**Utilizing Bioinformatics to Predict Diabetes**

Brandon Stevena, Jonathan Ivanb, Rayes Jordan Pradanac, Julius Ferdinandd

*a Computer Science Department, School of Computer Science, Bina Nusantara University, Tanggerang, Indonesia 15143*

*b Computer Science Department, School of Computer Science, Bina Nusantara University, Tanggerang, Indonesia 15143*

*c Computer Science Department, School of Computer Science, Bina Nusantara University, Tanggerang, Indonesia 15143*

*d Computer Science Department, School of Computer Science, Bina Nusantara University, Tanggerang, Indonesia 15143*

Abstract

Diabetes mellitus is a pressing global health concern associated with significant mortality, morbidity, and healthcare costs. This chronic disease disrupts glucose regulation, leading to severe complications such as blindness, kidney failure, myocardial infarction, stroke, and limb amputation. Diabetes management requires a deep understanding of the disease and its determinants, including diet, exercise, infection, and stress. This paper delves into the application of machine learning techniques for predicting and classifying diabetes, encompassing various types and hybrid models.

Keywords: Diabetes mellitus; healthcare; machine learning; predictive analysis

1. Introduction

Diabetes represents a significant global burden in terms of mortality, morbidity, and health-system costs in the world [1], it is a chronic disease which manifests itself into the body when it fails to regulate the amount of glucose in the blood, Leading to increased urine production by the kidney when the body does not produce enough insulin or does not consume it the way it should [2]. Diabetes is the main reason for other fatal diseases such as blindness, kidney failure, myocardial infarction, stroke, and amputation of the lower limbs [3]. Various factors including diet, exercise, infection, and stress influences diabetes. And hence why diabetic patients need a thorough understanding of the disease, how it affects the body, and its complications [4]. Consequently, effective management of our diet, food choices, and nutritional intake has been proven to be the long-term solution for diabetic patients [4].

Diabetes mellitus is a collection of various metabolic diseases, its most common characteristic among them is chronic high blood sugar level. There are three types of diabetes mellitus: type 1 diabetes, type 2 diabetes, and gestational diabetes (diabetes during pregnancy). Type 1 diabetes represents 5-10% of diabetic cases, being the primary cause of the immune system destroying the insulin producing pancreatic beta cells [5]. Type 2 diabetes is the most common one which amounts to 90% of cases of diabetes. Genetic and environmental factors participate in its development [5]. Gestational diabetes is defined as a condition where pregnant women have difficulty tolerating carbohydrates, it affects around 1-9% of pregnant women depending on the prevalence of type 2 diabetes in the population. And it does not stop when the mother gives birth, rather in the future the mother will be at risk of developing type 2 diabetes and additionally, in the future the child will have a higher chance of obesity and diabetes [6].

Machine learning, more specifically classification techniques are widely used in the medical field, it categorizes data into different classes based on specific criteria when comparing different individual classifiers [7]. There is a lot of research that has been done with similar topics, while utilizing various methods and techniques to experiment with detecting diabetes. One of such research [8] mentioned the utilization of several techniques of data preprocessing to process the dataset used, and the use of an ensemble of different classifier models  (KNN, AdaBoost, Naive Bayes, Decision Trees, Random Forest, and XGBoost) with the addition of MLP within the bundling of models used for the experiment. The result of the experiment, evaluated by utilizing ROC-AUC as the evaluation metric, ended with the proposed classifier outperforming the state-of-the-art result by around 2.00%. Another research [9] also mentioned the use of ensemble machine learning techniques for the research, utilizing several classifiers such as Random Forest, Support Vector Machine, Decision Tree, Multi-layer Perceptron, and Naïve Bayes to experiment. The result that was evaluated based on several evaluation metrics such as precision, recall, ROC, and F-Measure had shown that the proposed method has created a satisfying result, with only 8 recorded mistakes throughout the process of experimentation. The result of the experiment shows that the ensemble method had performed better than other singular base models, which means that the result of the experiment has fulfilled the goal of the research itself. One paper [10] also used a similar method as the ones mentioned previously, utilizing one of the most popular ensemble machine learning technique, Bagging and a meta-learner called DECORATE (short for Diverse Ensemble Creation by Oppositional Relabeling of Artificial Training Examples), comparing the result of the two while also comparing the final result with previous related works. The result gained from the performance of the two methods used in the experiment that has been evaluated using various performance metrics such as accuracy, MAE, RMSE, precision and recall (weighted average), etc. had shown that the proposed methods has far surpassed the results gained from experiments done on previous related researches, with the result of the DECORATE method being the best one amongst the two, reaching as high as 98.53% of accuracy percentage.

