School of Informatics

Informatics Research Review
A Review of Tree-based Classifiers for Diagnosing Diabetes

[redacted]

January 2021

Abstract

Data Mining is the process of collecting, exploring, and modelling large data set to identify patterns. When facing with a large number of data in the medical field, it is vital to come up with appropriate scientific technology to assist medical stuff identifying disease. Tree-based classifiers usually provide a good performance in the medical field. In this case, we will give a systematic view of the tree-based classifiers and ensemble methods used for diagnosing diabetes. We are focusing on ID3 Algorithm, C4.5 Algorithm, C5.0 Algorithm and CART Algorithm. Besides, ensemble methods such as Bagging Algorithm and Boosting Algorithm also be reviewed.

Date: Friday 29th January, 2021

Supervisor: [redacted]

1 Introduction

Diabetes, also called diabetes mellitus (DM), is a kind of metabolic disease characterized by high blood sugar. It is considered as one of the top 10 leading causes of death in 2019 [1]. Up to 2019, 463 million people in the world are living with diabetes, which means 1 in 11 people are suffering from this disease [2]. Diabetes can cause a large number of complications such as diabetic neuropathy, diabetic nephropathy. These complications are harmful to human nerves and organs. Once the complication is being diagnosed, it is hard to be cured. The earlier diabetes is diagnosed, the more complications can be reduced. Therefore, it is vital for human beings to find effective ways to diagnose and treat diabetes.

There are three main types of diabetes: Type 1 diabetes mellitus, Type 2 diabetes mellitus, and Gestational diabetes mellitus (GDM) [3]. Type 1 diabetes mellitus is caused by deficient insulin production, this kind of patients need a daily injection of insulin to maintain normal levels of insulin in their body. Type 2 diabetes mellitus is resulting from the body's insufficient use of insulin. More than 90% of diabetic patients suffering from this type of diabetes. Unlike Type 1 diabetes mellitus, Type 2 diabetic patients do not need a daily injection of insulin, instead, they need to stimulate the secretion of insulin in the body through certain drugs. Gestational diabetes mellitus is different from the previous two types, it only occurs in pregnant women. Most patients can recover after childbirth. All these types of diabetes can be diagnosed and treated through a urine test and blood test at an early stage. However, the huge number of patients mentioned above makes researchers searching for new technologies to speed up the diagnosis process.

Data Mining is the process of collecting, exploring, and modelling large data set to identify patterns. It is one of the most common and successful techniques in the medical field [4]. Applying data mining techniques can assist medical stuff identifying diabetes in a faster and more accurate way. A vast number of machine learning algorithms such as Decision Tree (DT) [5], Support Vector Machine (SVM) [6], Random Forest (RF) [7], k-Nearest Neighbors (kNN) [8] are being used in modeling process, which perform well in detecting diabetes.

A decision tree is a flow-chart-like tree structure, which is composed of nodes (internal nodes and leaf nodes) and directed edges [9]. While using the decision tree, the test process is starting from the root node, and then assign the instance to its child node according to its value. Each child node has criteria linked to them. Using a recursive method, the instance will be assigned to one of the leaf nodes in the end. This indicates which class the instance belongs to. Decision tree is easier to understand and implement compared to other classification algorithms. Once the decision tree model is built, it can be used repeatedly. Also, decision tree classifiers can obtain better performance in many cases [10]. Commonly used variations of decision tree algorithms applied on data to predict diabetes are Iterative Dichotomiser 3 (ID3 Algorithm), C4.5 Algorithm also called J48 decision tree, C5 Algorithm, and Classification and Regression Tree (CART Algorithm).

Saba [11] mentioned that no single classifier can consistently obtain the maximum accuracy on different data sets. Therefore, ensemble methods and other combination methods have been used in subsequent experiments related to diabetes. Commonly used ensemble methods are Bagging algorithm and Boosting algorithm.

