Pre-Print Version

Risk Assessment of Diabetes Mellitus Dataset Using Supervised Machine Learning Algorithms.

Nehl Roop[1], Munaf Rashid[2], Sidra Abid Syed[3], Fahad Shamim[4] ,Sarmad Shams[5*] , Shahzad Nasim[6]

Abstract: Diabetes mellitus is a chronic condition that can lead to serious health complications if not properly managed and this research paper focuses on its early diagnosis and risk assessment Machine Learning Support Vector and Machine Learning Random Forest are the two algorithms being used in this study to provide a comparative analysis of their predictive accuracies and efficiency. The research is conducted using a multivariate dataset, consisting of 520 instances and 17 attributes, obtained from the UCI Repository (Machine Learning). After thorough analysis, it is found that both SVM and Random Forest algorithms performs well in predicting diabetes mellitus risk. However, comparison of the accuracies of both algorithms shows that the RF classifier yielded greater accuracy and provided the most suitable output. This study is an effective demonstration of the importance and effectiveness for the early diagnosis by utilizing artificially learned patterns and risk assessment of diseases like diabetes mellitus. The findings also highlight the significance of comparative analysis to identify the most accurate and efficient algorithm for a given dataset.

***Keywords*: Early diagnosis, Risk assessment, Diabetes mellitus, SVM, Random Forest, Multivariate dataset, UCI Machine Learning Repository.**

INTRODUCTION:

The prevalent disorder of metabolism, Diabetes mellitus may results in severe hyperglycemia and poses a significant burden on healthcare globally [1]. It is of two main types: type 1, due to the destruction of beta cells directing to universal insufficiency of insulin, and Type 2, primary insulin objection with respective insulin insufficiency to convincing malfunctioning discharge [2, 3]. Other types of diabetes include diseases of the exocrine pancreas, endocrinopathies, β-cell function known as insulin action by inducing drug for genetic defects, contamination, abnormal formations of autoimmune-mediated diabetes, syndromes caused by genes that can be related with diabetes [4].

[1-2]Ziauddin University Karachi
[3]Sir Syed University of Engineering & Technology
[4-5]Liaquat University of Medical & Health Sciences
[6]The Begum Nusrat Bhutto Women University, Sukkur
Country : Pakistan,
Email: *sarmad.shams@lumhs.edu.pk

Type 1 diabetes is among the most common type in adults and is also known as insulin-dependent, whereas type 2 diabetes is called insulin-independent and is typically found in people aged 40 and above with risk factors such as obesity or a sedentary lifestyle [5, 6]. Early prediction of diabetes is important in maintaining the health of affected individuals, as it can help diagnose the disease and its associated complications, it will also reduce risk factors and the likelihood of developing severe complications and diseases like heart related problems, stroke, loss of sight, and organ failure among others [7-9].

Ongoing biomedical research is providing new insights into the mechanisms of diabetes development, which may lead to the development of recent diagnostic and therapeutic approaches [10]. Understanding the genetic and environmental factors that contribute to diabetes development is essential in preventing and managing this disease effectively.

To assist the prior onset of diabetes, the use of AI techniques may offer great benefits and therefore the findings are based on comparisons from two supervised algorithms: The Random Forest Classifier (RFC) and the Support Vector Machine (SVM).

RECENT WORK:

Many of scholars have used different machine learning algorithms to present their work on comparative analysis for diabetes risk assessment using prior stage diabetes risk prediction dataset, having 17 attributes of 520 subjects. Some of them used other datasets but representing similar problem and found out suitable accuracies by running different algorithms and performing comparative analysis for early diagnosis of diabetes. Some most relevant works are discussed in this section [11-19]. This work represents the idea that they have used some AI based algorithms like Machine learning and its detailed techniques to run comparative analysis for early detection of diabetes with the help of a dataset of early diabetes diagnosis having 17 features. After evaluation and carrying out diversified performance matrices, they came to know that XGBoost classifier outperformed almost all of the remaining algorithms and provided them with most accurate results approximately 100%, while some of the algorithms given outcome with about 90% accuracy. They have used here convolutional neural network approach too [13].
They have used various automation techniques for predicting

