A Comparative Analysis of Machine Learning Algorithms to Build a Predictive Model for Detecting Diabetes Complications

Ali A. Abaker
Department of Accounting Information Systems, Faculty of Computer Science
Al-Neelain University, Khartoum, Sudan
E-mail: aliabdallah@neelain.edu.sd

Fakhreldeen A. Saeed
Department of Software Engineering, Faculty of Computer Science, Al-Neelain University, Khartoum, Sudan
E-mail: fasaeed@neelain.edu.sd

Keywords: diabetes complications, machine learning, logistic regression, electronic health records, feature selection.

Received: April 4, 2020

Diabetes complications have a significant impact on patients' quality of life. The objective of this study was to predict which patients were more likely to be in a complicated health condition at the time of admission to allow for the early introduction of medical interventions. The data were 644 electronic health records from Alsukari Hospital collected from January 2018 to April 2019. We used the following machine learning methods: logistic regression, random forest, and k-nearest neighbor (KNN). The logistic regression algorithm performed better than the other algorithms achieving an accuracy of 81%, recall of 81%, and F1 score of 75%. Also, attributes such as infection years, swelling, diabetic ketoacidosis, and diabetic septic foot were significant in predicting diabetes complications. This model can be useful for the identification of patients requiring additional care to limit the complications and help practitioners in making decisions on whether the patient should be hospitalized or sent home. Furthermore, we used the sequential feature selection (SFS) algorithm which reduced the features to six, which is fewer than any model built before to predict diabetes complications. The primary goal of this study was achieved. The model had fewer attributes which means we have a simple and understandable model in addition to, it has a better performance.

Povzetek: Podana je analiza metod strojnega učenja za napovedovanje komplikacij pri sladkorni bolezni.

1 Introduction

Diabetes is a chronic disease and considered a serious global health challenge[1]. Diabetes Complications happen when diabetes is uncontrolled. That leads to serious health problems; the patient could suffer from a diabetic coma or die from a heart attack or stroke. The number of people with diabetes has risen to 422 million in 2014 which was about 17% of the world population[2]. More than 68% of diabetes-related deaths are caused by diabetes complications[3]. About 25% of people with diabetes are undiagnosed in the United States[4]. To address diabetic complications, doctors need to be allowed to identify and monitor patients at risk of complications[5]; [6]. Early discovery can prevent or delay diabetes-related complications and allow for effective intervention at both the individual and population levels, which are desperately needed to slow the diabetes epidemic[7].

The objective of this study was to predict which patients were more likely to be in a complicated health condition at the time of admission, which helps for the early introduction of medical interventions. We used the following machine learning methods: logistic regression, random forest, and KNN. The dataset was collected from Alsukari Hospital for building the machine learning model. The final model used the most six significant attributes and 644 records. This model is useful for predicting the likelihood that the patient should be hospitalized or sent home.

The rest of this paper is organized into Sections detailed as follows: section 2 literature review, data collection in section 3, feature selection in section 4, sequential feature selection section 5, model design in section 6, evaluation metrics section 7, results in section 8, section 9 conclusion, study limitations in section 10, last future work section 11.

2 Literature review

Various traditional methods, based on the physical and chemical tests are available for diagnosing diabetes. However, methods based on data mining techniques can be effectively implemented[8]. The authors agreed on the importance of developing machine learning algorithms to learn patterns and decision rules from data[9]. Although, many studies were conducted to assess the main causes of diabetes mellitus. But, a few were directed to discover the clinical risk factors[10]. A machine learning model was built to predict wound complications and mortality[11].

Electronic health records were used for many studies related to diabetes[12]; [13]. A method that enables risk assessment from electronic health records(EHR) on a large population, they also added administrative claims and pharmacy records[14]. Another study proposed a model that predicts the severity as a ratio interpreted as the impact of diabetes on different organs of the human body, the algorithm estimated the severity on different parts of the body like the heart and kidney[15]. A rapid model for glucose identification and prediction based on the idea of model migration[16]. Despite the data was collected from different sources and places but, some attributes were used in many studies such as age, gender, body mass index(BMI), glucose, blood pressure, time of diagnosis, and smoking [9]; [17]; [18]; [11].

Algorithms	2015 - 2019
ANN	[8],[19],[2],[20],[21],[11]
K-means	[22]
Logistic	[22],[9],[2],[20],[23],[24],[11]
Decision Tree	[1],[20],[3],[25],[24], [10],[26]
SVM	[27],[2],[1],[20],[3],[21],[24]
KNN	[1],[21],[24]
Random Forest	[1],[24]
Naive Bayes	[3],[23],[25],[24]
Designed algorithms	[28],[13],[5],[29],[9]

Table 1: Shows the trends of used algorithms in previous literature.

Table 1 above is the review of the popular algorithms used in diabetes studies related to machine learning from 20015 to 2019. And the figure below demonstrates the trend.

