Machine Learning Methods for Diabetes Prediction

by Nur Dzakiyullah

Submission date: 20-Jul-2022 09:34AM (UTC+0700)
Submission ID: 1872844675
File name: L29731081219_2.pdf (838.27K)
Word count: 5936
Character count: 34507

Machine Learning Methods for Diabetes Prediction

Nur Rachman Dzakiyullah, M.A. Burhanuddin, Raja Rina Raja Ikram, Khanapi Abdul Ghani, Winny Setyonugroho

Abstract*: *Machine Learning is one of the methods used for task prediction. In the diabetic's research field, the application of machine learning is emerging since the advantages of approximation on the prediction technique has significantly given insight for many health practitioners. Machine Learning is utilized in order to handle the uncontrollable risk factor by finding a relation between such a risk factor trough prediction. This study aims to review recent machine learning models that have been used in diabetes prediction with respect to the risk factors in order to prevent diabetes. This study compares the performance of the model by justified the accuracy as the baseline to evaluate the model. The result of this review shows that the Random Forest and Support Vector Machine are the most popular technique among researcher. Moreover, from this study, it can be seen that Type 2 Diabetes Mellitus (T2DM) has been a concern by researchers since the incidence of diabetes was increasing in worldwide today that happened from an uncontrollable risk factor.

Keywords: Machine Learning, Diabetes Prediction, Risk Factor, Accuracy.

I. INTRODUCTION

Diabetes Prediction (DP) is becoming one of the important topics in diabetic research. This is because there is a high incidence of Diabetes Mellitus (DM) associated with disease complications, cost of treatment and other factors that need to be controlled. DM is categorized as a Non-Communicable Disease (NCD) which is not considered contagious from one person to another. Commonly, hyperglycemia is considered one of the most significant impacts due to diabetes that affects the body in order to normalize blood glucose levels [1]. The latest estimation there are 451 million people with diabetes in 2017, and this is expected to increase to 693 million by 2045 [2]. Diabetes is a disease that changes metabolism in the body indicated by lack of resistance to insulin where the pancreas produces insulin in order to maintain the blood glucose level in the normal range. A person that having diabetes cannot remove glucose automatically from their bloodstreams thus this condition that giving an impact into serious health problems [3]. There are several types of diabetes such as Type 1 DM (T1DM), Type 2 DM (T2DM) and another type because of hyperglycemia that has been explained by the American Diabetes Association [4]. Recent trend shows that the incidence of DM is mostly due to T1DM and T2DM. In Type 1 DM (T1DM), the immune system attacks the insulin-producing pancreatic cells resulting in an absolute deficiency of insulin secretion, while Type 2 DM (T2DM) is categorized by increased resistance of the body cells to insulin, which often co-occurs with limited insulin secretion [5]. However, in order to control and prevent the incidence of diabetes, early detection can be one solution to managed diabetes. That's why diabetes prediction becoming one important research area in diabetic research.

Many researchers apply machine learning as a tool in diverse studies such as engineering, medicine, life science, and computer science especially. However, according to Harrington [6], ML is defined as a technique of turning data into meaningful information. In the term of diabetes, related diabetes information can be produced or collected through advanced sensor, digital machine, advance sequencing DNA, blood test, super-resolution digital microscopy, mass spectrometry, Magnetic Resonance Imagery (MRI) and other advanced medical tools where a huge mass production of data can be stored into Electronic Health Record (EHR). However, a lot of data still needs to be processed and analyzed in order to extract important information. ML can be utilized on large datasets related to diabetes to extract meaningful information and knowledge. Machine Learning (ML) is part of a statistic, artificial intelligence and mathematics that can be used to help the physician to make an effective clinical decision making in terms of diabetic prognosis and diagnosis. ML is a promising tool for dealing with learning ability from the data that is called "learn from the experience". As we can see in figure 1, the standard ML process in data modeling is shown below:

Revised Manuscript Received on October 05, 2019.

Nur Rachman Dzakiyullah, Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka (UTeM), Melaka, Malaysia and Faculty of Science and Technology, Department of Information Technology, Universitas 'Aisyiyah Yogyakarta (UNISA), Indonesia. Email: P031710013@student.utem.edu.my, nurrachmandzakiyullah@unisayogya.ac.id

M.A. Burhanuddin, Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka (UTeM), Melaka, Malaysia. Email: burhanuddin@utem.edu.my

Raja Rina Raja Ikram, Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka (UTeM), Melaka, Malaysia. Email: raja.rina@utem.edu.my

Khanapi Abdul Ghani, Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka (UTeM), Melaka, Malaysia. Email: khanapi@utem.edu.my

Winny Setyonugroho, Faculty of Medicine and Health Science Muhammadiyah University of Yogyakarta (UMY), Yogyakarta, Indonesia. Email: wsetyonugroho@umy.ac.id