diabetes i.e. Ada-Boost, Bagging and random forest algorithms, from which random forest came out to be with the highest accuracy, precision and F1-scoring. For feature scoring they have applied Chi-Square technique and cross-validation with 10 folds. This work is non same dataset with similar problem description but different techniques for predicting models with high accuracy and feature selection [14]. This paper focuses on the risk factors for type 2 diabetes. This study carried out literature surveys, cleaning dataset by eliminating missing values, splitting dataset within 80-20 i.e. 80% training and 20% testing dataset. They have matched outcomes of various artificial patterns like random forest, Naïve Bayes, KNN (K-nearest neighbor) etc. Afterwards represent different criterion like accuracy, sensitivity, recalled precision curve etc. By means of carrying out all such operations they have found out that prototype with random forest shown best finest results with upmost accuracies and considering other factors too [15]. This work represents that for feature selection of oversampling technique of synthetic minority is used and for the validation of results, K-fold cross-validation has been used to carry out prediction if diabetes type 2 by applying basic lifestyle indicators calculated by feature engineering technique. Following are some specific operations are accomplished such as the bagged decision tree accuracy score, F-1 score, receiver operating characteristic (ROC) curve, misclassification score, precision etc. Experts suggested to collect relevant data online as well as offline and have taken sample size duration of 2 years. The evaluation has been fulfilled by analyzing confusion matrix taken under consideration four indicators i.e. True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) [16]. This study determines early diagnosis of diabetes mellitus by assessing the results taken out through comparison of performances of eight machine learning algorithms on the data collected by 520 people and 16 classes.

Results from different algorithms along with the approach of convolutional neural network (CNN), they were being successful by creating various models which outperformed with highest scores. The data was cross validated by fivefold cross- validation. Finally, by applying XG-Boost or Convolutional Neural Networks satisfactory results were obtained [17].

Many algorithms i.e. multilayer perceptron, Naïve Bayes, random forest, SVM etc. were applied on the dataset of diabetes obtained from UCI repository of machine learning were examined and concluded that random forest came out to be with most precise values [18]. An open database of diabetes prognosis has been utilized to work out upon prior hazard evaluation by K-means clustering method, elbow method, silhouette method etc. following algorithms have been used for carrying out diabetes forecast i.e. multi-layer perceptron (MLP), random forest (RF), and K-Nearest Neighbors (KNN), random forest (RF) and support vector machine (SVM). For the extraction of prominent characteristics is XG-Boost [19].

Methodology:

Dataset, Features, and Software Tool:

The dataset on which this research is carried out for the purpose of comparative analysis using SVM and random forest classifier algorithms is multivariate. It possesses 520 count of instances and 17 count of attributes (table 1). Software used for this purpose is PYTHON and the process can be visualized using VISUAL STUDIO. The dataset containing the manifestation of data outcome that the subject would be fresh diabetic patient or would become diabetic in near future. Detail are attached in table 1.

Data Preprocessing:

First, from the dataset of risk assessment of diabetes mellitus, feature extraction should take place, taking into consideration the set of attributes involved (table 2). Complete database should be carried on by verifications. Complete database should be analyzed vertically and horizontally. All these are some of the basic steps involving preprocessing for this certain research to be carried out.

Attributes	Description	Data type
Age	Subjects' age in years	Numeric
Gender	Gender specification as Male/Female	Explicit
Polyuria	Yes or No	Explicit
Polydipsia	Yes or No	Explicit
Sudden weight loss	Yes or No	Explicit
Weakness	Yes or No	Explicit
Polyphagia	Yes or No	Explicit
Genital thrush	Yes or No	Explicit
Visual blurring	Yes or No	Explicit
Itching	Yes or No	Explicit
Irritability	Yes or No	Explicit
Delayed healing	Yes or No	Explicit
Partial paresis	Yes or No	Explicit
Muscle stiffness	Yes or No	Explicit
Alopecia	Yes or No	Explicit
Obesity	Yes or No	Explicit
Class	Positive or Negative	Explicit

Data Set :Features	Poly variations	Count of occurrenc-:es	520	:Place	Computer
Features Characteris-:tics	N/A	Count of :features	17	Date of sub-mis-:sion	2020-07-12

Assigned :works	Separa-tion	Absent ?counts	Yes	Count of Web :Hits	14756

Table 1. Details of prior onset of diabetes mellitus databank.