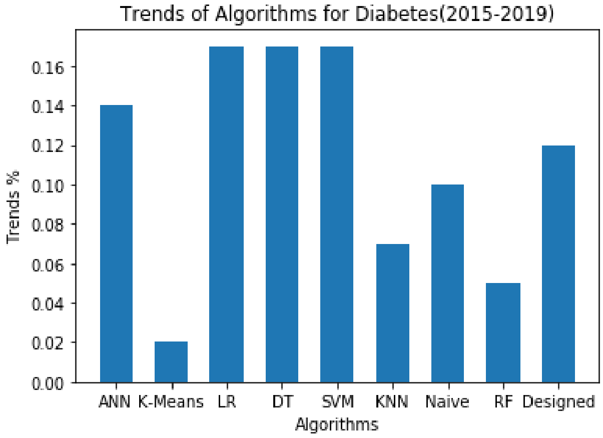

Figure 1: Shows the trend of algorithms used in the last five years.

According to the figure above, logistic regression (LR), decision tree (DT), and supervised machine learning (SVM) are the most popular algorithms used for diabetes studies among the research community with about 17%. Followed by, the artificial neural network (ANN) with 14%. According to, literature review researchers designed about 12% out of all used algorithms as new methods for diabetes problems.

We built a machine learning model to predict diabetes complications, using six attributes. This model introduced new attributes such as diabetic ketoacidosis, swelling, infection years, and diabetic septic foot were found to be significant. These attributes were not included in the previously mentioned studies[30]; [13]; [31]; [9]. Also, in this paper, we investigated five performance metrics such as F1 and recall.

A logistic regression model was used to assess the factors associated with glycemic control. The model indicated that patients older than 65 years old were more likely to have complications compared to the younger[32]. Demographic and treatment data were collected and logistic regression was used to predict complications[33]. Another study was conducted to predict 30-day complication rate using random forest and logistic regression, the analysis showed that age is the most significant attribute in predicting complications[34]. Random forest and simple logistic regression methods showed the best performance compared to the evaluated algorithms[35]. Diabetes complications prediction model was based on similarity measure. first, they assessed the similarity between textual medical records after data cleaning, then topic mining is conducted, and last building the model[36].

3 Data collection

The dataset was collected from Alsukari Hospital. Ethical approval to use the data for research was obtained both from the Ministry of Health (MOH) and the hospital. The dataset contained 29 attributes and 644 records of diagnosed diabetes patients who were admitted to the hospital in the period from January 2018 to April 2019.

4 Feature selection

Classification problems usually have a big number of features in the dataset, but not all of them are significant for classification. Irrelevant features may reduce the performance and even complicate the model. Feature selection aims to select a small number of relevant features to get similar or better classification performance than using a larger number of features[39].

Thus, it is usual to apply a preprocessing step to remove irrelevant features and reduce the dimensionality of the data[40]. The selection of the features can lead to an improvement of the learning algorithm, either in terms of learning speed, generalization, or simplicity of the model. Furthermore, there are other advantages associated with a reduced feature: low cost, clear model, and a better understanding of the domain knowledge[41].

The figure below illustrates the five approaches for feature selection.

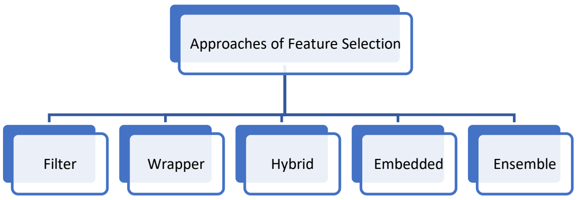

Figure 2: Shows feature selection approaches.

Attribute	Type	Scale	Description
Infection years	Continuous	0 – 35	Is the period from the patient was diagnosed with Diabetes
Sugar	Continuous	53 - 681	Blood sugar or glucose level in the body is measured in mg/dL or milligrams per deciliter.
Swelling	Category	Binary	Swollen body part.
DKA	Category	Binary	Diabetic ketoacidosis called [22]ketones.
DSF	Category	Binary	Diabetic septic Foot is a foot affected by ulceration of the lower limb in a patient with diabetes[37].
HR	Continuous	27 - 139	Is the speed of the heartbeat measured by the number of contractions (beats) of the heart per minute (bpm)[38].
Class	Binary	Binary	The Target variable.

Table 2: Shows the selected attributes and their descriptions to build the model.

4.1 Filter method

Filter based feature methods evaluate features as an individual assessment. Therefore, these methods first assign a score value for each feature using one of the statistical criteria, and then, all features are sorted according to the scores. Then, they select top-n features with the highest score as the final step[42]. Filter feature selection algorithms are useful due to their simplicity and fast speed. A common filter is to use mutual information to evaluate the relationships between each feature and the class variable[43].

4.2 Wrapper method

Feature selection algorithms are divided into two methods. Firstly, it depends on the outcome of the selection algorithm: whether it returns a subset of significant features or an ordered ranking of all the features, recognized as feature ranking. Secondly, feature selection methods are divided into three approaches based on the relationship between a feature selection algorithm and the learning method, which is used for building the model: filters, which rely on overall characteristics of the dataset and are independent of the learning algorithm; wrappers, which use the prediction of a classifier to estimate subsets of features; and embedded methods, which perform the selection in the process of training and are specific to the given learning algorithm.

Wrapper techniques depend on the classification algorithm, which is used to evaluate the subsets of features, but they are more expensive when it comes to computation[44]. Despite this weakness, they often provide better outcomes; wrappers are used widely in many applications[45]. Especially, in healthcare where we care about the accuracy more than the performance of the algorithm in many situations and allow for implementation in real-time systems when we have fewer attributes[46].