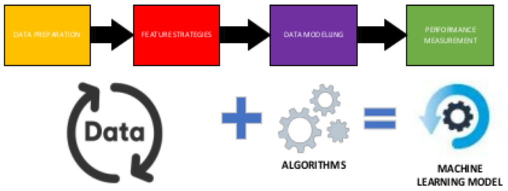

Figure 1. Machine Learning Process

However, this learning process of machine learning is typically classified into three different approaches such as supervised learning, unsupervised learning and semi-supervised learning [7], [8]. Firstly, a simple analogy, like student and teacher, supervised learning or active learning is a possibility of the learning process based on two variable called dependent and independent variable. The dependent variable defines as a target/class or label learning, then the independent variable can be explained as a predictor or attribute. In supervised learning, classification and regression techniques are two kinds of learning tasks. The different Classification and regression are the task of prediction where classification tries to predict categorical data, while regression seeks to predict numerical data. Some of the most common techniques are Decision Trees (DT), Rule Learning, and Instance-Based Learning (IBL), such as k-Nearest Neighbors (k-NN), Artificial Neural Networks (ANN), and Support Vector Machines (SVM). Second, unsupervised does not need an external teacher. It means that the process of learning depends on the structure of data or relations between variables and without any labels when training proceeds. Clustering is the technique that categorized into unsupervised learning and the ML that mostly used to do clustering, for example, K-Means, K-Medoids, Self-Organizing Map (SOM), Fuzzy C-Means, etc. Third, Semi-supervised is learning with a label and unlabeled data when training applied. Data that has classes is used to form a model (knowledge), unlabeled data is used to create boundaries between classes. Particularly, Semi-supervised learning is used when difficult to find the best feature on the data. Furthermore, another learning of ML is Reinforcement Learning. This learning type is quite different compared with standard learning as mention before which the training process is based on direct interaction with the environment through trial and failure (trial and error) without any knowledge to start while the actions produce the most significant rewards.

The advantage of these techniques is the ability process modeling of ML in the era different types of data in which can be more effective and create powerful insight when the combination based on data are extracted. According to Kavakiotis, et.al. [9] state that research in the relations between diagnosis, management, administration and other social impact give effect to the diabetes are a great concern to understand in medical science. The objective of the research in the medical field is to extend the quality of life people with diabetes when intervention is applied in many possible ways [10]. Moreover, ML are potential can be used to understand the relation according to a diabetes risk factor. In this case, as mention earlier, ML can be used to help medical doctor or physician to prevent or monitor diabetes can be more effective and efficient in term of decision making. Hence, in this effort of study, the current review of literature on diabetes prediction are presented. In this study, the rest of the paper is organized: Section 2 provides the recent publication reviewed in the study; Section 3 presents the result and discussion of the review; Section 4 finally, explain the conclusion of the review.

II. REVIEW MACHINE LEARNING METHOD FOR DIABETES PREDICTION

We conduct a literature review on Diabetes Prediction that has used a machine learning algorithm. In order to collect all the paper that related to our objective, we used the search term “Diabetes” and “Prediction” and “Machine Learning”. The database that conducted in this study such as Science Direct, PubMed, Springer link, IEEE Explore and Taylor and Francis. The limitation of this study only focuses on the performance ML model. For additional, we also consider the category of risk factor that available as a variable or attribute from the dataset or database in the study on each paper. The results of the literature review are shown in Table 1 the comparison Machine Learning Model for Diabetes Prediction application. In this study, we focus on the comparison of machine learning models based on the performance result of each model. Moreover, we have highlighted factors that are related to diabetes in the model as input or dataset and type of diabetes that are of research interest.

Based on previous studies, traditional ML and different variation of ML models are used in order to predict diabetes by achieving the best accuracy prediction of the model. Vignesh and Amalarethinam [11] proposed rule extraction algorithm for improving regular covering technique to achieve high classification accuracy. The experiment result shows an average accuracy of 85.55% for the Pima Indian diabetes (PID) dataset by 10 runs of 10-fold cross-validation. Another study [12] presents an approach for comparison of risk models to improve the prediction capacity using datasets that are provided by genome-wide association studies. Boosting algorithms have been used successfully to solve the case of genome-wide data sets due to linkage disequilibrium. Another comparison study that used genetic data has been done by [13]. The new variant Decision Tree (DT) has been proposed and compared with Random forest using Genome-wide association studies (GWASs) dataset to discover the genetic basis of complex phenotypes. [14] has proposed diabetes treatment, diagnosing, monitoring and educating patients on their medication and pragmatic technological resources for maintaining blood glucose level system through data mining technique on electronic medical records by utilizing classification technique. [15] proposed a new prognostic approach for type 2 diabetes mellitus based on electronic health records. The model was developed using random forest classifier and temporal features and feature selection. [16] developed a new SVM called WVKSVM approach by modifying the kernel matrix in order to distinguish between the true and noise variables.

Retrieval Number: L29731081219/2019©BEIESP
*DOI: 10.35940/ijitee.*L2973.*1081219*

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

The model was examined by comparing the random forest (RF) and the normal SVM, and the results show that WVKSVM has better prediction ability to improve the performance of SVM classifier when utilizing the metabolomics datasets.

