Towards Diabetes Mellitus Prediction Based on Machine-Learning

Houda El Bouhissi
LMED Laboratory
Faculty of Exact Sciences
University of Bejaia
Bejaia, 06000, Algeria
houda.elbouhissi@gmail.com

Rafa E. Al-Qutaish
Senior Researcher,
ADP
MBZ, , PO Box: 13027
Abu Dhabi, UAE
rafa@ieee.org

Amine Ziane
LMED Laboratory
Faculty of Exact Sciences
University of Bejaia
Bejaia, 06000, Algeria
amine.ziane@univ-bejaia.dz

Kamal Amroun
LIMED Laboratory
Faculty of Exact Sciences
University of Bejaia
Bejaia, 06000, Algeria
kamal.amroun@univ-bejaia.dz

Nabila Yaya
Faculty of Exact Sciences
University of Bejaia
Bejaia, 06000, Algeria
nabila.yaya@univ-bejaia.dz

Melissa Lachi
Faculty of Exact Sciences
University of Bejaia
Bejaia, 06000, Algeria
melissa.lachi@univ-bejaia.dz

2023 International Conference on Smart Computing and Application (ICSCA) | 979-8-3503-4705-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICSCA57840.2023.10087782

***Abstract*— Diabetes is a chronic pathology caused by a disorder of the pancreas, which leads to a high concentration of sugar in the blood and can affect the functioning of the body system at. This disease may cause damage to the heart, blood vessels, eyes, kidneys, and nerves. Therefore, the development of a suitable system for effectively earlier diagnosing diabetic patients using personal, historical, and medical information is required. This system can assist patients in preventing this disease and its complications. Several machine-learning techniques were used for the predictive analysis of diabetes. In this paper, we conduct a review of the most important works related to diabetes prediction and propose an approach for the prediction of gestational diabetes using Deep Neural Network (DNN), Support Vector Machine (SVM), and Random Forest (RF) classifiers. The experiment was conducted using a real dataset from Frankfurt Hospital indicating that the Random Forest algorithm provides more accuracy.**

Keywords*— *Machine-Learning, DNN, RF, SVM, Prediction, Diabetes, Pregnancy

I. Introduction and Motivation

Diabetes mellitus is a chronic pathology caused by a disorder of the pancreas according to the natural regulation of blood sugar levels. The World Health Organization (WHO) states that diabetes is one of the world's leading killers, along with high blood pressure and smoking [1]. This disease is a significant public health problem, and despite prevention efforts, the pandemic continues.

Two types of diabetes are distinguished [2]: type 1, called "insulin-dependent diabetes", which is accompanied by a lack of insulin that can develop into diabetic ketoacidosis. The second type is: Type 2 also called "non-insulin-dependent diabetes" or "adult-onset diabetes". In addition, there is another type of diabetes that concerns women, called "gestational diabetes". Gestational diabetes concerns pregnant women who have previously had no history of diabetes.

Gestational Diabetes is one of the most frequent maladies in women during their pregnancy due to different clinical, social, and lifestyle factors. Machine learning (ML) has the potential to predict the occurrence of this disease based on hidden features in the data.

Diabetes is diagnosed in the healthcare industry based on three blood glucose measurements: fasting, glucose tolerance, and random blood glucose levels [3]. Nowadays, diabetes is spreading quickly among people, especially young people, and has grown to be a significant concern for researchers, scientists, and educators [4].

However, as the number of patients suffering from this disease rises, it becomes more and more challenging to control it. In 2017, there were about 451 million adults in the world with diabetes receiving the necessary treatment. The healthcare community predicts that by 2045 there will be approximately 693 million diabetic patients worldwide and that half of this population will be undetected [5].

Researchers have implemented several systems that help to reduce the risk of infection with early prediction and knowledge of the most critical factors that control it.

ML algorithms are widely used for prediction and recommendation in healthcare and provide good results.

Recently, many algorithms have been employed to predict diabetes, including the main known ML approaches [6], such as Logistic Regression(LR), support vector machine (SVM), and Decision Tree (DT).

Indeed, in our work, we are interested in exploring existing solutions for predicting this disease using ML algorithms to help prospective people to predict whether he/she has diabetes in future or not based on personal features.