4.3 Hybrid method

Recently, researchers are concentrating on developing novel hybrid feature selection methods as they speed up the removal of irrelevant features and give greater classification accuracy compared to other methods[47]. Though various techniques were developed for selecting the perfect subset of features, these methods faced some problems such as instability, high processing time, and selecting a semi-optimal solution as a final result. In other words, they have not been able to fully extract the effective features. Hybrid methods were introduced as a solution to overcome the weaknesses of using a single algorithm[48].

4.4 Embedded method

Embedded feature selection is related to classification algorithms. This relation in embedded methods is stronger than that in wrapper methods. Embedded methods are a sort of combination of filter and wrapper methods[49]. By, embedding feature selection into the model learning. They return both the learned model and selected features and are frequently used for classification[50]. Inserting the feature selection step into the training process can improve the performance of the model.

4.5 Ensemble method

Ensemble learning is an effective method for machine learning. The objective is to attain better learning accuracy by combining different learning models[41]. Ensemble methods are better than using a single machine learning model. Recently, the development of ensemble feature selection is increasingly getting attention[51].

Combined feature selection aims to find multiple optimal features. The advantage of integration technology would produce a stable and efficient method; especially, with high-dimensional data[52].

5 Sequential feature selection (SFS)

The sequential feature selection(SFS) algorithm begins with a blank set and increases one feature for the first step which gives the best value for the model. On the second step onwards the remaining features are added separately to the existing subset and the new subset is assessed. The new feature is permanently added to the subset if it gives the maximum classification accuracy. The process is repeated until the required number of features are added[53]. SFS method has the advantage of improving the prediction performance of the classifier by excluding any characteristic that reduces the performance[54].

6 Model design

Working with methods for reasoning under uncertainty is now one of the most interesting areas of machine learning [20]; [21]. Machine learning has been used for several decades to tackle a broad range of problems in many fields of applications[57]. The machine learning model was built on the historical data using different algorithms for each model; we evaluated the results, and assessed the model accuracy on the testing data. Six attributes were selected out of 29. The dataset was divided into two parts: training and testing set consisting of 70 and 30 percent respectively. The features were selected to be used in the final model based on the training set which achieved the highest accuracy of 86%, with the six attributes as shown in table 2. Logistic regression, random forest, and KNN were selected because of their simplicity and good predictive capability. However, machine learning models are more accurate than normal statistical methods[58].

6.1 Logistic regression

Logistic regression models are a sort of widespread linear model that is used for datasets where the dependent variable is categorical[59]. The logistic model is used to estimate the probability of the response variable based on one or more predictor variables. Logistic regression is used when the dependent variable is categorical and generates output in terms of probabilities[60]. Logistic regression is an effective prediction algorithm. Its applications are efficient when the dependent variable of a dataset is binary[61].

$$f(x) = \frac{L}{1+e^{-k(x-x_0)}} \qquad (1)$$

Where

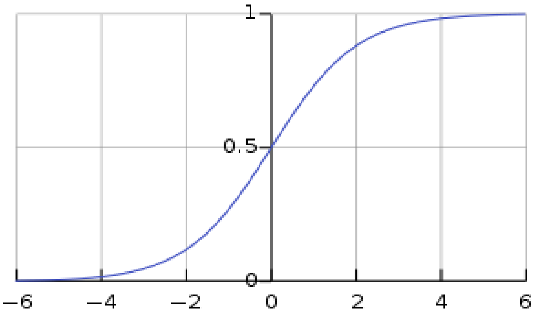

Figure 3: Shows logistic regression curve.

X_0 = the x value of the sigmoid's midpoint.

L = the curve's maximum value.

K = the logistic growth rate or steepness of the curve.

6.2 Random forest

Random Forest was proposed by Dr. Breiman in 2001. It is typically used in classification[62]. It is an algorithm based on statistical learning theory, which uses a bootstrap randomized re-sampling method to extract multiple versions of the sample sets from the original training datasets. Then it builds a decision tree model for each sample set, and finally combines all the results of the decision trees to predict via a voting mechanism[63].

Suppose we have the dataset D = {(x1, y1) … (xn, yn)} and the aim is to find the function f : X →Y where X is the inputs and Y is the produced outputs. Furthermore, let M be the number of inputs. Random forest randomly selects n observations from D with replacement to a bootstrap sample. Each tree is grown using a subset of m features from the overall M features. For regression, it is recommended to set the subset of features at M=3. Then at each node, m features are nominated at random and the best performing split among the M features is selected according to the impurity measure (Gini impurity). The trees are grown to a maximum depth without pruning[64].