The best method to diagnose T2DM has been investigated using data from Tabriz, Iran [17]. The algorithm such as support vector machine, artificial neural network, decision tree, nearest neighbors, and Bayesian network was chosen to handle diagnose of T2DM. The artificial neural network was performed to compare the better algorithm. Moreover, [18] proposed a novel ensemble method (AdaboostM1 with the random committee) to predict diabetes. The same algorithm is used to diagnose diabetes, particularly diabetes retinopathy through extracted features from heart rate (HR) [19]. Another study aims to monitor blood glucose using a sensor of continuous glucose monitoring (CGM) [20]. Classification Logistic regression model and two-step binary logistic regression model are used to analyze the feasibility of using CGM-based GV indices from impaired glucose tolerance

Table 1. Comparison of Machine Learning Model for Diabetes Prediction

Reference	ML Model	Methodology	Type of Diabetes	Focus Risk Factor	Performance Result
[11]	ERCT Algorithms	Quantitative	T2DM	• Psychological Factors • Medical History /Non-Modifiable Risk Factors	Accuracy: 85.55%
[12]	Boosting algorithm	Quantitative	T1DM	• Medical History /Non-Modifiable Risk Factors	Area Under the ROC Curve (AUC): 0.8805
[13]	A hybrid method combining T-Trees with the modeling of linkage disequilibrium	Quantitative	T1DM and T2DM	• Medical History /Non-Modifiable Risk Factors	T1DM AUC: 0.957 T2DM AUC: 0.961
[14]	Random Forest	Quantitative	T1DM and T2DM	• Psychological Factors • Medical History /Non-Modifiable Risk Factors	Accuracy: 94.66
[15]	Random Forest	Quantitative	T2DM	• Psychological Factors	AUC: 84.22
[16]	WVKSVM	Quantitative	T2DM	• Psychological Factors	Accuracy: 97.78%.
[17]	Artificial neural network	Quantitative	T2DM	• Psychological Factors • Medical History /Non-Modifiable Risk Factors	Accuracy: 97.18 %
[18]	Ada- boostM1 with a random committee	Quantitative	T1DM and T2DM	• Psychological Factors • Medical History /Non-Modifiable Risk Factors	Confusion Matrix: 81%
[19]	AdaBoost	Quantitative	T1DM and T2DM	• Psychological Factors	Accuracy: 86%.
[20]	Two-step binary logistic regression model	Quantitative	IGT&T2D	• Psychological Factors	Accuracy: 86.6%
[21]	Random forest	Quantitative	T2DM	• Health Behavior • Psychological Factors • Medical History /Non-Modifiable Risk Factors	Accuracy: 71.1%
[22]	Fully Corrective Binning	Quantitative	T2DM	• Psychological Factors • Medical History /Non-Modifiable Risk Factors	Accuracy: 85%
[23]	Decision Tree QUEST	Quantitative dan Qualitative	T2DM	• Health Behavior • Psychological Factors • Medical History /Non-Modifiable Risk Factors • Social Economic	Accuracy: 78%
[24]	Neural Network Back-propagation method and the Bayesian Regulation (BR) Algorithm	Quantitative	T1DM and T2DM)	• Psychological Factors • Medical History /Non-Modifiable Risk Factors	Accuracy: 92%

Reference	ML Model	Methodology	Type of Diabetes	Focus Risk Factor	Performance Result
[25]	The stacked generalization meta-learner	Quantitative	T1DM and T2DM	• Psychological Factors • Medical History /Non-Modifiable Risk Factors	Kappa coefficient: 0.95 (95%)
[26]	Hybrid SVM and Feature Selection	Quantitative	T2DM	• Psychological Factors • Medical History /Non-Modifiable Risk Factors	Accuracy: 97.87
[27]	Support vector machine (SVM) classifier with the wolf pack search (WPS) algorithm	Quantitative	T2DM	• Psychological Factors	Mean Squared Errors Function: N/A
[28]	Random Forests	Quantitative	T2DM	• Psychological Factors • Medical History /Non-Modifiable Risk Factors	Accuracy: 83.95%
[29]	E3-SVM	Quantitative	T2DM	• Psychological Factors • Medical History /Non-Modifiable Risk Factors	Accuracy: 80%
[30]	K-Nearest Neighbors	Quantitative	T2DM	• Psychological Factors • Medical History /Non-Modifiable Risk Factors	Sensitivity 100%
[31]	The improved K-means and logistic regression algorithms.	Quantitative	T2DM	• Psychological Factors • Medical History /Non-Modifiable Risk Factors	Accuracy: 95.42%

(IGT) or T2DM and only IGT. Then, [21] proposed decision tree and random forest techniques in order to identify the associated risk factors of T2DM using Mashhad Stroke and Heart Atherosclerotic Disorders (MASHAD). The random forest techniques performed best with 71.1% accuracy, 71.3% sensitivity, 69.9% specificity, and area under the ROC curve measuring 77.3% compare decision tree respectively.