A lot of research had also done comparative studies. Another research [11] also utilized similar methods, by using multiple machine learning models (SVM, KNN, Logistic Regression, XGBoost, Random Forest, and Decision Tree) with the addition of some deep learning models as well. With the addition of accuracy, precision, recall, and F1-score, as the evaluation metrics, the five evaluation metrics including the ROC-AUC had shown that the performance of the models used in this research is incredible, with the best performing model being XGradient Boost classifier. There is also another research [12] that also did a similar thing, utilizing multiple machine learning models (Naive Bayes, Decision Tree, and KNN with K=1 and K=3) to experiment with classifying diabetes, and with the result that is evaluated using accuracy metric, had shown that two of the proposed models has shown a performance that dwarfs all of the previous related research mentioned in the paper by quite an amount, reaching 94.44% and 93.79% accuracy percentage for Decision Tree and KNN (K=1) bootstrapped models respectively. Another research [13] also did similar things, using some of the most commonly used machine learning classifiers such as SVM and Random Forest,  and a deep learning method which is CNN. The overall performance shown by the models used, which is evaluated by accuracy, has shown quite a result, obtaining 76.81%, 65.38% and 83.67% accuracy percentage for DL (CNN), SVM, and RF respectively. Another paper [14] mentioned a similar method, utilizing a different set of machine learning models (Support Vector Machines, K-Nearest Neighbors, Decision Trees, Logistic Regressions, Discriminant Analysis) and their variants, which is accessed using MATLAB software, or to be exact, the MATLAB Classification Learner Tool (MCLT). Utilizing 10-fold cross validation for the data splitting process, the performance shown through the average accuracy that averages within the range of 65.5% and 77.9% entails that it is a pretty successful research, with the Logistic Regression model being the top performing model, reaching 77.9% average accuracy.

A lot of studies also did experiments which were done using several different models and methods, looking for one single best performing model or method. One of such papers [16] used three models, which are Logistic Regression, Naive Bayes, and K-Nearest Neighbor. The result gained from the experiment shows that out of the three, Logistic Regression is the best performing model, reaching as high as 94% precision percentage, a huge difference when compared to the other two, which reaches merely 79% and 69% precision percentage for Naive Bayes and KNN respectively. Another research [17] did something similar, using four different machine learning classifiers such as J48, Naïve Bayes (NB), Logistic Regression (LR), and Random Forest (RF). The experiment was done using several different cross-validation values (K = 5, 10, 15, and 20), and was evaluated using several performance metrics such as accuracy, precision, recall, F1-score, and AUC. The result shows that Logistic Regression has the highest accuracy percentage amongst all four K values used for the cross-validation, reaching 77% accuracy percentage, while also being above average for the other metrics.

There are also some papers which proposed the use of a single model that has been modified, enhanced, and/or improved. One of such papers [18] proposed the use of a tuned Multi-Layer Perceptron (MLP) model for the task of diabetes classification. The result of the experiment showed that the proposed model had shown better performance than other models such as Random Forest and Logistic Regression, reaching 86.083% accuracy percentage, a huge improvement when compared to the other two comparative models, which only averages around 75% accuracy percentage. Another paper [19] proposed the use of SVM and MLP for the same task mentioned previously. The result of both models shows that both models are usable for the task of classifying diabetes, with MLP being the better one amongst the two, reaching 77.474% accuracy percentage. Another paper [20] also proposed two models, SVM and Naive Bayes, comparing the results of the two to find which one is the better one. The result shows that SVM has a better performance than Naive Bayes, reaching 82% accuracy percentage, a bigger number compared to the performance of the Naive Bayes model, which only reaches 62.5%.