This paper tends to provide a comprehensive literature review of commonly used tree-based classifiers (ID3 Algorithm, C4.5 Algorithm, C5 Algorithm, and CART Algorithm) as well as ensemble methods (Boosting methods and Averaging methods) in diabetes diagnosis field. The rest of paper is organized as follows: Section 2 gives an overview of several commonly used tree-

based classifiers and ensemble methods. Their advantages and disadvantages will be mentioned in this part. Section 3 provides a literature review based on tree-based classifiers and Section 4 provides a literature review for ensemble methods. Section 5 will summaries the entire review and concludes.

2 Overview of Tree-based Classifiers and Ensemble Methods

2.1 ID3 Algorithm

ID3 Algorithm is a simple decision tree algorithm proposed by Quinlan Ross in 1986 [12]. The core of this algorithm is using Information Gain as the baseline to select features recursively on each internal nodes. However, only categorical features can be accepted. It is serially implemented and based on Hunt's algorithm. The ID3 Algorithm follows Occam's razor principle. It attempts to create the smallest possible decision tree in a very short time.

Although ID3 Algorithm is easy to understand and simple to implement, it has four main drawbacks shown as below:

- Classifying continues features is expensive. It has to turn continuous features into discrete features.
- Using Information Gain to split the data is not accurate.
- Missing values are hard to handle.
- The structure of the decision tree might be complex without using pruning technique, which may lead to an over-fitting problem.

2.2 C4.5 Algorithm

C4.5 Algorithm (known in WEKA as J48 decision tree, which is an open-source Java implementation of the C4.5 Algorithm) is developed based on ID3 Algorithm by Ross Quinlan in 1993 [13]. It is serially implemented and based on Hunt's algorithm. Instead of using Information Gain as the baseline to select features, C4.5 Algorithm uses information gain rate to split attributes to overcome bias problem. This algorithm has solved problems occurred in ID3 Algorithm mentioned above. There are several improvements compared to ID3 Algorithm.

- Unlike ID3 Algorithm, C4.5 Algorithm can handle both continuous features and discrete features.
- C4.5 Algorithm can handle missing values by marking them as '?'.
- Pruning technique is also introduced in C4.5 Algorithm after creating the tree. This is done by replacing the internal node with a leaf node, which can reduce the impact of noise in the data set.

2.3 C5.0 Algorithm

C5.0 Algorithm is one of the improved versions of C4.5 Algorithm. C5.0 Algorithm offers several improvements on C4.5 Algorithm.

- C5.0 Algorithm has faster processing speed compared to the C4.5 Algorithm.
- C5.0 Algorithm can build smaller decision tree compare to the C4.5 Algorithm but with the same result.
- C5.0 Algorithm is more memory efficient than C4.5 Algorithm.

2.4 CART Algorithm

CART Algorithm usually refers to the 'decision tree' which can be applied in both classification and regression problems. It was introduced by Breiman in 1984. The same as C4.5 Algorithm, it is an improved version of ID3 Algorithm. It is serially implemented and based on Hunt's algorithm. Unlike ID3 Algorithm and C4.5 Algorithm, CART Algorithm uses Gini Impurity as the feature selection measurement to split the data set into a binary decision tree.

Compared to ID3 Algorithm, CART Algorithm can handle continuous features and discrete features at the same time. Besides, it uses cost complexity pruning to remove the unreliable branches from the decision tree to improve the accuracy. The CART Algorithm provides a foundation for important algorithms like bagged decision trees, random forest and boosted decision trees.

2.5 Ensemble Methods

The idea of Ensemble methods is combining multiple classifiers to improve accuracy. Usually, the combined model has better performance in accuracy than a single classifier. However, complex ensemble models are slower and less efficient compare to a single classifier.

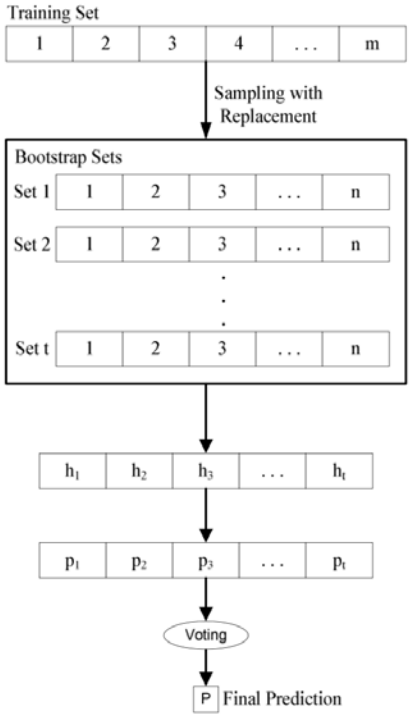

Figure 1: Bagging ensemble classifier framework

At present, there are mainly two common ensemble methods, one is Bagging-based methods, the other is Boosting-based methods.