6.3 K-nearest neighbor (KNN)

It is a classification algorithm that classifies data based on similarity measure or distance measure[18]. This algorithm can be used in both classification and regression problems[21], [65]. KNN classifies an instance by finding its nearest neighbors[66]. The KNN classifier applies the Euclidean distance or cosine similarity for differentiating the training tuple and test tuples. The same Euclidean distance between tuples Xi and Xt (t = 1,2,3 …n) can be explained as[67]:

$$d(yi,yt) = \sqrt{(y_{i1} - y_{t1})^2 + (x_{i2} - x_{t2})^2 + \cdots + (x_{is} - x_{ts})^2} \qquad (2)$$

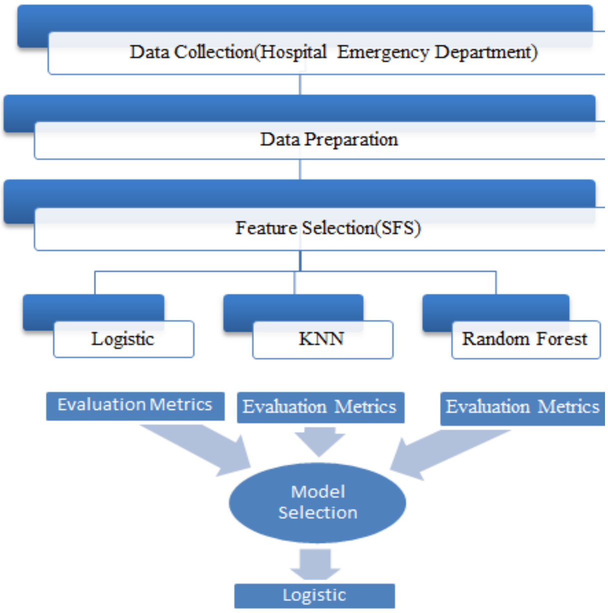

Figure 4: Shows the model selection process.

where yi, n and s be the tuple constant. It can be written as:

$$Dist(y1, y2 = \sqrt{\sum_{i=1}^{n}(y_{1i} - y_{2i})^2} \qquad (3)$$

This equation is prepared based on KNN algorithms, every neighboring point that is closest to the test tuple, which is encapsulated and on the nearby space to the test tuple[68].

7 Evaluation metrics

The three algorithms, Logistic Regression, Random Forest Classifier, and KNN were compared in terms of accuracy, recall, specificity, precision, and F1scores as demonstrated below. The level of efficiency of the classification model is measured with the number of correct and incorrect classifications in each potential value of the variables being classified. From the outcomes gained. The following equations are used to measure the Accuracy, Sensitivity, and Specificity, Precision, and F1 score [69].

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad (4)$$

Recall or also known as sensitivity refers to the percentage of total relevant results correctly classified by your algorithm.

$$\text{Recall} = \frac{TP}{TP+FN} \qquad (5)$$

Specificity is defined as the proportion of actual negatives, which got predicted as the negative (or true negative).

$$\text{Specificity} = \frac{TN}{TN+FP} \qquad (6)$$

The precision of all the records we predicted positive.

Algorithm/ Metrics	Accuracy	Recall	Specificity	Precision	F1
Logistic Regression	81%	81%	81%	70%	75%
Random Forest Classifier	78%	57%	89%	74%	64%
KNN	76%	62%	84%	68%	65%

Table 3: Shows the three algorithms and their metrics scores using the default threshold.

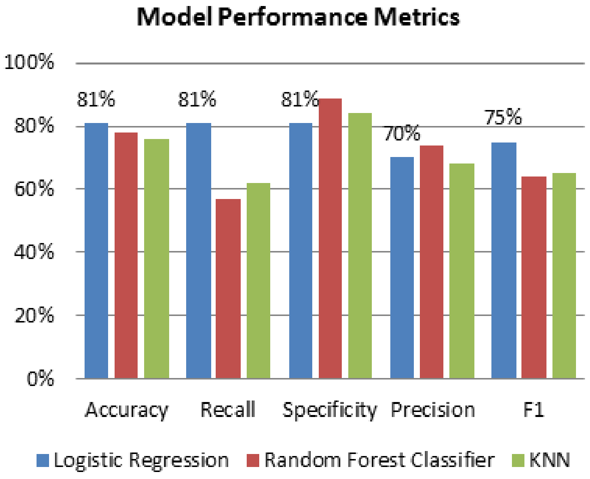

Figure 5: Shows model performance metrics.

$$\text{Precision} = \frac{TP}{TP+FP} \qquad (7)$$

F1-score, which is simply the harmonic mean of precision and recall.

$$\text{F1 Score} = \frac{2TP+TN}{2TP+FP+FN} \qquad (8)$$

From table 3 above the logistic regression achieved the highest accuracy of 81%. According to the medical objective, the model was designed to be more sensitive in predicting true positive which was calculated by recall and F1 of 81 and 79 percent respectively. As this model is used in healthcare our interest in predicting the positive class, which is more important. It is acceptable for the model to fall into a false positive error (type 1 error). But it is very costly for the model to commit the (false negative error) type 2 error, it means that the patient might be in complicated health status and needs special and immediate medical care and intervention, but the model tells us that the patient will not be in a complicated situation. And that could result in more health complications and even the patient's life.

8 Results

New attributes were found to be significant in predicting diabetes complications such as infection years, swelling, diabetic ketoacidosis, diabetic septic foot, which were found to be vital in predicting diabetes complications.

But, blood pressure, body temperature, cholesterol, protein level, oxygen level, and gender were not significant in predicting diabetes complications because they were excluded by the feature selection algorithm. We compared the following three algorithms: logistic regression, random forest classifier, and KNN to find the best algorithm for predicting diabetes complications. Accuracy, recall, specificity, precision, and F1; used as performance metrics. These metrics were used for selecting the best model. The logistic regression algorithm achieved the highest recall score of 81%, followed by KNN and random forest of 62% and 57% respectively, as shown in Table 3 above. This means that the model is more sensitive in predicting the positive class. Also, the model achieved the highest F1score of 75% as shown in Table 3 above. The model was designed to be more sensitive in predicting true positive class which means the diabetes status is complicated. It was calculated by the recall score of 81%. As this model is used in healthcare our interest in predicting the positive class, which is significant.