Some other papers with the same topic of diabetes classification also mentioned the use of hybrid machine learning models. One paper [21] used an ensemble method to combine various models such as ANN, SVM. KNN, and Naive Bayes, utilizing tools such as MATLAB and WEKA ver. 3.6.13 to do so. The result of the proposed model reached a very high percentage number, reaching 98.6% accuracy percentage, showing that this pathway of research is worth researching and experimenting further. Another paper [22] mentioned the use of an ensemble approach for classifying diabetes using a soft voting classifier, creating an ensemble of machine learning models such as Logistic Regression, Naive Bayes, and Random Forest. The result, although shown as not the best performing models when compared to the cutting-edge models of machine learning classifiers, such as XGBoost and CatBoost, had shown quite an improvement when compared to the models used to create the proposed ensemble soft voting classifier model (Logistic Regression, Naive Bayes, and Random Forest).

Another paper [23] also researched the topic of data mining techniques that can be used to do classifying tasks, which also includes the topic of diabetes classification. There is also another paper [24] that mentioned the use of the Random Forest algorithm for predicting diabetes. The research proved through theories gathered from related works that the utilization of Random Forest algorithm is possible. One paper [25] mentioned the use of Logistic Regression, split into 6 different models based on the features selected. The preferred specification, which is model 6, shows a very good performance, gaining 78.26% accuracy percentage, evidence that it could potentially be used for future research and experimentation to improve the model even further.

Some papers focus on experimentation, utilizing various metrics and techniques to evaluate the performance of the models. One of such papers [26] mentioned the use of several different models, such as Logistic Regression, K-Nearest Neighbour, Support Vector Machine, Naïve Bayes, Decision Tree, and Random Forest. The use of various performance metric, such as Accuracy, Error Rate, Sensitivity, Specificity, Precision, F-Measure, and MCC shows that the Random Forest model is the best performing model among the selected models, reaching as high as 100% AUC and 94.1% accuracy percentage. Another paper [27] compares machine learning-based models such as XGBoost, LightGBM, Glmnet, and Random Forest to a simple regression model on the task of predicting and classifying diabetes. The result, which is evaluated mainly using RMSE, shows that the simple regression model has the lowest amount of RMSE, only having 0.838 RMSE, which also means that it is the best performing model based on RMSE, and after more data was added, the average performance of Glmnet increases by around 3.4%, while LightGBM was mentioned to have the highest level of variable selection stability.

A comparative study between various machine learning models was done in various papers. one was done in this paper [28] between three machine learning models for a dataset of female patients with minimum twenty one year age of Pima Indian population from UCI machine learning repository, the models that were used was Support vector machine (SVM), Radial Basis Function (RBF) Kernel Support SVM, k-Nearest Neighbour (k-NN), Artificial neural network (ANN), and Multifactor Dimensionality Reduction (MDR). Results showed SVM-linear achieved the best accuracy and precision (0.89 and 0.88) for diabetes prediction, while k-NN had the highest recall and F1 score (0.90 and 0.88). Another one is this paper [29] Pima Indian Diabetes Database was the dataset that they had used, with Logistic Regression (LR), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Random Forest (RF), and their proposed framework logistic regression which was the use of grid search for hyper-parameter tuning in logistic regression. The results were that their proposed framework logistic regression came out to be the most accurate followed by RF, SVM, KNN, and LR. Next, is this paper [30] that compares their fused machine learning model where the model utilizes Support Vector Machines (SVM) and Artificial Neural Networks (ANNs) for classification with other models from other papers. The result was that their model did the best accuracy wise, with 94.87% accuracy, followed by [12] with their decision tree after bootstrapping approach that reached 94.4% accuracy, and [31] with their neural network approach that reached 87.88% accuracy. And then this paper [32] various machine learning models for type 2 DM which includes logistic regression, random forest, support vector machine, and confusion matrix-based classifier integration approach. Which SVM attained the best score overall between the performance metrics followed by RF that came by a really close second. And lastly, this paper [33] compares the performance of support vector machine (SVM) and random forest (RF), which showed that RF was more effective for classification of the diabetes in all rounds of experiments and produced overall accuracy for diabetic prediction to be 83.67%, while the prediction accuracy for SVM reached 65.38% on our dataset.