Table 2: Input dataset attributes

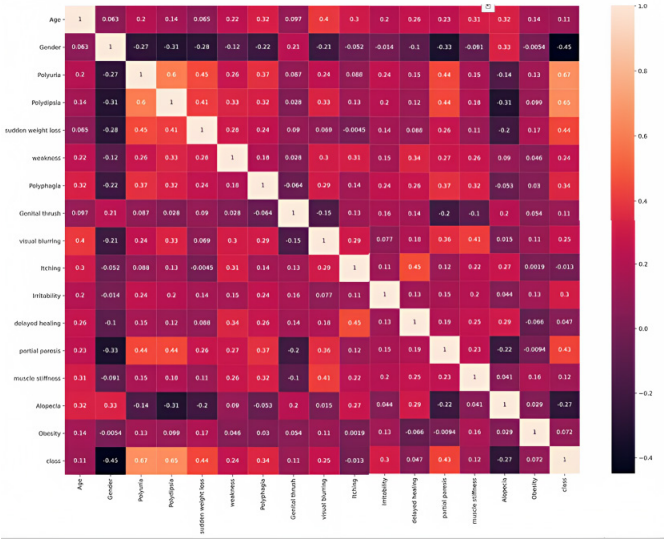

Figure 1. Encoding the features from categorical to numerical values so as to hint correlation between the characteristics of the graph.

From the heat map obtained by running commands for diabetes dataset with machine learning algorithms above in figure 1, it is evident that urea carbamide and waterlessness have a strong match with the selective member variable. It stated that both of the scenarios conclude about the foremost diagnosis of Diabetes. To be discussed about the characteristic of "Gender" most often comprises of negative values, that is the reason it is the irrelevant of all other attributes. If concluding, most of all other attributes contains similar combinations.

Dataset Splitting:

Machine learning models can only be judged after being put through their paces on many different data sets. The data file is separated into leaning and evaluating sets to eliminate overfitting under the technique of supervised machine learning algorithm. Taking under supervision, appropriately settled model and evaluating its performance by providing input of similar percentages of training and testing sets will not provide with proper outcomes for the model performance. It is intended to construct such an appropriate model; it is important to save as much of the training set as possible. Due to the similarities between the training and test data, we are able to clean up the data and better comprehend the model.

To avoid overfitting, we implemented data splitting in this study. For example, when a constructed model suits the training data properly, it will be inappropriate then to use on new particulars. As a result, we utilized 70% information for our training set.

Specifically, we used the train Control combination in the R data processing language and the rotation estimation or out-of-sample technique on the testing dataset. In all, 30% of the dataset is used to verify the accuracy of the suggested model. Utilizing specific practice, we can overcome mold outranging issue.

Features

Sequential Variable Selection:

Sequential Variable selection is a type of variable selection algorithm that selects a subunit of attributes from a huge set of attributes based on their relevance to a given problem. It works by iteratively adding or removing features from the feature set until a desired performance level is achieved [20, 21]. The script begins with a null bench of attributes, and at each step, a new attribute is added or removed based on a criterion such as accuracy, precision, or recall. The procedure is set on till the condition of stopping is reached, such as extreme value of features or a desired level of performance. There are two important variables of sequential attribute specific algorithms: Onward specification and reverse deletion. In onward sp0ewcification, the procedure begins with a null set of attributes and repetitive addition to the most suitable attribute till a stopping term is reached. In backward depletion, the script begins with the full bunch of attributes and repetitively eliminates the least suitable attribute until a stopping condition is reached. [22]. Sequential feature selection is a commonly used technique in Artificial intelligence based machine learning technique and data analysis, as it can help to reduce the dimensionality of the feature space, improve pattern accuracy, and reduce overfitting. However, it can be computationally expensive, especially for large feature sets [20-22].

Seeking Attribute Significance using XGBoost

XGBoost is a famous gradient boosting library that can be used for feature selection and feature importance analysis.