In the next section, we summarize the most important related works in the literature. In section 3, we describe in detail the data and proposed method with a focus on processing methods and ML algorithms used to create the predictive model. Section

4 presents the results of the proposed system and discusses its evaluation. Finally, Section 5 concludes the paper and outlines future projects.

II. Related Works

Prediction is a mathematical process for predicting future events or outcomes by analyzing patterns in a data set. The prediction has interested researchers in many fields such as road safety [7], business, and health. Diabetes prediction is the most discussed topic by researchers, as there are several works on it. The authors in [2] [4] conducted a comprehensive review of numerous diabetes prediction models. In this section, we have summarized the most important works on diabetes. In these studies, they used accuracy as a factor of comparison between the algorithms.

This section is dedicated present the main diabetes prediction algorithms for healthcare.

VijiyaKumar et al. [8] proposed a framework that can perform early prediction of diabetes with higher accuracy by using the RF algorithm. According to the authors, the proposed model gives the best results for diabetes prediction.

Wolde Michael et al. [9] suggested data mining techniques with the SVM algorithm for diabetes prediction. The proposed neural network architecture involves three different layers. The first layer called input layer includes eight neurons, the second is a hidden layer contains six neurons, and the last layer called output layer.

Mujumdar et al. [10] proposed an approach based on LR. The authors used this methodology to detect diabetes risk factors. The authors used four classifiers: Naive Bayes (NB), DT, Adaboost (AB), and RF to predict diabetic patients.

Islam et al. [11] present a novel technique for predicting early diabetes based on data mining techniques. For training purposes, the authors used percentage split and 10-fold cross-validation approaches. Through questionnaires, they gathered information from 529 patients, both diabetic and not, at a hospital in Bangladesh.

To predict the early appearance of diabetes in women, Malik et al [12] compared first data mining and machine learning proposals.

To predict the early appearance of diabetes in women, Malik et al. [12] compared first data mining and machine learning approaches. Next, the authors proposed a model for diabetes prediction with conventional ML algorithms using a German hospital's diabetes dataset. The empirical results demonstrate that K-nearest neighbor, RF, and DT are promising regarding the other traditional algorithms.

Cherradi et al. [13] applied and evaluated four ML algorithms (DT, K-Nearest Neighbors, Artificial Neural Network, and DNN to predict diabetes mellitus. These techniques were trained and tested on the Pima Indian database [14].

Karimi Darabi et al. [15] used eight different ML algorithms (LR, Nearest Neighbor, DT, RF, SVM, Naive Bayesian, Neural Network, and Gradient Boosting). The authors evaluated their proposal using parameters such as accuracy, sensitivity, and specificity. The model based on the gradient boosting algorithm showed better performance with a prediction accuracy of 95.50%.

Trivedi et al. [16] proposed a healthcare recommendation system that can predict health status by assessing a patient's lifestyle, and physical and mental health aspects using a deep learning model.

Talha et al. [17] implemented the methods: artificial neural network (ANN), RF, and K-means clustering to predict whether a patient has diabetes or not. The results indicate a strong association of diabetes with body mass index (BMI) and glucose levels.

Sarwar et al. [18] used six ML algorithms to predict diabetes in patients such as Nearest Neighbors (KNN), SVM classifier, LR, DT Classifier (DT), Gaussian Naive Bayes (NB), and RF. All these algorithms were applied to Dataset PIMA Indian, comprising 768 records and 9 attributes.

Many approaches using ML and deep learning algorithms have been proposed for early diabetes prediction. These works analyze the data to achieve better results in terms of precision, recall and accuracy. The studies considered blood glucose level, age and body mass index (BMI).

The significant advantages of the ML algorithms used for diabetes prediction are that it is easy to use and understand, do not require much data for proper functioning and that training is easy and efficient for different models. However, these approaches have several drawbacks, such as the model being challenging to analyze and understand how it works.

Among the techniques used, the DNN has shown the highest accuracy. The major advantages of this method are efficient execution and improvement in performance. On the other hand, it is an expensive technology to implement and requires a large amount of computing power.