In another study, [22] proposed prediction scoring method 5-year type 2 diabetes remission (DR) called 5y-Ad-DiaRem using machine learning fully corrective binning. The 5y-Ad-DiaRem model robustness was validated in three independent RYGB cohorts from three European countries. The 5y-Ad-DiaRem was more accurate compared to the previous studies such as DiaRem and Ad-DiaRem. The study [23] utilized data Tehran Lipid and Glucose Study (TLGS) to develop prediction models for the incidence of T2DM by utilizing 3 types of Decision Tree algorithms. The performances of the models were assessed using sensitivity, specificity, an area under the ROC curve (AUC), geometric mean (G-Mean) and F-Measure. The proposed model called The Quick Unbiased Efficient Statistical Tree (QUEST) algorithm resulting highest sensitivity and G-Mean among all the models for men and women.

In addition, [24] develop a model Artificial Neural Network with 8 input, 2 hidden layers with 10 neurons and 1 output layer to diagnose diabetes in pregnancies using PID dataset. The model is trained using back-propagation to correcting errors and Bayesian Regulation (BR) Algorithm are utilized to avoid overfitting the data set. Furthermore, the model was implemented by building a web-based system for diagnosing diabetes. [25] presented a comparison of ML model using a private dataset from EHR at Hospital Italiano de Buenos Aires, Buenos Aires, Argentina. The stacked generalization meta-learner was the greatest model compared with another model through Kappa coefficient value of 0.95 (95% CI 0.91, 0.98). A different study was provided using characteristic features of toe photoplethysmogram (PPG) as a data set for detection T2DM [26]. PPG signals were collected from 58 healthy and 83 type-2 DM subjects during routine checkup of patients at Medicine out-patient Department (OPD) of Dr. B.R.A Memorial Hospital, Raipur, India. A model was designed based on Hybrid Support Vector Machine and Feature Selection with an accuracy of 97.87% to predict T2DM based on data PPG. A similar study that has been done before has been presented by [27]. They proposed a model support vector machine (SVM) classifier with the wolf pack search (WPS) algorithm by conducting private dataset from Chi Mei Hospital, Tainan, Taiwan. Hence, the difference between this study is the proposed model is improved using optimization technique and feature selection technique [26] [27].

A risk prediction model has also been developed in order to understand the Single Nucleotide Polymorphisms (SNPs) feature related to T2DM using Random Forest (RF) [28]. The study also compared RF with SVM and Logistic Regression (LR) in order to show the best performance based on the area under the ROC curve and RF successfully perform the best result of AUC 0.89 on risk prediction without any problem such as complexity of features' interactions, overfitting, and unknown attribute values. Another study [29] proposed a classification model called E3-SVM that former as efficient and effective ensemble model using the real-world dataset of the anti-diabetic drug failure for predict type 2 diabetes. This model is designed using the bootstrapping technique or bagging technique which bootstrap samples were drawn from a given dataset to improve accuracy.

The result of the experiment from the proposed model E3-SVM show about 80% classification accuracy based on a dataset from Seoul National University Hospital in the Republic of Korea from 2003 to 2013.

Behadada et al. [30] applied a k-Nearest Neighbour (k-NN) for predicting the relation of metabolic syndrome with physiological parameters (age, BMI, level of glucose in the blood, etc.). The model has been compared with Naïve Bayes (NB) and Artificial Neural Network (ANN) based on sensitivity that indicated the performance of the K-NN Model achieved 100% fit to this type of data. However, they not presenting accuracy for the performance evaluation which is become standard measurement for ML. Finally, [31] was developing a new model for predicting T2DM using public dataset PID and generalize the model by giving a training procedure into another dataset that provided by medical expertise also from the questionnaire that has been prepared. Moreover, the model is compared with another researcher that used PID dataset to justify the model. The model was developed based on an improved K-means algorithm as clustering technique that used to the partitioning class of data with the same cluster then logistic regression as a classification to afford extracted information that related with significant data for predicting T2DM. The result has concluding three reasons that justify the model has significant improvement. The model show accuracy with 95.42% compares from the previous model. Then, the result of prediction from two different datasets shows the accuracy value of 0.907 and 0.935 that indicate more than 90% arcuately. Yet, the time of computation must be considered if the learning process from the model are implemented into the real system, especially electronic Health record is applicable in today hospital database.

III. DISCUSSION

In this study, a previous literature study was reviewed with respect to Machine Learning application for Diabetes Prediction. From the review, we can see there are several ML that has been used for instance RF, SVM, LG, ANN, Ensemble Learning and a hybrid model that combine in different strategies. Recently, the focus development by some researcher for predicting T1DM and T2DM. However, preventing T2DM has become more concern in every country in the word because epidemiology of T2DM significantly increases as monitored by the International Diabetes Federation and WHO to regularly update the prevalence of T2DM [32]. Additional, according to Larsson, et. al. stated that T1DM and T2DM are associated with the incidence of seven cardiovascular diseases (CVD) that can develop complication or death [33].