Here's how you can use XGBoost to find feature importance :

- First, load your database into a Pandas Framing and cut it into learning and evaluating station.
- Create an instance of the XGBoost model with your desired keyed up framework. The keyed up framework include the count of trees to use, the utmost deepness of the trees, the absorbing figure, and the subsampling ratio.
- Fit the XGBoost station to the absorbed set and evaluate its conduct on the verified set.
- After training the XGBoost station, you can extract the feature importance scores utilizing the characteristics_

performance_attribute of the learned station. This attribute returns a list of performance evaluation for each attribute in the database. The feature importance scores using a bar chart or a heat map. This will allow you to quickly identify the most significant attributes in the data file.

The code will load the data file, break it into training and testing sets, define XGBoost model with hyper parameters, fit the model to the training data, extract feature importance scores, and finally, visualize the feature importance scores using a bar chart. The resulting chart will show the relative importance of each feature in the dataset.

Classification Model:

Support Vector Machine:

Support Vector Machine (SVM) is a powerful and Artificial intelligence structured set that is widely used for identification and regression assessment [23]. It is a type of self-learned dataset that is capable of handling both direct and indirect classification and regression issues. The database in separated into various sets and identifies outcomes for evaluation of regression implementation, as by disclosing a hyperplane in a large proportions feature station area serve on huge separation of database, this is the plan where SVM projects. [24]. The most important aspect of SVM is to recognize most fit and appropriate hyperplane boundary that splits the data sheet into various stages defined. The hyperplane is defined as the bundles of targets in the feature workstation where the decision boundary lines. The SVM identifies the hyperplane that expands the boundary lines amongst two categories, that varies among the hyperplane and the recent data end from other categories. As far as data is concerned data linearity is not separable, SVM utilizes an approach called kernel trick to point out data to an expended dimensional order where a linear borderline can be established. [25]. The kernel activity premeditated the matching index between pairs of data labels in characteristic pool, and the aligning of the data to expended version capacity is done implicitly, without the need for explicit computation of the coordinates. SVM is known for its ability to generalize well and handle high-dimensional data with ease. It is also robust to noise and outliers, making it suitable for real-world applications [24-26]. Although, SVM can be intuitive to the selection of kernel activity and its relevant criterion, and it would be estimating extortionate for huge data sheets.

Random Forest:

A group of learning techniques that comprises of analysis of various decision trees to enhance results and outcome and avoid overfitting [27]. Random forest pattern determines the formulation of a set of decision branches on arbitrarily subsets of the learned sheets and arbitrarily sheets of the respective features. This task is pursued by a database on decision grouped branches with each tree in the forest is taught on varied subset of the specific attributes, consequently a huge set of decision trees with various strategies and deficiencies. [28] While making them learn, each decision tree is formulated by recurrent splitting the feature coverage into zones, such that each zone is as analogues as feasible with the respective spot terms. The separation is achieved by choosing the correct most attribute and cut spot at each point, based on a characteristic so as to obtain information gain or Gini impurity [29, 30]. While diagnosis, the random forest pattern sums up the diagnostics of each decision tree in the compilation to reach to a near to perfect assessment. For Classification issues, the utmost relevant divisions are evaluated by the decision branches which is selected in order to reach finalized assessments, meanwhile for regressing issues, the mean or median of the evaluated points is considered as last assessment. Random forest is called to be most compatible to huge-valued database, noisy collection of data points and with absent counts. It also facilitates with overfitting relevance and can help overcome attribute significance, that could be utilized for attribute characterization [31].

Outcome and Validation:

Results:

It can be concluded that there is a moderate positive correlation between the features “Polyuria” and “Polydipsia” with the target variable “class”, as indicated by the correlation coefficients of 0.67 and 0.65, respectively. This suggests that an increase in the occurrence of polyuria and polydipsia is associated with a higher probability of the positive class. The fact that these two specifications are also positively correlated with each other (association value of 0.60) suggests that they are likely to occur together in the same individuals.

The attribute “Gender” has almost all turn down association sets, which means it is not strongly associated with other features in the dataset. However, it is still possible that gender may have an independent association with the target variable, and should not be ignored. The remaining features in the dataset have the same association on average, which means they are roughly equally correlated with the target variable. The results with both circumstances are represented in the form of relationship between the coefficients with entities are 0.67 and 0.65, respectively. Both of the features also corresponds with the value of 0.60.