9 Conclusion

In this paper, we created a dataset from Alsukari Hospital for building our machine learning model. The best model was built using six out of 29 attributes. Three algorithms were compared in selecting the best model as follows: Logistic regression, random forest, and KNN. Furthermore, new attributes were investigated and included in the model. Finally, the best accuracy was obtained using logistic regression. The overall accuracy does not guarantee that the model will perform better and serve the specific domain interest. According to the medical objective, the recall score was used besides general accuracy. The higher recall score indicates that the model is more sensitive in predicting positive cases or medically patients with diabetes complications.

10 Study limitations

The accuracy is not very high as we are working in the medical field higher accuracy is needed. Second, the size of the dataset is small; machine learning models need more data for producing stable and well-trained models.

11 Future work

There are several future research directions. Firstly, for predicting diabetes complications more features could be included which were not included in this study. Secondly, more work should be directed toward identifying the risk factors associated with diabetes complications. Last we are interested in adopting this model to other chronic diseases.

12 Acknowledgement

The authors would like to thank the Ministry of Health (MOH) and Alsukari Hospital for giving us the ethical approval to use the data for this work. Also, we would like to take this opportunity to thank the reviewers for their constructive comments towards improving our manuscript.

13 Conflict of interest

The authors declare no conflict of interest.

References

[1] J. P. Kandhasamy and S. Balamurali, “Performance Analysis of Classifier Models to Predict Diabetes Mellitus,” *Procedia - Procedia Comput. Sci.*, vol. 47, pp. 45–51, 2015. https://doi.org/10.1016/j.procs.2015.03.182

[2] S. Malik, R. Khadgawat, S. Anand, and S. Gupta, “Non-invasive detection of fasting blood glucose level via electrochemical measurement of saliva,” *SpringerPlus*, vol. 5, no. 1. 2016. https://doi.org/10.1186/s40064-016-2339-6

[3] D. Sisodia and D. S. Sisodia, “ScienceDirect Prediction of Diabetes using Classification Algorithms,” *Procedia Comput. Sci.*, vol. 132, no. Iccids, pp. 1578–1585, 2018. https://doi.org/10.1016/j.procs.2018.05.122

[4] A. E. Anderson, W. T. Kerr, A. Thames, T. Li, J. Xiao, and M. S. Cohen, “Electronic health record phenotyping improves detection and screening of type 2 diabetes in the general United States population: A cross-sectional, unselected, retrospective study,” *J. Biomed. Inform.*, vol. 60, no. December, pp. 162–168, 2016. https://doi.org/10.1016/j.jbi.2015.12.006

[5] A. Anand and D. Shakti, “Prediction of diabetes based on personal lifestyle indicators,” *Proc. 2015 1st Int. Conf. Next Gener. Comput. Technol. NGCT 2015*, no. September, pp. 673–676, 2016. https://doi.org/10.1109/NGCT.2015.7375206

[6] G. Peddinti *et al.*, “Early metabolic markers identify potential targets for the prevention of type 2 diabetes,” *Diabetologia*, vol. 60, no. 9, pp. 1740–1750, 2017. https://doi.org/10.1007/s00125-017-4325-0

[7] T. P. A. Debray, Y. Vergouwe, H. Koffijberg, D. Nieboer, E. W. Steyerberg, and K. G. M. Moons, “ORIGINAL ARTICLES A new framework to enhance the interpretation of external validation studies of clinical prediction models,” *J. Clin. Epidemiol.*, vol. 68, no. 3, pp. 279–289, 2015. https://doi.org/10.1016/j.jclinepi.2014.06.018

[8] M. Komi, J. Li, Y. Zhai, and Z. Xianguo, “Application of data mining methods in diabetes prediction,” *2017 2nd Int. Conf. Image, Vis. Comput. ICIVC 2017*, no. S Ix, pp. 1006–1010, 2017. https://doi.org/10.1109/ICIVC.2017.7984706.

[9] A. Dagliati *et al.*, “Machine Learning Methods to Predict Diabetes Complications,” 2017. https://doi.org/doi: 10.1177/1932296817706375

[10] Purushottam, K. Saxena, and R. Sharma, “Diabetes mellitus prediction system evaluation using C4.5 rules and partial tree,” *2015 4th Int. Conf. Reliab.*

Infocom Technol. Optim. Trends Futur. Dir. ICRITO 2015, pp. 1–6, 2015. https://doi.org/10.1109/ICRITO.2015.7359272

[11] J. S. Kim *et al.*, "Examining the Ability of Artificial Neural Networks Machine Learning Models to Accurately Predict Complications Following Posterior Lumbar Spine Fusion," *Spine (Phila. Pa. 1976).*, vol. 43, no. 12, pp. 853–860, 2018. https://doi.org/10.1097/BRS.0000000000002442

[12] M. Kumar, N. K. Rath, A. Swain, and S. K. Rath, "Feature Selection and Classification of Microarray Data using MapReduce based ANOVA and K-Nearest Neighbor," *Procedia Comput. Sci.*, vol. 54, pp. 301–310, 2015. https://doi.org/10.1016/j.procs.2015.06.035

[13] B. Liu, Y. Li, Z. Sun, S. Ghosh, and K. Ng, "Early Prediction of Diabetes Complications from Electronic Health Records : A Multi-Task Survival Analysis Approach," pp. 101–108.