Other than type 1 and 2 diabetes, there are also research about gestational diabetes mellitus: the first one is [34] they used SVM as their model, and they concluded that they discovered that the coagulation levels of PAT-PT and REF-APTT were notably elevated in individuals who eventually developed GDM (gestational diabetes mellitus) after the initial 19 weeks. These biomarkers exhibited exceptional sensitivity (92.5%) and specificity (99.2%) in forecasting GDM occurrence. Their discovery proposed a fresh method for early GDM prediction. And then next is this paper [35] using the K-Means algorithm aided by the elbow technique and then clustering the data into optimal clusters. Where the Mahalanobis distance method was selected for the most relevant cluster. After that classification techniques (DT, RF, NB, KNN, SVM, LR) were used and they achieved accuracies of 92% and 90% in ensemble methods. And then the next 2 papers talks about early detection of GDM for Chinese women,  where [36] used various machine learning model, such as logistic regression (LR), k-nearest neighbor (KNN), support vector machine (SVM), and deep neural network (DNN) and they determined that they discovered 17 significant factors, such as lipoprotein(a), that require additional scrutiny. A pragmatic LR model consisting of 7 variables was developed for practical clinical application. Further research is necessary to comprehend the correlations among TT4, FT4, and GDM, as well as extremely low BMI and GDM within the Chinese populace. While [37] compares two machine learning models that are logistic regression and XGBoost. The outcome is that XGBoost performed more effectively in terms of accuracy, achieving 75.5%, while logistic regression obtained a lower score of 68.9%.

**2. Data**

The dataset used for this paper is the dataset predicting diabetes taken from Kaggle [38]. The National Institute of Diabetes and Digestive and Kidney Diseases provided this dataset, which aims to predict the presence of diabetes in patients using specific diagnostic measurements included in the dataset. The proportion for the training set and testing set is 80:20. This dataset has 9 features and 1 output with a total of 768 female patients. Table 1 describes its attributes and description.

Table 1. Attribute Information

|  |  |
| --- | --- |
| Attribute | Description |
| Pregnancies | To indicate the count of pregnancies |
| Glucose | To indicate the Glucose level in blood |
| BloodPressure | To indicate the Blood pressure |
| SkinThickness | To indicate the thickness of the skin |
| Insulin | To indicate the Insulin level in blood |
| BMI | To indicate the Body mass index |
| DiabetesPedigreeFunction | To indicate the Diabetes percentage |
| Age | To indicate the age |
| Outome | To indicate the result 1 for Yes and 0 for No |

**3. Methodology**

***3.1 Data Preprocessing***

One weakness of this dataset is that it exhibits an imbalanced label distribution, where the number of individuals labeled with diabetes is twice as high as those without diabetes.To address this issue, the SMOTE technique will be employed. The SMOTE technique is effective for addressing imbalanced datam and since all the features in the dataset are numerical, it is suitable for applying SMOTE. The fundamental concept of SMOTE involves randomly inserting new samples between minority samples and their neighboring instances [39].

***3.2 Proposed Method***

The major contribution of this research is to forecast diabetes using machine learning classification algorithms. Several ML algorithms were trained in this paper. To enhance the performance of each model, hyperparameter optimization will be used namely GridSearchCV. Grid search is a widely used method for tuning hyperparameters. In this technique, we generate a model for every conceivable combination of hyperparameter values provided, evaluate each model's performance, and then choose the combination that yields the optimal results [40].

Workflow menyusul

***3.3 Machine Learning Algorithm***

***3.3.1 Random Forest (RF)***

The RF algorithm is a machine learning technique that can be applied to various problems. It can be described as an ensemble of decision trees, where predictions are made by combining the outputs of multiple trees. Random forest is particularly beneficial when dealing with large datasets or data with many variables, as it is capable of producing accurate predictions in such scenarios [41].

***3.3.2 Logistic Regression (LR)***

LR, as it can be seen from the name, is a machine learning model that is often used for regression tasks. However, this model can also be used for binary classification tasks, where there are only two classes for the result. The way this model works is that it calculates the probability of a data being one of the two classes available by utilizing logistic/sigmoid function. The function will optimize any values found in-between the two classes available as the classification labels, which would result in the model being able to predict the probability of the target data being one of the two classes.

