Method’, Proceedings - 2019 Innovations in Intelligent Systems and Applications Conference, ASYU 2019, pp. 1–4. doi: 10.1109/ASYU48272.2019.8946380.
8. Katarya, R. and Srinivas, P. (2020) ‘Predicting Heart Disease at Early Stages using Machine Learning: A Survey’, Proceedings of the International Conference on Electronics and Sustainable Communication Systems, ICESC 2020, (Icesc), pp. 302–305. doi: 10.1109/ICESC48915.2020.9155586.
9. Kushwaha, P. K. and Kumaresan, M. (2021) ‘Machine learning algorithm in healthcare system: A Review’, Proceedings of International Conference on Technological Advancements and Innovations, ICTAI 2021, pp. 478–481. doi: 10.1109/ICTAI53825.2021.9673220.
10. Rathore, D. K. and Mannepalli, P. K. (2021) ‘A Review of Machine Learning Techniques and Applications for Health Care’, Proceedings of International Conference on Advances in

Technology, Management and Education, ICATME 2021, pp. 4–8. doi: 10.1109/ICATME50232.2021.9732761.

11. Shaheamlung, G., Kaur, H. and Singla, J. (2019) 'A Comprehensive Review of Medical Expert Systems for Diagnosis of Chronic Liver Diseases', Proceedings of 2019 International Conference on Computational Intelligence and Knowledge Economy, ICCIKE 2019, pp. 731–735. doi: 10.1109/ICCIKE47802.2019.9004438.
12. Sonar, P. and Jaya Malini, K. (2019) 'Diabetes prediction using different machine learning approaches', Proceedings of the 3rd International Conference on Computing Methodologies and Communication, ICCMC 2019, (Iccmc), pp. 367–371. doi: 10.1109/ICCMC.2019.8819841.
13. Yahyaoui, A. et al. (2019) 'A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques', 1st International Informatics and Software Engineering Conference: Innovative Technologies for Digital Transformation, IISEC 2019 - Proceedings, pp. 1–4. doi: 10.1109/UBMYK48245.2019.8965556.