Bagging algorithm, sometimes called bootstrap aggregation, is one of the machine learning ensemble algorithms which can be used to improve stability and accuracy. It was developed by Leo Breiman in 1996 [14]. Bagging algorithm conducts sampling of the training set in a random order and then trains new models through repeated sampling. The final result would be obtained by averaging among these models. The framework of Bagging ensemble classifier is shown in Fig 1.

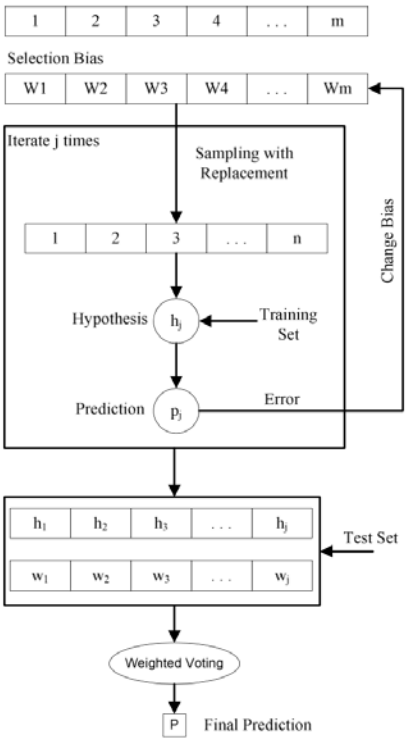

Figure 2: Boosting ensemble classifier framework

Boosting methods was developed by Schapire [15]. It firstly uses a basic algorithm to do the prediction, and then use the results of the previous algorithm in other subsequent algorithms to mainly focus on processing error data. Therefore, it can reduce the error rate continuously. One algorithm in Boosting methods is called Adaboost, it is developed for binary classification models by Freund and Schapire [16]. The framework of Boosting ensemble classifier is shown in Fig 2.

3 Literature Review of Tree-based Classifiers

In this section, we will discuss commonly used extensions of decision tree algorithm. They are Iterative Dichotomiser 3 (ID3 Algorithm), C4.5 Algorithm also called J48 decision tree, C5 Algorithm, and Classification and Regression Tree (CART Algorithm).

3.1 ID3 Algorithm

Weifeng Xu [17] uses medical database generated from the University of Virginia to predict the risk of diabetes. After applying dimensionality reduction and discretization to improve the performance of algorithms, they use RF, Naïve Bayes Algorithm (NB), ID3 Algorithm and AdaBoost Algorithm with 10-fold cross-validation to evaluate. The result shows that the ID3 Algorithm has a lower value in Accuracy compared to the rest three methods. It turns out the ID3 Algorithm has several drawbacks which can affect the accuracy of prediction.

3.2 C4.5 Algorithm

To identify the performance of one algorithm, researchers start from applying the algorithm to the different data set for accuracy testing. According to previous experiments, 10-fold cross validation is the best choice to get better accuracy among other test methods. In this case, Asma [5] uses Pima Indian Diabetes Data set (PIDD) with 10-fold cross validation as their test option to predict diabetes. They also use attributes identification and selection, missing values handling methods, and numerical discretization to improve the quality of the data set. The result shows that the accuracy of applying the J48 decision tree is 78.1768%. However, more data generated from different sources should be considered as one direction of improvement. Later, Chen [18] use almost the same methods in the evaluation process as Asma to diagnose Type 2 diabetes. The experiment is performed on PIDD with J48 decision tree and 10-fold cross-validation method. This experiment achieves better accuracy than the previous experiment and other models mentioned in the paper. The accuracy of the J48 decision tree in this experiment can reach to 90.04%. Since the model is only applied to Type 2 diabetes which is defined as a binary classification problem, authors want to extend the scope of application to the multi-class classification problem. Also, features of the data set used in the experiment are all numerical data, they suggest that the model should be able to process other types of data, like images or signals.