[14] N. Razavian, S. Blecker, A. M. Schmidt, A. Smith-mclallen, S. Nigam, and D. Sontag, "Population-Level Prediction of Type 2 Diabetes From Claims Data and Analysis of Risk Factors," vol. 3, no. 4, 2015. https://doi.org/10.1089/big.2015.0020

[15] V. R. Balpande and R. D. Wajgi, "Prediction and severity estimation of diabetes using data mining technique," *IEEE Int. Conf. Innov. Mech. Ind. Appl. ICIMIA 2017 - Proc.*, no. Icimia, pp. 576–580, 2017. https://doi.org/10.1109/ICIMIA.2017.7975526

[16] C. Zhao and C. Yu, "Rapid model identification for online subcutaneous glucose concentration prediction for new subjects with type i diabetes," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 5, pp. 1333–1344, 2015. https://doi.org/10.1109/TBME.2014.2387293

[17] O. Geman, I. Chiuchisan, and R. Toderean, "Application of Adaptive Neuro-Fuzzy Inference System for diabetes classification and prediction," *2017 E-Health Bioeng. Conf. EHB 2017*, no. Dm, pp. 639–642, 2017. https://doi.org/10.1109/EHB.2017.7995505

[18] N. Sneha and T. Gangil, "Analysis of diabetes mellitus for early prediction using optimal features selection," *J. Big Data*, vol. 6, no. 1, 2019. https://doi.org/10.1186/s40537-019-0175-6

[19] S. Joshi and M. Borse, "Detection and prediction of diabetes mellitus using back-propagation neural network," *Proc. - 2016 Int. Conf. Micro-Electronics Telecommun. Eng. ICMETE 2016*, pp. 110–113, 2016. https://doi.org/10.1109/ICMETE.2016.11

[20] H. Y. Tsao, P. Y. Chan, and E. C. Y. Su, "Predicting diabetic retinopathy and identifying interpretable biomedical features using machine learning algorithms," *BMC Bioinformatics*, vol. 19, no. Suppl 9, 2018. https://doi.org/10.1186/s12859-018-2277-0

[21] H. Kaur and V. Kumari, "Predictive modelling and analytics for diabetes using a machine learning approach," *Appl. Comput. Informatics*, no. December, 2019. https://doi.org/10.1016/j.aci.2018.12.004

[22] 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, pp. 100–107, 2018. https://doi.org/10.1016/j.imu.2017.12.006

[23] B. J. Lee and J. Y. Kim, "Identification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on Machine Learning," *IEEE J. Biomed. Heal. Informatics*, vol. 20, no. 1, pp. 39–46, 2016. https://doi.org/10.1109/JBHI.2015.2396520

[24] T. Zheng *et al.*, "A Machine Learning-based Framework to Identify Type 2 Diabetes through Electronic Health Records," *Int. J. Med. Inform.*, 2016. https://doi.org/10.1016/j.ijmedinf.2016.09.014

[25] P. Songthung and K. Sripanidkulchai, "Improving type 2 diabetes mellitus risk prediction using classification," *2016 13th Int. Jt. Conf. Comput. Sci. Softw. Eng. JCSSE 2016*, 2016. https://doi.org/10.1109/JCSSE.2016.7748866

[26] S. Perveen, M. Shahbaz, A. Guergachi, and K. Keshavjee, "Performance Analysis of Data Mining Classification Techniques to Predict Diabetes," *Procedia Comput. Sci.*, vol. 82, no. March, pp. 115–121, 2016. https://doi.org/10.1016/j.procs.2016.04.016

[27] J. Zhang *et al.*, "Diagnostic Method of Diabetes Based on Support Vector Machine and Tongue Images," *Biomed Res. Int.*, vol. 2017, 2017. https://doi.org/10.1155/2017/7961494

[28] J. Li *et al.*, "Feature selection: A data perspective," *ACM Comput. Surv.*, vol. 50, no. 6, 2017. https://doi.org/10.1145/3136625

[29] K. Zarkogianni, M. Athanasiou, and A. C. Thanopoulou, "Comparison of Machine Learning Approaches Toward Assessing the Risk of Developing Cardiovascular Disease as a Long-Term Diabetes Complication," *IEEE J. Biomed. Heal. Informatics*, vol. 22, no. 5, pp. 1637–1647, 2018. https://doi.org/10.1109/JBHI.2017.2765639

[30] L. Liu, "Forecasting Potential Diabetes Complications," 2014.