***3.3.3 Gradient Boosting (GB)***

GB is widely recognized as one of the most effective supervised ML algorithms due to its impressive performance in solving complex classification and regression problems. This versatile algorithm works by iteratively training a collection of weak predictive models, likely decision tree to create a single strong predictor [43]. By iteratively training weak predictive models, such as decision trees, and combining their outputs, gradient boosting creates a powerful ensemble model. This approach allows gradient boosting to effectively capture complex patterns in the data and improve prediction accuracy. Additionally, gradient boosting incorporates a clever optimization technique that minimizes a loss function, further enhancing its performance.

***3.3.4 Extreme Gradient Boosting (XGB)***

XGB, an implementation of ensemble learning, utilizes a combination of weak algorithms to enhance prediction accuracy. It is a gradient boosting method that leverages decision trees and can be applied to a wide range of problems. During training, XGB progressively adds trees and splits features in each iteration to build an improved model [44]. To prevent overfitting and improve the performance of the XGB (Extreme Gradient Boosting) algorithm, techniques such as sub-sampling and controlling the maximum depth of decision trees are employed. These strategies help to ensure that the model generalizes well to unseen data and achieves better accuracy by reducing the risk of overfitting [45].

***3.3.5 Ensemble Voting***

The classifier described here is a metaclassifier that combines multiple machine learning models to make predictions using a majority voting approach. It employs two types of voting techniques: hard voting and soft voting. In hard voting, the final prediction is determined by selecting the class prediction that appears most frequently among the base models. On the other hand, soft voting requires the base models to have the Predict\_proba method, and the final prediction is based on the aggregated probabilities from the base models [46].

***3.4 Performance Metrics***

Once the model is trained and tested, the results of the testing phase will be displayed, showcasing the model's performance. The evaluation metrics employed in this study include accuracy, precision, recall, and F1-score. These metrics provide quantitative representations of the model's performance, allowing for a comprehensive assessment of its effectiveness [47]. To calculate the performance metrics that we will be using for the models, we need to know what a confusion matrix is. Confusion matrix is used to represent each outcome of the model, it has four components:

* True Positives (TP): Number of samples correctly predicted as “positive.”
* False Positives (FP): Number of samples wrongly predicted as “positive.”
* True Negatives (TN): Number of samples correctly predicted as “negative.”
* False Negatives (FN): Number of samples wrongly predicted as “negative.”

#### Accuracy

Accuracy is the metric to determine how much a model has accurately predicted the outcome of a data based on the total number of data instances, it takes the number of True predictions and then divides it with all of the number of data whether True or False [48]. Mathematically, it defines as follows:

*Accuracy =*

#### Precision

Precision is the metric to determine what proportion of positive identifications was actually correct based on the training data. It takes the number of correct predictions and then divides it with the same number of correct predictions plus the amount of data that is falsely predicted as positive[48]. Mathematically, it defines as follows:

*Precision =*

#### Recall

Recall is the metric to determine what proportion of actual positives was identified correctly based on the training data. It takes the number of correct predictions and then divides it with the same number of correct predictions plus the amount of data that is falsely predicted as negative[48]. Mathematically, it defines as follows:

*Recall =*

#### F1 Score

F1 score serves as a performance metric that functions as a harmonic mean between accuracy and recall. It can be considered the optimal choice among the three methods, as increased precision comes at the expense of recall, and vice versa. Therefore, maximizing the F1 Score implies the simultaneous maximization of accuracy and recall scores [48]. Mathematically, it is defined as follows:

|  |  |
| --- | --- |
|  |  |

or

|  |  |
| --- | --- |
|  |  |

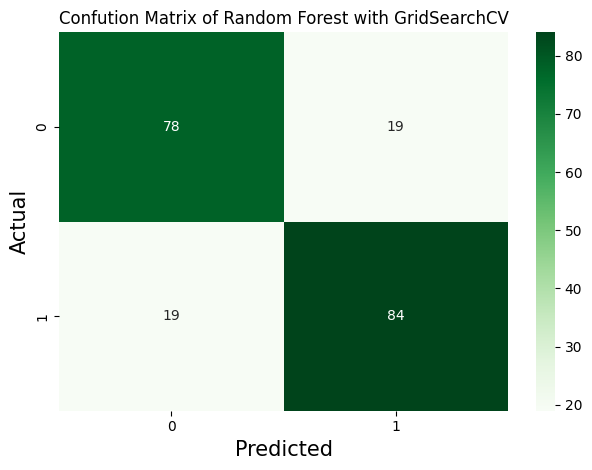
**4. Result and Analysis**

All the calculations that have been done in this paper, were done on a laptop with the following detail : AMD Ryzen 7 5800H with Radeon Graphics, 3.8GHz and 8 GB RAM, using Google Colab with the python programming language.Table 2 shows comparation between models before and after HPO.