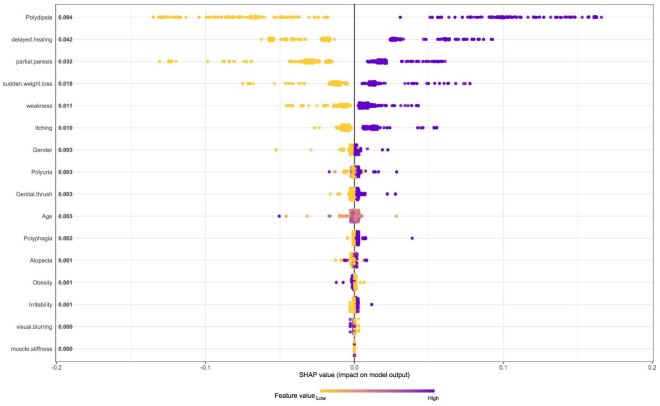

Figure 2. Feature importance graph

The topmost feature score observed by using algorithm XGBoost for predicting diabetes is "Polydipsia". This leads to the understanding of Polydipsia as the most reasonable cause for having a subject the onset of diabetes.

Another cause after polydipsia appears to be linked with diabetes prediction in a subject is "delayed healing". It may be assumed that this can serve as a most commonly occuring practice for diabetic patients.

However, the feature "Polyuria" has a very least effect being lower on board, showing that it has the least effect on diabetes onset. This suggests that other features may be more important for predicting diabetes, and that Polyuria may not be a strong indicator of the disease on its own.

Overall, the feature importance graph shown in Fig. 8 provides valuable insights into the relative importance of each feature for predicting diabetes, and can help guide further analysis and model development.

Table 3. Precision and Kappa test results of the classification prototype on the database.

	Accuracy	Kappa
SVM	98.7%	97.5%
Random Forest	66.6%	65.7%

DISCUSSION

This study suggests different studies that have been conducted to predict diabetes using artificial intelligence vased features. Various datasheets and artificial patterns were used by different scientist to make their prototypes, and the precision was of each prototype was examined. The beginning study by Victor Chang et al. handed down the "Pima Indians Diabetes Dataset" from Kaggle and make prototypes passed down J48 DT, RF, and NB separators to reach down predictions regarding diabetes mellitus. The reliability of each model was evaluated, and RF was found to have the best accuracy of 79.57% [32]. The second study by J.J. Khanam et al. used the "Pima Indians Diabetes Dataset" from the UCI Machine Learning Repository and applied various algorithms such as Decision Tree, RF, Naive Bayes, Logistic Regression, KNN, AdaBoost, and SVM to assess diabetes. The accuracy of each algorithm was evaluated, and the outcome were considered to match with other relevant works [33]. Another study by M.A. Sarwar et al. used Logistic Regression, SVM, KNN, Naive Bayes Decision Tree, and RF to predict early diabetes at a starting stage. The efficiency of each artificial pattern was examined, and the outcome were assisted with other specific works [34]. The fourth study by L. Alturki et al. used RF, Logistic Regression, XGBoost, SVM, and KNN to assess the readings of onset for diabetic patients. The precision of each algorithm was tested, and the predictions were matched with those of other relevant researches. The study found that XGBoost had the best accuracy of 94.86%, while RF had the second best accuracy of 94.82% [35].

The XGBoost as the most significant feature is suggested to be the most helped pattern for carrying out meaning research for predicting Early diabetes. The results of this study are consistent with prior research on the topic. For example, a study [36] used machine learning algorithms to predict diabetes and found that polydipsia was the most important feature for predicting the disease [36]. This is consistent with the results of the current study, which also found that polydipsia was the most important feature for predicting diabetes. Similarly, a study [37] used machine learning algorithms to predict diabetes and found that delayed wound healing was an important predictor of the disease. This is also consistent with the results of the current study, which found that delayed healing was the second most important feature for predicting diabetes.

CONCLUSION

In summary, the discussed study is advantageous as it offers valuable insights into the topic. Interrelationships between various characteristics and data points play a role in forecasting diabetes.