An improved J48 decision tree is then proposed by Kaur [19] in 2014. The modified J48 decision tree is aiming to improve the accuracy rate of the data mining procedure. The same as before, they use PIDD to evaluate the performance of the new model. It has been proved that the newly proposed J48 decision tree has accuracy up to 99.8% while the previous version of the J48 decision tree has accuracy 73.8%.

Apart from focusing on the testing performance of a single classifier, researchers are keening to compare various classification algorithms. Emrana [20] compares the performances of kNN and C4.5 Algorithm based on the statistical medical data. During the testing phase, the C4.5 Algorithm provides 90.43% accuracy while kNN provides 76.96% accuracy. In addition to this, by comparing previous experiments using PIDD, C4.5 Algorithm in this paper achieves the highest accuracy.

Pradeep [21] uses J48 decision tree, kNN, RF and SVM to classify diabetes patients. The data set is obtained from UCI machine learning data repository. Authors performed two tasks based on different data set separately, one is the data set with noise data (perform before pre-processing), the other is the data set without noise data (perform after pre-processing). For the data set with noise data, the J48 decision tree achieves the highest accuracy (73.82%) among other classifiers. However, for data set without noise data, both KNN with k=1 ad RF achieve 100% accuracy. It turns out that removing noise data before applying machine learning algorithms can provide better results.

To further improve accuracy, researchers then tends to modify pre-processing steps. Zheng [22] using improved filtering criteria to evaluate the performance of kNN, NB, J48 decision tree, RF, SVM, and Logistic Regression (LR). Obviously, the recall rate has been improved with a low false positive rate. There are still some limitations in this paper, for example, the number of samples is limited, models have relatively low specificity. Sun [23] also proposed an optimal decision tree model using Expectation-maximization (EM) clustering algorithm to remove the incorrect classified data. The proposed model can achieve 92.81% accuracy. However, more data pre-processing steps can be applied to improve model performance. The other problem in this experiment is the same as the problem mentioned in Chen's paper. The proposed model can only test on numerical data, we need to extend this model to other types of data.

3.3 C5.0 Algorithm

Since the C5.0 Algorithm is just a simple improvement of C4.5 Algorithm, few papers use this algorithm to compare with other classification algorithms. Meng [24] proposed a study to compare the performance of LR, Artificial Neural Networks (ANNs), and C5.0 Algorithm for predicting diabetes. The data set is obtained by questionnaire, which has 735 instances and composed of 12 features and one output. The experiment result shows that the C5.0 Algorithm achieves the best classification accuracy among the rest two classification algorithm.

3.4 CART Algorithm

To evaluate the performance of CART Algorithm, Anand [25] uses several factors in people lifestyle combined with some indicators such as BMI, weight as the data set, using CART Algorithm with 10-fold cross-validation to predict. It shows that CART Algorithm can provide 75% accuracy. Meanwhile, Saravananathan [26] compared the J48 decision tree with CART Algorithm. The results for the accuracy of these two classifiers are 67.15%, 62.28%. This indicates that the J48 decision tree provides higher accuracy than CART Algorithm. However, a relatively small number of features in the current data set limits the performance of classifiers. It is important to find more attributes as the features in the training set. Besides, they mentioned that one way of performance improvement is applying a hybrid classification method.

Hathaway [27] compares LR, Linear Discriminant Analysis (LDA), Gaussian Naive Bayes (NB), SVM, and CART Algorithm using 10-fold cross validation. It shows that CART Algorithm achieves the highest accuracy through these methods, which is 75% area under the ROC curve.

When it comes to pre-processing, Nilashi [28] uses CART Algorithm to generate a fuzzy rule. During the pre-processing procedure, they have applied fuzzy rule, clustering algorithm and noise removal methods to improve the quality of data set. Obviously, the result shows that the CART Algorithm can give better performance after pre-processing data set before applying machine learning algorithms in disease prediction.