Generally, all ML method is quantitative approaches. Nevertheless, qualitative approaches can be added to some procedure in collecting data, then this makes the ML model suitable quantitative or qualitative depending on the research process and problem that can be solved. Furthermore, all the model has designed through the public or private dataset. The public dataset in this term of diabetes mostly used PID dataset because available in UCI ML Repository and biological data genome-wide association study (GWAS, https:// gwas.nih.gov/) data set that is based on single nucleotide polymorphism (SNP) [34]. The public dataset is often to use since the researcher can develop model easily by comparing performance and other parameters in the process of experimentation. Then, the competition among the researcher in improving knowledge and state of the art method can be more competitive. Moreover, private dataset mostly focusses in the real problem of data, for instance, Electronic Health Record (EHR), case study, questionnaire and so on. This condition becoming pro and cont. since the application of diabetes prediction model is needed practically in the term of clinical decision making. In this study, we are pointing out focus on risk factors such as Health behaviors: Tobacco use, Alcohol consumption, Physical inactivity, Sedentary activity, Fruit/vegetable intake, Simple carbohydrate etc.; Social and Economic Factors: Household income, Wealth, Social interaction, Food security, Workforce, Education, etc.; Psychological Factors: Obesity/Overweight, Systolic, Total cholesterol, triglyceride, Blood glucose, etc.; Medical History: Race/ethnicity, History of gestational diabetes, Age, Family history, DNA, HBA1c, gender, etc., that are considered in order to understand typical of data and possibility relation between the risk factors when used ML for Diabetes Prediction.

For evaluation, usually, performance measurement of the model depends on the learning process, techniques, and type of data. Numerous performance measurements that has been used in previous research is Accuracy, Sensitivity, Specificity, Peirce skill score (PSS), Heidke skill score (HSS), AUC/ROC, Precision, Recall, Kappa Statistic, Confusion matrix, Mean Square Error (MSE), Mathews correlation coefficient (MCC) and more. Hence, in this study, we select accuracy as our focus because more general and most of the researcher using this evaluation. The used of performance evaluation mostly for justification of the model when achieved the improvement result after a new strategy is applied and for comparing several models. However, in the term of diabetes, the prediction accuracy of the model is needed not only when the model well trained. It must have the ability to handle big data or EHR with consistent accuracy, reliability, and optimized computational time.

IV. CONCLUSION

ML is becoming on important tools for screening, diagnostic, treatment, prognosis, monitoring, and management of diabetic research [35], [36]. Based on the literature above, the accuracy of the diabetic prediction models is above 80%. In additional, supervised learning are the most commonly used and have been employed in the task of prediction such as classification and regression. In this study shows Random Forest and Support Vector Machine are popular ML approaches used among the researchers. Yet, both techniques have strengths and weaknesses that are needed to be considered [37]. Even the model shows the best accuracy for the specific dataset but the problem will arise if the strategies on dealing with another dataset are not well defined. The potential solution on predictive modeling based on a machine learning technique can improve management diabetes and it shows a big challenged to consider not only clinical data hence numerous factors are uncontrolled are identified.

For future research, we suggest optimization technique and feature selection be integrated into such a learning schema or improved performance of the model. Moreover, it is important to define clearly information required, such as dimensionality, number of features or even the clinical, genetic, or another data type in line with the objective of the prediction, to improve the performance of the ML diabetic prediction model and can be more robust.

ACKNOWLEDGMENT

The authors would like to acknowledge Advanced Manufacturing Centre, Faculty of Information and Communication Technology, Centre for Research and Innovation Management, Universiti Teknikal Malaysia Melaka and Universitas ‘Aisyiyah Yogyakarta for providing all facilities and financial support in this study. This project supported by FRGS/2018/FTMK-CACT/F00388. The authors are very thankful to reviewer for their valuable feedback and comments to improve the contents of this article.