Our approach is an analysis study inspired by the related works. In this proposal, we select the most important ML algorithms that provide best results in the literature namely: DNN, SVM and RF and identify which of these three algorithms provides good results. We aim to predict early gestational diabetes with better accuracy. In addition, to our knowledge, few studies have addressed gestational diabetes, since it is a (temporary) disease and only affects pregnant women.

III. Proposed Methodology

Our work consists of early prediction of gestational diabetes for a person to discover if he/she is at risk of developing diabetes in the future with a well-defined prediction rate.

The diabetic prediction process is depicted in Figure 1. Furthermore, it highlights the ML methods used for diabetes prediction.

The proposal involves different steps for better results. These steps are: "Data collection", which concerns the dataset used for our experiment. "Data Preprocessing", involves sub-steps for processing data such as cleaning, and transformation. "Data selection" involves preparing data for the next step. "Data Modeling" where will use 3 ML models and then choose the best-performing model.

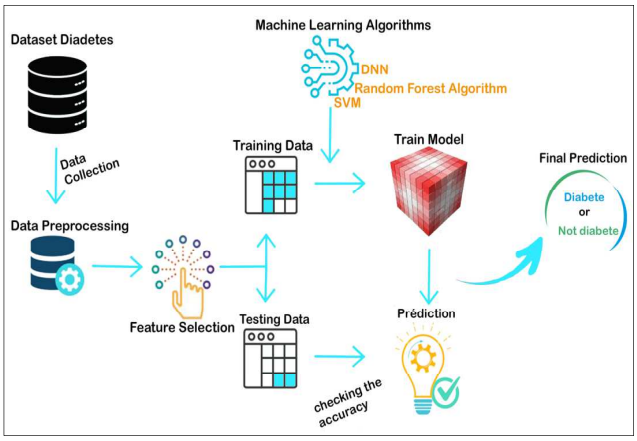

Fig. 1. Proposed System Architecture

A. Data Collection

Data collection is a crucial step for diabetes prediction. This step allows good processing and evaluation of the proposal to achieve quality results.

The diabetes dataset used for our proposal consists of several predictor variables (Table 1) and is in CSV format. The dataset is initially collected from the hospital of Frankfurt, Germany [19], is 62.06 kB, and involves 2000 persons between diabetic and healthy persons.

The dataset structure, which has seven (7) attributes, is shown in table1 and the names of the attributes are Glucose, Pregnancy, Blood Pressure, Skin Thickness, Insulin, Diabetes Paradigm Function, and Age.

TABLE I. ATTRIBUTES DESCRIPTION

Attribute	Description
Glucose	Called also blood sugar.
Pregnancies	Number of pregnant times
Blood Pressure	If a diastolic BP>90 means high blood pressure which increases the probability of diabetes. BP <60 means low blood pressure (low probability of diabetes)
Skin Thickness	The estimated value for body fat. Normal splicing of the skin fold in women is 23 mm. A higher splice leads to obesity, which raises the risk of developing diabetes.
Insulin	Weight (in kg/height in m2): BMI between 18.5 and 20 is normal and between 25 and 30 is in the overweight range and 30 or more is in the obese range
Diabete Predigme Function	Provides information on the parental history and genetic and genetic relationship to patients. This function means that the patient is more likely to suffer from diabetes
Age	Age of a person in years

B. Data Preprocessing

After data collection, the next step is preprocessing, which is very important to extract a perfect dataset for providing quality results. Datasets may contain fluctuating data as well as missing or noisy information. The ML algorithms fail to provide better results if the data quality is poor.

This pre-processing involves cleaning, and transformation.

- The Data cleaning involves different phases, such as removing noisy data and missing values, deleting duplicates and irrelevant data. To address discrepancies [20], we eliminate anomalies in noisy data. For instance, if an attribute such as glucose had zero values, all the values were replaced with the median value of that attribute.
- Data transformation concerns data smoothing, normalization, and aggregation [21]. It is essential in this phase to transform and prepare data for the training phase. For example, normalizing an attribute involves converting each upper case letter to lower case, removing numbers and punctuations from letters. Data smoothing Data smoothing is an important statistical technique used when we deal with ML algorithms. This technique removes outliers from a dataset to make a model more visible.

Significant attributes are selected and converted into bins to achieve the preprocessing task after data cleansing.