One of the breakthroughs in predicting diabetes using classification algorithms is applying hybrid classification methods. Mira Kania Sabariah [29] combined CART Algorithm with RF as a new prediction model for predicting diabetes. The result shows that the combined model has higher accuracy (83.8%) than the single CART Algorithm.

4 Literature Review of Ensemble Methods

In this section, we will discuss the performance differences among the tree-based classifiers using various ensemble methods. There are many ensemble methods such as Majority Voting, Adaboost, Bayesian Boosting, Stacking, Bagging and so on. We only conduct an exploratory review about Boosting Algorithm and Bagging Algorithm.

To investigate the impact of using ensemble techniques, Perveen [30] compare the performance of the J48 decision tree, Bagging with the J48 decision tree, and AdaBoosting with J48 decision tree. The evaluation indicates that Adaboosting with J48 decision tree outperforms than Bagging with J48 decision tree and J48 decision tree. As for future works, they suggest we can use more techniques such as NB, SVM and so on as the base learners in an ensemble framework to evaluate the performance. In 2014, Saba [31] examine the performances of ID3 Algorithm, C4.5 Algorithm and CART Algorithm using Majority Voting, Adaboost, Bayesian Boosting, Stacking and Bagging as ensemble methods. In this case, we only focus on two boosting algorithms (Adaboost and Bayesian Boosting) and Bagging algorithms. The overall experiment is based on two data set, one is PIDD, the other is BioStat data set. For both data set, they all show that classifiers with bagging algorithm have the highest accuracy than using a single classifier and the rest ensemble methods. One improvement of this paper is to use more individual classifiers as the base classifiers with ensemble methods, such as NB, SVM. This illustration is the same as Perveen mentioned in his paper.

In the following experiments, many tree-based classifiers are used in the experiment to evaluate ensemble methods. Tama [32] uses CART Algorithm, C4.5 Algorithm, Reduced Error Pruning Tree (REPT), Random Tree, Naïve Bayes Tree (NB), Functional Tree, Best-First Decision Tree (BFDT) and Logistic Model Tree (LMT) as base classifiers in five different ensemble methods, i.e. bagging, boosting, random subspace, DECORATE, and rotation forest. The experimental result indicates that LMT is the best classifier, no matter what ensemble methods are being used and whether it uses ensemble methods. The paper also suggests that more ensemble techniques should be considered in future study.

Instead of only using tree-based classifiers as the base classifiers with ensemble methods, Nongyao Nai-arun [33] conduct a study using Decision Tree, Artificial Neural Networks, Logistic Regression and Naive Bayes as base classifiers with Bagging methods and Boosting methods as ensemble methods to predict the risk of diabetes for everyone without the need of blood test or going to a hospital. After applying all these classifiers with ensemble methods, SVM also used to evaluate. The data set used in this study was collected from 26 Primary Care Units (PCU) in Sawanpracharak Regional Hospital. The result reveals that RF has the highest accuracy (85.558%) and ROC Curve among these classifiers.

5 Summary & Conclusion

Medical diagnosis is a serious topic in medical the field, which is closely related to human lives. During the past few years, applying data mining and machine learning techniques in this filed becomes a hot topic. At the same time, the increasing number of articles are focusing on the comparison between different algorithms. Researchers make efforts to investigate which model can provide high accuracy in diagnosis.

The creation of a particular model is to solve existing problems in the field. They are developed to improve the accuracy of prediction. Take tree-based classifiers as an example, the C4.5

Algorithm and CART Algorithm are created based on ID3 Algorithm. These two algorithms are aiming to solve problems in ID3 Algorithm. The same with C5.0 Algorithm, which is designed as an improvement version of C4.5 Algorithm.