[31] G. Huzooree, "Glucose Prediction Data Analytics for Diabetic Patients Monitoring," no. i, 2017. https://doi.org/10.1109/NEXTCOMP.2017.8016197

[32] M. Almetwazi *et al.*, "Factors associated with glycemic control in type 2 diabetic patients in Saudi Arabia," *Saudi Pharm. J.*, vol. 27, no. 3, pp. 384–388, 2019. https://doi.org/10.1016/j.jsps.2018.12.007

[33] M. YimamAhmed, S. H. Ejigu, A. Z. Zeleke, and M. Y. Hassen, "Glycemic control, diabetes complications and their determinants among ambulatory diabetes mellitus patients in southwest ethiopia: A prospective cross-sectional study,"

Diabetes, Metab. Syndr. Obes. Targets Ther., vol. 13, pp. 1089–1095, 2020. https://doi.org/10.2147/DMSO.S227664

[34] B. N. Armstrong, A. Renson, L. C. Zhao, and M. A. Bjurlin, "Development of novel prognostic models for predicting complications of urethroplasty," *World J. Urol.*, vol. 37, no. 3, pp. 553–559, 2019. https://doi.org/10.1007/s00345-018-2413-5

[35] V. Rodriguez-Romero, R. F. Bergstrom, B. S. Decker, G. Lahu, M. Vakilynejad, and R. R. Bies, "Prediction of Nephropathy in Type 2 Diabetes: An Analysis of the ACCORD Trial Applying Machine Learning Techniques," *Clin. Transl. Sci.*, vol. 12, no. 5, pp. 519–528, 2019. https://doi.org/10.1111/cts.12647

[36] S. Ding, Z. Li, X. Liu, H. Huang, and S. Yang, "Diabetic complication prediction using a similarity-enhanced latent Dirichlet allocation model," *Inf. Sci. (Ny).*, vol. 499, pp. 12–24, 2019. https://doi.org/10.1016/j.ins.2019.05.037

[37] K. Alexiadou and J. Doupis, "Management of diabetic foot ulcers," *Diabetes Ther.*, vol. 3, no. 1, pp. 1–15, 2012. https://doi.org/10.1007/s13300-012-0004-9

[38] Y. J. van de Vegte, B. S. Tegegne, N. Verweij, H. Snieder, and P. van der Harst, "Genetics and the heart rate response to exercise," *Cell. Mol. Life Sci.*, no. 123456789, 2019. https://doi.org/10.1007/s00018-019-03079-4

[39] B. Xue, M. Zhang, S. Member, and W. N. Browne, "Particle Swarm Optimization for Feature Selection in Classification : A Multi-Objective Approach," *Ieee Trans. Cybern.*, pp. 1–16, 2012. https://doi.org/10.1109/TSMCB.2012.2227469

[40] M. A. Sulaiman and J. Labadin, "Feature selection based on mutual information for machine learning prediction of petroleum reservoir properties," *2015 9th Int. Conf. IT Asia Transform. Big Data into Knowledge, CITA 2015 - Proc.*, pp. 2–7, 2015. https://doi.org/10.1109/CITA.2015.7349827

[41] V. Bolón-Canedo and A. Alonso-Betanzos, "Ensembles for feature selection: A review and future trends," *Inf. Fusion*, vol. 52, pp. 1–12, 2019. https://doi.org/10.1016/j.inffus.2018.11.008

[42] R. Cekik and A. K. Uysal, "A novel filter feature selection method using rough set for short text data," *Expert Syst. Appl.*, vol. 160, p. 113691, 2020. https://doi.org/10.1016/j.eswa.2020.113691

[43] E. Hancer, B. Xue, and M. Zhang, "Differential evolution for filter feature selection based on information theory and feature ranking," *Knowledge-Based Syst.*, vol. 140, pp. 103–119, 2018. https://doi.org/10.1016/j.knosys.2017.10.028

[44] M. Monirul Kabir, M. Monirul Islam, and K. Murase, "A new wrapper feature selection approach using neural network," *Neurocomputing*, vol. 73, no. 16–18, pp. 3273–3283, 2010. https://doi.org/10.1016/j.neucom.2010.04.003

[45] V. F. Rodriguez-Galiano, J. A. Luque-Espinar, M. Chica-Olmo, and M. P. Mendes, "Feature selection approaches for predictive modelling of groundwater nitrate pollution: An evaluation of filters, embedded and wrapper methods," *Sci. Total Environ.*, vol. 624, pp. 661–672, 2018. https://doi.org/10.1016/j.scitotenv.2017.12.152

[46] J. González, J. Ortega, M. Damas, P. Martín-Smith, and J. Q. Gan, "A new multi-objective wrapper method for feature selection – Accuracy and stability analysis for BCI," *Neurocomputing*, vol. 333, pp. 407–418, 2019. https://doi.org/10.1016/j.neucom.2019.01.017

[47] D. Jain and V. Singh, "Feature selection and classification systems for chronic disease prediction: A review," *Egypt. Informatics J.*, vol. 19, no. 3, pp. 179–189, 2018. https://doi.org/10.1016/j.eij.2018.03.002

[48] J. Pirgazi, M. Alimoradi, T. Esmaeili Abharian, and M. H. Olyaee, "An Efficient hybrid filter-wrapper metaheuristic-based gene selection method for high dimensional datasets," *Sci. Rep.*, vol. 9, no. 1, pp. 1–15, 2019. https://doi.org/10.1038/s41598-019-54987-1