REFERENCES

1. C. C. Thomas and L. H. Philipson, “Update on Diabetes Classification,” Med. Clin. North Am., vol. 99, no. 1, pp. 1–16, 2015.
2. N. H. Cho et al., “IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045,” Diabetes Res. Clin. Pract., vol. 138, pp. 271–281, Apr. 2018.
3. Y. Hayashi and S. 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 Med. Unlocked, vol. 2, pp. 92–104, 2016.
4. American Diabetes Association., “Summary of Revisions: Standards of Medical Care in Diabetes—2018,” Diabetes Care, vol. 41, no. Supplement 1, pp. S4–S6, Jan. 2018.
5. Z. Tao, A. Shi, and J. Zhao, “Epidemiological Perspectives of Diabetes,” Cell Biochem. Biophys., vol. 73, no. 1, pp. 181–185, 2015.
6. P. Harrington, Machine Learning in Action. Manning Publications Co, 2012.
7. M. Nilashi, O. bin Ibrahim, H. Ahmadi, and L. Shahmoradi, “An analytical method for diseases prediction using machine learning techniques,” Comput. Chem. Eng., vol. 106, pp. 212–223, 2017.
8. J. Liu et al., Foundations of Machine Learning, vol. 17, no. 4. 2012.
9. I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, “Machine Learning and Data Mining Methods in Diabetes Research,” Comput. Struct. Biotechnol. J., vol. 15, pp. 104–116, 2017.
10. World Health Organization, “Global Report on Diabetes,” 2016.
11. N. A. Vignesh and D. I. G. Amalarethinam, “Rule Extraction for Diagnosis of Diabetes Mellitus Used for Enhancing Regular Covering Technique,” in 2017 World Congress on Computing and Communication Technologies (WCCCT), 2017, pp. 111–114.
12. V. Potenciano, M. M. Abad-Grau, A. Alcina, and F. Matesanz, “A comparison of genomic profiles of complex diseases under different models,” BMC Med. Genomics, vol. 9, no. 1, pp. 1–16, 2016.
13. C. Sinoquet, “A method combining a random forest-based technique with the modeling of linkage disequilibrium through latent variables, to run multilocus genome-wide association studies,” BMC Bioinformatics, vol. 19, no. 1, pp. 1–24, 2018.
14. S. Poonkuzhali, J. Jeyalakshmi, and S. Sreesubha, “Diabetes Mellitus Risk Factor Prediction Through Resampling and Cost Analysis on Classifiers,” vol. 836, Springer Singapore, 2018, pp. 212–225.
15. A. Pimentel, A. V. Carreiro, R. T. Ribeiro, and H. Gamboa, “Screening diabetes mellitus 2 based on electronic health records using temporal features,” Health Informatics J., vol. 24, no. 2, pp. 194–205, Jun. 2018.
16. X. Huang, Q.-S. Xu, Y.-H. Yun, J.-H. Huang, and Y.-Z. Liang, “Weighted variable kernel support vector machine classifier for metabolomics data analysis,” Chemom. Intell. Lab. Syst., vol. 146, pp. 365–370, Aug. 2015.
17. M. Heydari, M. Teimouri, Z. Heshmati, and S. M. Alavinia, “Comparison of various classification algorithms in the diagnosis of type 2 diabetes in Iran,” Int. J. Diabetes Dev. Ctries., vol. 36, no. 2, pp. 167–173, Jun. 2016.
18. R. Ali, M. H. Siddiqi, M. Idris, B. H. Kang, and S. Lee, “Prediction of Diabetes Mellitus Based on Boosting Ensemble Modeling,” in Ubiquitous Computing and Ambient Intelligence. Personalisation and User Adapted Services, vol. 8867, 2014, pp. 25–28.
19. U. R. Acharya et al., “An integrated diabetic index using heart rate variability signal features for diagnosis of diabetes,” Comput. Methods Biomech. Biomed. Engin., vol. 16, no. 2, pp. 222–234, Feb. 2013.
20. G. Acciaroli et al., “Diabetes and Prediabetes Classification Using Glycemic Variability Indices From Continuous Glucose Monitoring Data,” J. Diabetes Sci. Technol., vol. 12, no. 1, pp. 105–113, Jan. 2018.
21. H. Esmaily, M. Tayefi, H. Doosti, M. Ghayour-Mobarhan, H. Nezami, and A. Amirabadizadeh, “A Comparison between Decision Tree and Random Forest in Determining the Risk Factors Associated with Type 2 Diabetes.,” J. Res. Health Sci., vol. 18, no. 2, p. e00412, Apr. 2018.
22. J. Debédat et al., “Long-term Relapse of Type 2 Diabetes After Roux-en-Y Gastric Bypass: Prediction and Clinical Relevance,” Diabetes Care, vol. 41, no. 10, pp. 2086–2095, Oct. 2018.
23. A. Ramezankhani, E. Hadavandi, O. Pournik, J. Shahrabi, F. Azizi, and F. Hadaegh, “Decision tree-based modelling for identification of potential interactions between type 2 diabetes risk factors: a decade follow-up in a Middle East prospective cohort study,” BMJ Open, vol. 6, no. 12, p. e013336, Dec. 2016.
24. O. M. Alade, O. Y. Sowunmi, S. Misra, R. Maskeliūnas, and R. Damaševičius, “A Neural Network Based Expert System for the Diagnosis of Diabetes Mellitus,” in Advances in Intelligent Systems and Computing, vol. 724, 2018, pp. 14–22.
25. S. Esteban et al., “Development and validation of various phenotyping algorithms for Diabetes Mellitus using data from electronic health records,” Comput. Methods Programs Biomed., vol. 152, pp. 53–70, Dec. 2017.
26. N. Nirala, R. Periyasamy, B. K. Singh, and A. Kumar, “Detection of type-2 diabetes using characteristics of toe photoplethysmogram by applying support vector machine,” Biocybern. Biomed. Eng., vol. 39, no. 1, pp. 38–51, 2019.
27. C.-M. Li et al., “Synchronizing chaotification with support vector machine and wolf pack search algorithm for estimation of peripheral vascular occlusion in diabetes mellitus,” Biomed. Signal Process. Control, vol. 9, no. 1, pp. 45–55, Jan. 2013.
28. B. López, F. Torrent-Fontbona, R. Viñas, and J. M. Fernández-Real, “Single Nucleotide Polymorphism relevance learning with Random Forests for Type 2 diabetes risk prediction,” Artif. Intell. Med., vol. 85, pp. 43–49, Apr. 2018.
29. S. Kang, P. Kang, T. Ko, S. Cho, S. Rhee, and K.-S. Yu, “An efficient and effective ensemble of support vector machines for anti-diabetic drug failure prediction,” Expert Syst. Appl., vol. 42, no. 9, pp. 4265–4273, Jun. 2015.
30. O. Behadada, M. Abi-Ayad, G. Kontonatsios, and M. Trovati, “Automatic Diagnosis Metabolic Syndrome via a k-Nearest Neighbour Classifier,” vol. 10232, M. H. A. Au, A. Castiglione, K.-K. R. Choo, F. Palmieri, and K.-C. Li, Eds. Cham: Springer International Publishing, 2017, pp. 627–637.
31. H. Wu, S. Yang, Z. Huang, J. He, and X. Wang, “Type 2 diabetes mellitus prediction model based on data mining,” Informatics Med. Unlocked, vol. 10, no. August 2017, pp. 100–107, 2018.
32. P. Z. Zimmet, D. J. Magliano, W. H. Herman, and J. E. Shaw, “Diabetes: A 21st century challenge,” Lancet Diabetes Endocrinol., vol. 2, no. 1, pp. 56–64, 2014.
33. S. C. Larsson, A. Wallin, N. Håkansson, O. Stackelberg, M. Bäck, and A. Wolk, “Type 1 and type 2 diabetes mellitus and incidence of seven cardiovascular diseases,” Int. J. Cardiol., no. 2017, pp. 1–5, 2018.
34. L. B. Holder, M. M. Haque, and M. K. Skinner, “Machine learning for epigenetics and future medical applications,” Epigenetics, vol. 12, no. 7, pp. 505–514, 2017.
35. N. Esfandiari, M. R. Babavalian, A. M. E. Moghadam, and V. K. Tabar, “Knowledge discovery in medicine: Current issue and future trend,” Expert Syst. Appl., vol. 41, no. 9, pp. 4434–4463, 2014.
36. I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, “Machine Learning and Data Mining Methods in Diabetes Research,” Comput. Struct. Biotechnol. J., vol. 15, pp. 104–116, 2017.
37. C. L. C. Koo, M. M. J. Liew, M. S. Mohamad, and A. H. M. Salleh, “A review for detecting gene-gene interactions using machine learning methods in genetic epidemiology,” Biomed Res. Int., vol. 2013, p. 13, 2013.