When it comes to identifying how good a model is, researchers usually analysis its performance from the following aspects:

- Compared with the previous version of the algorithm. The comparison should be based on the same data set and the same pre-processing methods.
- For a group of classifiers with similar characteristics, researchers often make a comparison between newly developed classifier with others.
- Use more comprehensive pre-processing methods to improve the quality of data set, hence improve the model accuracy.
- It is often the case that the accuracy of a single classifier has an upper bound. Therefore, ensemble methods can be used to compare those single classifier to improve the robustness and accuracy of the model.

In this literature review, we discovered the development of tree-based classifiers, and particularly ID3 Algorithm, C4.5 Algorithm, C5.0 Algorithm as well as CART Algorithm. We critically reviewed each algorithm almost from the above 4 aspects. However, there are still many aspects that need to be improved. Different experiments usually use different data set to train and use the different pre-processing method before modelling. It turns out we can conclude with a different result. Besides, problems related to inadequate data set and features, used classifiers are limited should be solved in the future. Moreover, we need to try more single classifier as the base classifier with ensemble methods.

However, all these diagnoses using data mining and machine learning techniques is only an initial diagnosis. People who found that they are in the diabetes risk group should go to see a doctor for a formal diagnosis to prevent themselves from serious diabetes.

References

[1] The top 10 causes of death. `https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death`. Retrieved 9 December, 2020.

[2] Idf diabetes atlas ninth edition 2019. `www.diabetesatlas.org`. Retrieved 18 May, 2020.

[3] Diabetes fact sheet n°312. WHO. Retrieved 25 March, 2014.

[4] Abdullah A Aljumah, Mohammed Gulam Ahamad, and Mohammad Khubeb Siddiqui. Application of data mining: Diabetes health care in young and old patients. *Journal of King Saud University-Computer and Information Sciences*, 25(2):127–136, 2013.

[5] Asma A Al Jarullah. Decision tree discovery for the diagnosis of type ii diabetes. In *2011 International conference on innovations in information technology*, pages 303–307. IEEE, 2011.

[6] Nahla Barakat, Andrew P Bradley, and Mohamed Nabil H Barakat. Intelligible support vector machines for diagnosis of diabetes mellitus. *IEEE transactions on information technology in biomedicine*, 14(4):1114–1120, 2010.

[7] Enrique Monte Moreno, Maria José Anyo Luján, Montse Torrres Rusinol, Paqui Juárez Fernández, Pilar Nunez Manrique, Cristina Aragon Trivino, Magda Pedrosa Miquel, Marife Alvarez Rodriguez, and M José González Burguillos. Type 2 diabetes screening test by means of a pulse oximeter. *IEEE Transactions on Biomedical Engineering*, 64(2):341–351, 2016.

[8] Eleonora Maria Aiello, Chiara Toffanin, Mirko Messori, Claudio Cobelli, and Lalo Magni. Postprandial glucose regulation via knn meal classification in type 1 diabetes. *IEEE control systems letters*, 3(2):230–235, 2018.

[9] Surjeet Kumar Yadav and Saurabh Pal. Data mining: A prediction for performance improvement of engineering students using classification. *arXiv preprint arXiv:1203.3832*, 2012.

[10] Anuja Priyam, GR Abhijeeta, Anju Rathee, and Saurabh Srivastava. Comparative analysis of decision tree classification algorithms. *International Journal of current engineering and technology*, 3(2):334–337, 2013.

[11] Saba Bashir, Usman Qamar, Farhan Hassan Khan, and Lubna Naseem. Hmv: a medical decision support framework using multi-layer classifiers for disease prediction. *Journal of Computational Science*, 13:10–25, 2016.

[12] J. R. Quinlan. Introduction of decision tree. *Journal of Machine learning*, pages 81–106.

[13] Mr Brijain, R Patel, MR Kushik, and K Rana. A survey on decision tree algorithm for classification. 2014.

[14] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[15] Robert E Schapire. The boosting approach to machine learning: An overview. *Nonlinear estimation and classification*, pages 149–171, 2003.

[16] Nongyao Nai-Arun and Punnee Sittidech. Ensemble learning model for diabetes classification. In *Advanced Materials Research*, volume 931, pages 1427–1431. Trans Tech Publ, 2014.