It aligns with prior studies on the subject. These findings can help direct additional examination and research. The creation of models can be beneficial for healthcare professionals and researchers working in the field. The realm of diagnosing and treating diabetes in general, research indicates that RF is commonly used. An algorithm that is efficient in predicting diabetes, consistently delivering high accuracy across various scenarios and collections of data. The present research contributes to this knowledge base by utilizing clinical data for construction and evaluation the framework for forecasting and evaluating diabetic individuals. This research study is considered highly valuable contribution to the healthcare industry, offering a new method for predicting and handling. Diabetes brings a new method to assess and confirm diabetic

cases in the healthcare industry and searches that are relevant to the subjects. Healthcare professionals can benefit from the findings of this study, forecasting and detecting diabetes, and for scientists to create more precise and reliable methods, efficient machine learning models to forecast diseases. Moreover, the research underscores the importance of identifying and assessing the key characteristics to anticipate diabetes which can provide direction for future research in this field.

REFERENCES

[1] S. Abhari, S. R. N. Kalhori, M. Ebrahimi, H. Hasannejadasl, and A. Garavand, "Artificial intelligence applications in type-2 diabetes mellitus care: Focus on machine learning methods," Healthc. Inform. Res, vol. 25, pp. 248–261, 2019.

[2] A. Allalou et al., "A predictive metabolic signature for the transition from gestational diabetes mellitus to type 2 diabetes," Diabetes, vol. 65, no. 9, pp. 2529–2539, 2016.

[3] A. Mujumdar and V. Vaidehi, "Diabetes Prediction using Machine Learning Algorithms," Procedia Comput. Sci., vol. 165, pp. 292–299, 2019.

[4] R. Andrews, J. Diederich, and A. B. Tickle, "Survey and critique of techniques for extracting rules from trained artificial neural networks," Knowl. Based Syst., vol. 8, no. 6, pp. 373–389, 1995.

[5] M. K. Hasan, M. A. Alam, D. Das, E. Hossain, and M. Hasan, "Diabetes prediction using ensembling of different machine learning classifiers," IEEE Access, vol. 8, pp. 76516–76531, 2020.

[6] A. N. Soni, "Diabetes mellitus prediction using ensemble machine learning techniques," SSRN Electron. J., 2020.

[7] S. L. Chiu, "Fuzzy model identification based on cluster estimation," J. Intell. Fuzzy Syst., vol. 2, no. 3, pp. 267–278, 1994.

[8] A. Yahyaoui, A. Jamil, J. Rasheed, and M. Yesiltepe, "A decision support system for diabetes prediction using machine learning and deep learning techniques," in 2019 1st International Informatics and Software Engineering Conference (UBMYK), 2019.

[9] P. Cunningham, J. Carney, and S. Jacob, "Stability problems with artificial neural networks and the ensemble solution," Artif. Intell. Med., vol. 20, no. 3, pp. 217–225, 2000.

[10] D. Dutta, D. Paul, and P. Ghosh, "Analysing feature importances for diabetes prediction using machine learning," in 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2018.

[11] A. T. Kharroubi, "Diabetes mellitus: The epidemic of the century," World J. Diabetes, vol. 6, no. 6, p. 850, 2015.

[12] Y. Wu, Y. Ding, Y. Tanaka, and W. Zhang, "Risk factors contributing to type 2 diabetes and recent advances in the treatment and prevention," Int. J. Med. Sci., vol. 11, no. 11, pp. 1185–1200, 2014.

[13] M. A. R. Refat, M. A. Amin, C. Kaushal, M. N. Yeasmin, and M. K. Islam, "A comparative analysis of early stage diabetes prediction using machine learning and deep learning approach," in 2021 6th International Conference on Signal Processing, Computing and Control (ISPCC), 2021.

[14] U. E. Laila, K. Mahboob, A. W. Khan, F. Khan, and W. Taekeun, "An ensemble approach to predict early-stage diabetes risk using machine learning: An empirical study," Sensors (Basel), vol. 22, no. 14, p. 5247, 2022.

[15] V. Chang, M. A. Ganatra, K. Hall, L. Golightly, and Q. A. Xu, "An assessment of machine learning models and algorithms for early prediction and diagnosis of diabetes using health indicators," Healthcare Analytics, vol. 2, no. 100118, p. 100118, 2022.