AUTHORS PROFILE

Nur Rachman Dzakiyullah Received S.Kom from Informatics Engineering in Universitas Islam Indonesia, Yogyakarta and Master of Science in Information and Communication Technology from Universiti Teknikal Malaysia Melaka (UTeM), Malaysia. Currently working as a lecturer in Universitas 'Aisyiyah Yogyakarta, Indonesia and Ph.D. student in Universiti Teknikal Malaysia Melaka (UTeM), Malaysia. Research areas of interest include Industrial Computing, Operation Research, Modelling and Decision Technology, Data Mining, Artificial Intelligence, Health Informatics, Diabetics Research. He is a member of APTIKOM and IAENG.

Assoc. Prof. Dr. M.A. Burhanuddin is a lecturer in the Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka (UTeM). Burhanuddin is a former Deputy Dean of Research and Postgraduate studies in the Faculty of Information and Communication Technology, UTeM. He also is a former Head of Industrial Computing Department in the Faculty of Information and Communication Technology, UTeM. He has been actively involved in all levels of the university development and one of the pioneers in building the faculties in UTeM, Malaysia. Also, he was one of the pioneers in developing the faculties of Computing and Information Technology Rabigh, King Abdulaziz University in Rabigh, Kingdom of Saudi Arabia. He has huge experience in managing research grants and supervises many students, i.e. under-graduate and postgraduate levels. Burhanuddin teaching, research and consultancy work focus on the areas of multiple criteria decision-making models, decision support system, optimization techniques, operational research, soft computing and artificial intelligence.