[49] H. Liu, S. Member, M. Zhou, I. Qing, and G. Liu, "An Embedded Feature Selection Method for Imbalanced Data Classification," *IEEE/CAA J. Autom. Sin.*, vol. PP, pp. 1–13. https://doi.org/10.1109/JAS.2019.1911447

[50] M. Lu, "Embedded feature selection accounting for unknown data heterogeneity," *Expert Syst. Appl.*, vol. 119, pp. 350–361, 2019. https://doi.org/10.1016/j.eswa.2018.11.006

[51] R. Saifan, K. Sharif, M. Abu-Ghazaleh, and M. Abdel-Majeed, "Investigating algorithmic stock market trading using ensemble machine learning methods," *Inform.*, vol. 44, no. 3, pp. 311–325, 2020. https://doi.org/10.31449/INF.V44I3.2904

[52] J. Wang, J. Xu, C. Zhao, Y. Peng, and H. Wang, "An ensemble feature selection method for high-dimensional data based on sort aggregation," *Syst. Sci. Control Eng.*, vol. 7, no. 2, pp. 32–39, 2019. https://doi.org/10.1080/21642583.2019.1620658

[53] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Comput. Electr. Eng.*, vol. 40, no. 1, pp. 16–28, 2014. https://doi.org/10.1016/j.compeleceng.2013.11.024

[54] J. Lee, D. Park, and C. Lee, "Feature selection algorithm for intrusions detection system using sequential forward search and random forest classifier," *KSII Trans. Internet Inf. Syst.*, vol. 11, no. 10, pp. 5132–5148, 2017. https://doi.org/10.3837/tiis.2017.10.024

[55] O. F.Y, A. J.E.T, A. O, H. J. O, O. O, and A. J, "Supervised Machine Learning Algorithms: Classification and Comparison," *Int. J. Comput. Trends Technol.*, vol. 48, no. 3, pp. 128–138, 2017. https://doi.org/10.14445/22312803/ijctt-v48p126

[56] B. J. Frey, S. Member, and N. Jojic, "freyJojicTutorial_pami_sep05.pdf," vol. 27, no. 9, pp. 1392–1416, 2005.

[57] C. M. Bishop, "Model-based machine learning Author for correspondence :," *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, vol. 371, no. 1984, p. 20120222, 2013.

[58] M. M. Churpek, T. C. Yuen, C. Winslow, D. O. Meltzer, M. W. Kattan, and D. P. Edelson, "Multicenter Comparison of Machine Learning Methods and Conventional Regression for Predicting Clinical Deterioration on the Wards," *Crit. Care Med.*, vol. 44, no. 2, pp. 368–374, 2016. https://doi.org/10.1097/CCM.0000000000001571

[59] B. Heung, H. C. Ho, J. Zhang, A. Knudby, C. E. Bulmer, and M. G. Schmidt, "An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping," *Geoderma*, vol. 265, pp. 62–77, 2016. https://doi.org/10.1016/j.geoderma.2015.11.014

[60] M. Maniruzzaman *et al.*, "Accurate Diabetes Risk Stratification Using Machine Learning: Role of Missing Value and Outliers," *J. Med. Syst.*, vol. 42, no. 5, pp. 1–17, 2018. https://doi.org/10.1007/s10916-018-0940-7

[61] C. Zhu, C. U. Idemudia, and W. Feng, "Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques," *Informatics Med. Unlocked*, vol. 17, no. January, p. 100179, 2019. https://doi.org/10.1016/j.imu.2019.100179

[62] M. Jena and S. Dehuri, "Decision tree for classification and regression: A state-of-the art review," *Inform.*, vol. 44, no. 4, pp. 405–420, 2020. https://doi.org/10.31449/INF.V44I4.3023

[63] W. Xu, J. Zhang, Q. Zhang, and X. Wei, "Risk prediction of type II diabetes based on random forest model," 2017. https://doi.org/10.1109/AEEICB.2017.7972337

[64] B. Baba, "Borsa _ Istanbul Review Predicting IPO initial returns using random forest," 2020. https://doi.org/10.1016/j.bir.2019.08.001

[65] D. Panda, S. R. Dash, R. Ray, and S. Parida, "Predicting the causal effect relationship between copd and cardio vascular diseases," *Inform.*, vol. 44, no. 4, pp. 447–457, 2020. https://doi.org/10.31449/INF.V44I4.3088

[66] K. Saxena, Z. Khan, and S. Singh, "Diagnosis of Diabetes Mellitus using K Nearest Neighbor Algorithm," vol. 2, no. 4, pp. 36–43, 2014.

[67] G. G. Várkonyi and A. Gradišek, "Data protection impact assessment case study for a research project using artificial intelligence on patient data," *Inform.*, vol. 44, no. 4, pp. 497–505, 2020. https://doi.org/10.31449/INF.V44I4.3253

[68] S. K. Nayak, M. Panda, and G. Palai, "Realization of optical ADDER circuit using photonic structure and KNN algorithm," *Optik (Stuttg).*, vol. 212, no. March, p. 164675, 2020. https://doi.org/10.1016/j.ijleo.2020.164675

[69] N. Nai-Arun and R. Moungmai, "Comparison of Classifiers for the Risk of Diabetes Prediction," *Procedia Comput. Sci.*, vol. 69, pp. 132–142, 2015. https://doi.org/10.1016/j.procs.2015.10.014