Table 2. Comparation between models before and after HPO.

|  |  |  |  |
| --- | --- | --- | --- |
| Method | Metrics | Default | Tuning |
| Random Forest | Accuracy (%) | 0.8150 | 0.8100 |
| Precision (%) | 0.8155 | 0.8155 |
| Sensitivity (%) | 0.8235 | 0.8155 |
| F1 – Score (%) | 0.8195 | 0.8155 |
| Gradient Boosting | Accuracy (%) | 0.8050 | 0.8000 |
| Precision (%) | 0.8058 | 0.8350 |
| Sensitivity (%) | 0.8137 | 0.7890 |
| F1 – Score (%) | 0.8098 | 0.8113 |
| XGBoost | Accuracy (%) | 0.7900 | 0.8000 |
| Precision (%) | 0.8252 | 0.8155 |
| Sensitivity (%) | 0.7798 | 0.8000 |
| F1 – Score (%) | 0.8019 | 0.8077 |
| Logistic Regression | Accuracy (%) | 0.7450 | 0.7700 |
| Precision (%) | 0.7184 | 0.7961 |
| Sensitivity (%) | 0.7708 | 0.7664 |
| F1 – Score (%) | 0.7437 | 0.7810 |
| Ensemble Voting | Accuracy (%) | 0.8100 | 0.8100 |
| Precision (%) | 0.7864 | 0.8058 |
| Sensitivity (%) | 0.8351 | 0.8218 |
| F1 – Score (%) | 0.8100 | 0.8137 |

According to table 2, it can be concluded that HPO with GridSearchCV has been shown to boost the metrics of LR, XGB and, Ensemble Voting. While GridSearchCV has been shown to enhance accuracy in machine learning models, there is still room for improvement, such as broadening the range of parameters to be explored. According to table 2, it can be concluded that RF and Ensemble Voting are the best models for predicting diabetes dataset. Figure 2 and 3 shows confusion metrics of RF and Ensemble Voting.

 A picture containing text, screenshot, rectangle, diagram

Description automatically generated

Fig 2. Confusion metrics of RF with HPO. Fig 3. Confusion metrics of EV with HPO.

Figure 2 depicts True Negative number 78, False Positive number 19, False Negative number 19, True Positive number 84. On the other side, figure 3 depicts True Negative number 79, False Positive number 18, False Negative number 20, True Positive number 83. Where there are zeros and ones on the X and Y axes, zero stands for a label without diabetes and one, a label with diabetes.

**5. Conclusion**

The experiment findings led to several conclusions that align with the problem description. These conclusions are outlined below:

1. RF and Ensemble Voting are the best models for this dataset with gain accuracy of 81.00% respectively. Despite the modest accuracy of 81.00%, it is important to consider the context of the experiment discussed in the paper. The dataset used in this study is unbalanced, indicating higher complexity compared to balanced datasets.
2. Hyperparameter tuning with GridSearchCV has been proved can boost the machine learning model’s performance, eventhough not statistically significant, this approach performs well.

This experiment shows that machine learning algorithms can be used to forecast diabetes in real life conditions. However, this experiment still needs several improvements, especially for the dataset. Future research should consider more thorough HPO in addition to using more powerful machine learning models like neural network and others to gain better performance and utilizing diverse datasets to broaden the scope of the study population.

References:

[1] Zhou, Bin, et al. "Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4· 4 million participants." *The lancet* 387.10027 (2016): 1513-1530.

[2] Cockram, Clive S. "Diabetes mellitus: perspective from the Asia–Pacific region." *Diabetes research and clinical practice* 50 (2000): S3-S7.

[3] Suryasa, I. Wayan, María Rodríguez-Gámez, and Tihnov Koldoris. "Health and treatment of diabetes mellitus." *International Journal of Health Sciences* 5.1 (2021).