Dr. Raja Rina Raja Ikram is a lecturer in the Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka (UTeM). She earned his Bachelor in the University of Melbourne. She received Master (Health Informatics) and Doctor of Philosophy (Ph.D.) (Health Information Systems), from Universiti Teknikal Malaysia Melaka. Her areas of interest are Medical Informatics, Health Informatics, Healthcare Informatics, Telemedicine, Health Information Management, eHealth, EMR, Clinical Informatics, Hospital Information Systems

Dr. Mohd Khanapi Abd Ghani is a Professor at the Department of Software Engineering, Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka (UTeM). He earned his Diploma in Computer Science, from Universiti Teknologi MARA. He then received his Bachelor Degree in Computer Science and Master of Science in Real-Time Software Engineering from Universiti Teknologi Malaysia. He then pursues the third degree in Biomedical Computing and earned a Doctor of Philosophy (Ph.D.) from Coventry University, the United Kingdom in 2009. His research areas of interest include electronic healthcare systems, telemedicine, healthcare knowledge management, system architecture and software reuse. He was appointed as Telehealth Associate Consultant for AIH Group (Malaysia) Sdn Bhd for developing telehealth application nationally and internationally. He is also a Principal Consultant for BIT Group for Health Information System development. He is a founder of Biomedical Computing and Engineering Technologies Research Group and teaching Advanced Software Engineering, Software Analysis and Design, Software Testing and Quality Assurance and, Software Project Management.

Winny Setyonugroho is a lecturer in the Faculty of Medicine and Health Science Muhammadiyah University of Yogyakarta (UMY). He received Bachelor of Medicine (S.Ked) from Faculty of Medicine, Gadjah Mada University (UGM), Indonesia, Master of Information Technology, Electrical Engineering Department, Engineering Faculty, Gadjah Mada University (UGM), Indonesia and Ph.D. from Medical Informatics & Education, School of Medicine, National University of Ireland, Galway. Research areas of interest include Health Informatics, Psychometric analysis, eLearning (health informatics in medical education), Technology Management.

Machine Learning Methods for Diabetes Prediction

ORIGINALITY REPORT

8%	7%	6%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Www.intechopen.com Internet Source	<1%
2	philpapers.org Internet Source	<1%
3	scholar.uwindsor.ca Internet Source	<1%
4	Longquan Jiang, Bo Zhang, Qin Ni, Xuan Sun, Pingping Dong. "Prediction of SNP Sequences via Gini Impurity Based Gradient Boosting Method", IEEE Access, 2019 Publication	<1%
5	edepot.wur.nl Internet Source	<1%
6	www.ls2n.fr Internet Source	<1%
7	Reema Shyamsunder Shukla, Yogender Aggarwal. "Fourier Transform and Autoregressive HRV Features in Prediction and Classification of Breast Cancer", IETE Journal of Research, 2021	<1%

	Publication	
8	deepai.org Internet Source	<1 %
9	e-spacio.uned.es Internet Source	<1 %
10	Svetlana Girs, Severine Sentilles, Sara Abbaspour Asadollah, Mohammad Ashjaei, Saad Mubeen. "A Systematic Literature Study on Definition and Modeling of Service-Level Agreements for Cloud Services in IoT", IEEE Access, 2020 Publication	<1 %
11	eprints.covenantuniversity.edu.ng Internet Source	<1 %
12	bmcmedinformdecismak.biomedcentral.com Internet Source	<1 %
13	doi.org Internet Source	<1 %
14	care.diabetesjournals.org Internet Source	<1 %
15	Ying Wang, Baichun Hu, Shasha Feng, Jian Wang, Fengjiao Zhang. "Target recognition and network pharmacology for revealing anti-diabetes mechanisms of natural product", Journal of Computational Science, 2020 Publication	<1 %

16	"Intelligent Healthcare", Springer Science and Business Media LLC, 2021 Publication	<1 %
17	www.springerprofessional.de Internet Source	<1 %
18	dblp.dagstuhl.de Internet Source	<1 %
19	www.hindawi.com Internet Source	<1 %
20	www.jetir.org Internet Source	<1 %
21	Kumar, Ashok, Priyanka Thakur, Kanika Gupta, and Amit Pal. "Text Mining Approach to Analyse the Relation between Obesity and Breast Cancer Data", International Letters of Natural Sciences, 2015. Publication	<1 %
22	www.imagesensors.org Internet Source	<1 %
23	pure.qub.ac.uk Internet Source	<1 %
24	Yunus EMRE Isik, Yasin Gormez, Zafer Aydin, Burcu Bakir-Gungor. "The Determination of Distinctive Single Nucleotide Polymorphism Sets for the Diagnosis of Behçet's Disease",	<1 %

IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2021
Publication

25 Aicha Boutorh, Ahmed Guessoum. "Complex diseases SNP selection and classification by hybrid Association Rule Mining and Artificial Neural Network—based Evolutionary Algorithms", Engineering Applications of Artificial Intelligence, 2016
Publication
<1 %

26 U. RAJENDRA ACHARYA, HAMIDO FUJITA, SHREYA BHAT, JOEL EW KOH et al. "AUTOMATED DIAGNOSIS OF DIABETES USING ENTROPIES AND DIABETIC INDEX", Journal of Mechanics in Medicine and Biology, 2016
Publication
<1 %

27 livrepository.liverpool.ac.uk
Internet Source
<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		