[17] Weifeng Xu, Jianxin Zhang, Qiang Zhang, and Xiaopeng Wei. Risk prediction of type ii diabetes based on random forest model. In *2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, pages 382–386. IEEE, 2017.

[18] Wenqian Chen, Shuyu Chen, Hancui Zhang, and Tianshu Wu. A hybrid prediction model for type 2 diabetes using k-means and decision tree. In *2017 8th IEEE International conference on software engineering and service science (ICSESS)*, pages 386–390. IEEE, 2017.

[19] Gaganjot Kaur and Amit Chhabra. Improved j48 classification algorithm for the prediction of diabetes. *International journal of computer applications*, 98(22), 2014.

[20] Emrana Kabir Hashi, Md Shahid Uz Zaman, and Md Rokibul Hasan. An expert clinical decision support system to predict disease using classification techniques. In *2017 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pages 396–400. IEEE, 2017.

[21] J Pradeep Kandhasamy and SJPCS Balamurali. Performance analysis of classifier models to predict diabetes mellitus. *Procedia Computer Science*, 47:45–51, 2015.

[22] Tao Zheng, Wei Xie, Liling Xu, Xiaoying He, Ya Zhang, Mingrong You, Gong Yang, and You Chen. A machine learning-based framework to identify type 2 diabetes through electronic health records. *International journal of medical informatics*, 97:120–127, 2017.

[23] Zhen Sun, Songsen Yu, and Yang Zhang. An optimal decision tree model for diabetes diagnosis. In *2019 4th International Conference on Computational Intelligence and Applications (ICCIA)*, pages 83–87. IEEE, 2019.

[24] Xue-Hui Meng, Yi-Xiang Huang, Dong-Ping Rao, Qiu Zhang, and Qing Liu. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. *The Kaohsiung journal of medical sciences*, 29(2):93–99, 2013.

[25] Ayush Anand and Divya Shakti. Prediction of diabetes based on personal lifestyle indicators. In *2015 1st International Conference on Next Generation Computing Technologies (NGCT)*, pages 673–676. IEEE, 2015.

[26] K Saravananathan and T Velmurugan. Analyzing diabetic data using classification algorithms in data mining. *Indian Journal of Science and Technology*, 9(43):1–6, 2016.

[27] Quincy A Hathaway, Skyler M Roth, Mark V Pinti, Daniel C Sprando, Amina Kunovac, Andrya J Durr, Chris C Cook, Garrett K Fink, Tristen B Cheuvront, Jasmine H Grossman, et al. Machine-learning to stratify diabetic patients using novel cardiac biomarkers and integrative genomics. *Cardiovascular diabetology*, 18(1):1–16, 2019.

[28] Mehrbakhsh Nilashi, Othman bin Ibrahim, Hossein Ahmadi, and Leila Shahmoradi. An analytical method for diseases prediction using machine learning techniques. *Computers & Chemical Engineering*, 106:212–223, 2017.

[29] MT Mira Kania Sabariah, ST Aini Hanifa, and MT Siti Sa'adah. Early detection of type ii diabetes mellitus with random forest and classification and regression tree (cart). In *2014 International Conference of Advanced Informatics: Concept, Theory and Application (ICAICTA)*, pages 238–242. IEEE, 2014.

[30] Sajida Perveen, Muhammad Shahbaz, Aziz Guergachi, and Karim Keshavjee. Performance analysis of data mining classification techniques to predict diabetes. *Procedia Computer Science*, 82:115–121, 2016.

[31] Saba Bashir, Usman Qamar, Farhan Hassan Khan, and M Younus Javed. An efficient rule-based classification of diabetes using id3, c4. 5, & cart ensembles. In *2014 12th International Conference on Frontiers of Information Technology*, pages 226–231. IEEE, 2014.

[32] Bayu Adhi Tama and Kyung-Hyune Rhee. Tree-based classifier ensembles for early detection method of diabetes: an exploratory study. *Artificial Intelligence Review*, 51(3):355–370, 2019.

[33] Nongyao Nai-arun and Rungruttikarn Moungmai. Comparison of classifiers for the risk of diabetes prediction. *Procedia Computer Science*, 69:132–142, 2015.