[16] S. M. Ganie and M. B. Malik, "An ensemble Machine Learning approach for predicting Type-II diabetes mellitus based on lifestyle indicators," Healthcare Analytics, vol. 2, no. 100092, p. 100092, 2022.

[17] Ö. N. Ergün and H. O.ilhan, "Early stage diabetes prediction using machine learning methods," European Journal of Science and Technology, 2021.

[18] S. Patel, "Predicting a risk of diabetes at early stage using machine learning approach," Turkish Journal of Computer and Mathematics Education (TURCOMAT), vol. 12, no. 10, pp. 5277–5284, 2021.

[19] M. M. Hassan, S. Mollick, and F. Yasmin, "An unsupervised cluster-based feature grouping model for early diabetes detection," Healthcare Analytics, vol. 2.

[20] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," Comput. Electr. Eng., vol. 40, no. 1, pp. 16–28, 2014.

[21] "Feature selection algorithm for intrusions detection system using sequential forward search and random forest classifier," KSII Trans. Internet Inf. Syst., vol. 11, no. 10, 2017.

[22] A. A. Abaker and F. A. Saeed, "A comparative analysis of machine learning algorithms to build a predictive model for detecting diabetes complications," Informatica (Ljubl.), vol. 45, no. 1, 2021.

[23] J. Li, J. Ding, D. Zhi, K. Gu, and H. Wang, "Identification of type 2 diabetes based on a ten-gene biomarker prediction model constructed using a support vector machine algorithm BioMed Res," BioMed Res. Int, 2022.

[24] P. K. Upadhyay and C. Nagpal, "Wavelet based performance analysis of SVM and RBF kernel for classifying stress conditions of sleep EEG Sci," Technol, vol. 23, no. 3, pp. 292–310, 2020.

[25] B. Schoslkopf and A. Smola, Learning with Kernels, Support Vector Machines. London: MIT Press, 2002.

[26] O. A. Rasheed, E. Mohammed, and S. Iris, "Comparison of supervised machine learning classification techniques in prediction of locoregional recurrences in early oral tongue cancer," Int J Med Inform, vol. 136, 2020.

[27] M. M. Hassan, Z. J. Peya, S. Mollick, M. A. Billah, M. M. Hasan Shakil, and A. U. Dulla, "2021 12th International Conference on Computing Communication and Networking Technologies, ICCCNT," pp. 1–05, 2021.

[28] P. Ghosh, A. Karim, S. T. Atik, S. Afrin, and M. Saifuzzaman, "Expert cancer model using supervised algorithms with a LASSO selection approach Int," Int. J. Electr. Comput. Eng. (IJECE), vol. 11, no. 3, pp. 2632–2640, 2021.

[29] M. Jena and S. Dehuri, "DecisionTree for classification and regression: A state-of-the art review," Informatica (Ljubl.), vol. 44, no. 4, 2020.

[30] W. Xu, J. Zhang, Q. Zhang, and X. 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), 2017.

[31] B. Baba and G. Sevil, "Predicting IPO initial returns using random forest," Borsa Istanb. Rev., vol. 20, no. 1, pp. 13–23, 2020.

[32] V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms Neural Comput," Neural Comput. Appl, pp. 1–17, 2022.

[33] J. J. Khanam and S. Y. Foo, A comparison of machine learning algorithms for diabetes prediction ICT Express. Express, 2021.

[34] M. A. Sarwar, N. Kamal, W. Hamid, and M. A. Shah, "Prediction of diabetes using machine learning algorithms in healthcare," in 2018 24th International Conference on Automation and Computing (ICAC), 2018.

[35] L. Alturki, K. Aloraini, A. Aldughayshim, and S. Albahli, "Predictors of readmissions and length of stay for diabetes related patients," in 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA), 2019.

[36] H. Yun, J. Choi, and J. H. Park, "Prediction of critical care outcome for adult patients presenting to emergency department using initial triage information: An XGBoost algorithm analysis," JMIR Med. Inform., vol. 9, no. 9, p. e30770, 2021.

[37] X. Mo et al., "Early and accurate prediction of clinical response to methotrexate treatment in juvenile idiopathic arthritis using machine learning," Front. Pharmacol., vol. 10, p. 1155, 2019.