[4] Shivashankar, Murugesh, and Dhandayuthapani Mani. "A brief overview of diabetes." *International Journal of Pharmacy and Pharmaceutical Sciences* 3.4 (2011): 22-27.

[5] Czichos, Horst, and Tetsuya Saito. *Springer handbook of materials measurement methods*. Ed. Leslie Smith. Vol. 978. Berlin: Springer, 2006.

[6] Ozdemir, M., et al. "Risk factors for ocular surface disorders in patients with diabetes mellitus." *Diabetes research and clinical practice* 59.3 (2003): 195-199.

[7] Sisodia, Deepti, and Dilip Singh Sisodia. "Prediction of diabetes using classification algorithms." *Procedia computer science* 132 (2018): 1578-1585.

[8] Hasan, Md Kamrul, et al. "Diabetes prediction using ensembling of different machine learning classifiers." *IEEE Access* 8 (2020): 76516-76531.

[9] Soni, Ankit Narendrakumar. "Diabetes Mellitus Prediction Using Ensemble Machine Learning Techniques." *Available at SSRN 3642877* (2020).

[10] Islam, Md Tanvir, et al. "Diabetes mellitus prediction using ensemble machine learning techniques." *Advances in Computational Intelligence, Security and Internet of Things: Second International Conference, ICCISIoT 2019, Agartala, India, December 13–14, 2019, Proceedings 2*. Springer Singapore, 2020.

[11] Refat, Md Abu Rumman, et al. "A comparative analysis of early stage diabetes prediction using machine learning and deep learning approach." *2021 6th International Conference on Signal Processing, Computing and Control (ISPCC)*. IEEE, 2021.

[12] Saru, S., and S. Subashree. "Analysis and prediction of diabetes using machine learning." *International journal of emerging technology and innovative engineering* 5.4 (2019).

[13] Yahyaoui, Amani, et al. "A decision support system for diabetes prediction using machine learning and deep learning techniques." *2019 1st International informatics and software engineering conference (UBMYK)*. IEEE, 2019.

[14]  Al-Zebari, Adel, and Abdulkadir Sengur. "Performance comparison of machine learning techniques on diabetes disease detection." *2019 1st international informatics and software engineering conference (UBMYK)*. IEEE, 2019.

[16] Khaleel, Fayroza Alaa, and Abbas M. Al-Bakry. "Diagnosis of diabetes using machine learning algorithms." *Materials Today: Proceedings* 80 (2023): 3200-3203.

[17] Battineni, Gopi, et al. "Comparative machine-learning approach: A follow-up study on type 2 diabetes predictions by cross-validation methods." *Machines* 7.4 (2019): 74.

[18] Butt, Umair Muneer, et al. "Machine learning based diabetes classification and prediction for healthcare applications." *Journal of healthcare engineering* 2021 (2021)

[19] Thaiyalnayaki, K. "Classification of diabetes using deep learning and svm techniques." *International Journal of Current Research and Review* 13.01 (2021): 146.

[20]Nadeem, Muhammad Waqas, et al. "A fusion-based machine learning approach for the prediction of the onset of diabetes." Healthcare. Vol. 9. No. 10. MDPI, 2021.

[21]Sarwar, Abid, et al. "Diagnosis of diabetes type-II using hybrid machine learning based ensemble model." International Journal of Information Technology 12 (2020): 419-428.

[22] Kumari, Saloni, Deepika Kumar, and Mamta Mittal. "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier." International Journal of Cognitive Computing in Engineering 2 (2021): 40-46.

[23] Preetha, S., et al. "Diabetes Disease Prediction Using Machine Learning." *Int. J. Recent trends Eng. Res.* 6.5 (2020).

[24] VijiyaKumar, K., et al. "Random forest algorithm for the prediction of diabetes." 2019 IEEE international conference on system, computation, automation and networking (ICSCAN). IEEE, 2019.

[25] Joshi, Ram D., and Chandra K. Dhakal. "Predicting type 2 diabetes using logistic regression and machine learning approaches." International journal of environmental research and public health 18.14 (2021): 7346.

[26]Tigga, Neha Prerna, and Shruti Garg. "Prediction of type 2 diabetes using machine learning classification methods." Procedia Computer Science 167 (2020): 706-716.

[27] Kopitar, Leon, et al. "Early detection of type 2 diabetes mellitus using machine learning-based prediction models." *Scientific reports* 10.1 (2020): 11981.

[28] Kaur, H., & Kumari, V. (2020). Predictive modelling and analytics for diabetes using a machine learning approach. Applied Computing and Informatics, ahead-of-print(ahead-of-print). doi:10.1016/j.aci.2018.12.004

[29] Krishnamoorthi, Raja, et al. "A novel diabetes healthcare disease prediction framework using machine learning techniques." *Journal of Healthcare Engineering* 2022 (2022).

[30] Ahmed, Usama, et al. "Prediction of diabetes empowered with fused machine learning." *IEEE Access* 10 (2022): 8529-8538.

[31] Jain, Bhavini, et al. "WITHDRAWN: A machine learning perspective: To analyze diabetes." (2021).

[32] Deberneh, H. M., & Kim, I. (2021). Prediction of Type 2 Diabetes Based on Machine Learning Algorithm. International Journal of Environmental Research and Public Health, 18(6), 3317. doi:10.3390/ijerph18063317

[33] Yahyaoui, A., Jamil, A., Rasheed, J., & Yesiltepe, M. (2019). A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques. 2019 1st International Informatics and Software Engineering Conference (UBMYK). doi:10.1109/ubmyk48245.2019.8965556

[34] Xiong, Yan, et al. "Prediction of gestational diabetes mellitus in the first 19 weeks of pregnancy using machine learning techniques." *The journal of maternal-fetal & neonatal medicine* 35.13 (2022): 2457-2463.

[35] Jader, Rasool, and Sadegh Aminifar. "Predictive Model for Diagnosis of Gestational Diabetes in the Kurdistan Region by a Combination of Clustering and Classification Algorithms: An Ensemble Approach." *Applied Computational Intelligence and Soft Computing* 2022 (2022).

[36] Wu, Yan-Ting, et al. "Early prediction of gestational diabetes mellitus in the Chinese population via advanced machine learning." *The Journal of Clinical Endocrinology & Metabolism* 106.3 (2021): e1191-e1205.

[37] Liu, Hongwei, et al. "Machine learning risk score for prediction of gestational diabetes in early pregnancy in Tianjin, China." *Diabetes/metabolism research and reviews* 37.5 (2021): e3397.

[38] <https://www.kaggle.com/datasets/whenamancodes/predict-diabities>

[39] Alex, Suja A., et al. "Deep LSTM Model for Diabetes Prediction with Class Balancing by SMOTE." Electronics 11.17 (2022): 2737.

[40] Kumar, P. Manoj, et al. "Diabetes Prediction Using Different Ensemble Learning Classifiers in Machine Learning."

[41] Katarya, Rahul, and Sunit Kumar Meena. "Machine learning techniques for heart disease prediction: a comparative study and analysis." Health and Technology 11 (2021): 87-97.

[43] Muppalaneni, Naresh Babu, et al. "Machine learning algorithms with ROC curve for predicting and diagnosing the heart disease." Soft computing and medical bioinformatics (2019): 63-72.

[44] Valarmathi, R., and T. Sheela. "Heart disease prediction using hyper parameter optimization (HPO) tuning." Biomedical Signal Processing and Control 70 (2021): 103033.

[45] Budholiya, Kartik, Shailendra Kumar Shrivastava, and Vivek Sharma. "An optimized XGBoost based diagnostic system for effective prediction of heart disease." Journal of King Saud University-Computer and Information Sciences 34.7 (2022): 4514-4523.

[46] Kumari, Saloni, Deepika Kumar, and Mamta Mittal. "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier." International Journal of Cognitive Computing in Engineering 2 (2021): 40-46.

[47] Mienye, Ibomoiye Domor, Yanxia Sun, and Zenghui Wang. "An improved ensemble learning approach for the prediction of heart disease risk." Informatics in Medicine Unlocked 20 (2020): 100402.

[48] Goutte, Cyril, and Eric Gaussier. "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation." *Advances in Information Retrieval: 27th European Conference on IR Research, ECIR 2005, Santiago de Compostela, Spain, March 21-23, 2005. Proceedings 27*. Springer Berlin Heidelberg, 2005.