C. Data Selection

This method consists in dividing the dataset into two parts: the training part on which the model is trained and the testing part on which the model is tested and the performance of the selected classifiers is evaluated. If a model performs better in both datasets, then the expected accuracy is better.

D. Data Modeling

After data splitting, the next step is data modeling for diabetes prediction. This step involves the implementation of various ML algorithms. Three models were used for early gestational diabetes prediction (DNN, SVM and RF).

For early diabetes prediction, our proposal includes the following three known and powerful ML [22]:

1. DNN

DNN has become a promising solution for injecting artificial intelligence into our daily lives,

from autonomous cars, smartphones, games, drones, etc.

DNNs are improved versions of the conventional ANN with multiple layers. DNN models have recently become very popular due to their excellent performance in learning nonlinear input-output mapping, and the underlying structure of the input data vectors.

DNN is a widely used model given its results and ability to perform complex tasks and the efficient processing of very large data.

2. SVM Classifier

For classification and regression issues, SVM is one of the most widely used supervised learning algorithms.

SVM is a classifier representing data as points and using the kernel method to classify efficiently nonlinear data.

In the future, to classify efficiently new data points, the SVM algorithms determine the optimal boundary or decision line that can divide the n-dimensional space into classes.

3. RF classifier

The RF approach, which combines three predictors, is a robust, rapid, and easy ML algorithm. Most of the time, RF yields satisfactory outcomes.

RF is an ML-based classifier for building decision trees. RF is a supervised ML algorithm widely used in classification and regression problems.

The advantage of this classifier is that it solves the overfitting problem because the output is based on the majority or average voting.

IV. Implementation and Evaluation

To validate our proposal, we implement a software tool in python. The dataset used for testing and validation is collected from the hospital of Frankfurt, Germany, and includes 2000 diabetic and non-diabetic patients.

Different classification methods were used on the dataset, and the outcomes for each methodology varied due to the distinct working requirements of each algorithm.

The model results were evaluated for performance evaluation, a necessary step in testing model quality to ensure the effectiveness of predictive model results.

To evaluate the efficiency of our approach, we use metrics that are widely used in the literature [24] [25]: "Precision" and "Recall". Both precision and recall are based on an understanding and measure of relevance.

Recall (formula 1) focuses only on customers who completed and indicates the share of false negatives. False negatives are customers who cancel but are not identified by the score. In concrete terms, these are customers that you do not detect and for whom you cannot act to prevent them from leaving.

$$\text{Recall} = \frac{\text{Correctly classified person}}{\text{Total used classified}} \quad (1)$$

Precision (formula 2) is the second indicator, complementing recall, and focuses only on persons for whom the model has predicted a termination and gives an indication of false positives. False positives are persons for whom the score indicated an ending but remained subscribed.

$$\text{Precision} = \frac{\text{Correctly classified person}}{\text{Total classified}} \quad (2)$$

Another interesting metric called the f1-score (formula 3) is the combination of precision and recall. The F1 score is considered a harmonic mean of precision and recall, with the best value being 1 and the poorest being 0.

$$F1 - Score = 2 * \frac{(Precision * Recall)}{(Precision + Recall)} \quad (3)$$

The table below (table 2) represents the precision, recall, and F1-score for the RF, SVM classifiers, and DNN. The Precision, Recall, and F1-Score columns show the classification records and are computed according to formulas 1, 2, and 3.

The prediction column is the main interesting column, and returns the prediction result: Positive indicates that the person is likely to have diabetes, and Negative means not.

TABLE II. Comparative table of precision, recall, and F1 score of the 3 classifiers RF, DNN, and SVM.

Classifier	Prediction	Precision	Recall	F1-Score
DNN	Positive	0.81	0.88	0.84
	Negative	0.93	0.89	0.91
RF	Positive	0.94	0.96	0.95
	Negative	0.98	0.97	0.97
SVM	Positive	0.73	0.58	0.65
	Negative	0.80	0.89	0.84

The figures 2, 3, and 4 show the ROC curves of the SVM, RF, and the DNN classifiers. The ROC curve of the RF classifiers covers 96% of the site beyond the diagonal baseline. The ROC curve of the SVM classifiers covers 73% of the area beyond the diagonal baseline. The ROC curve of the DNN Algorithm covers 89% of the site that exceeds the slanted baseline.

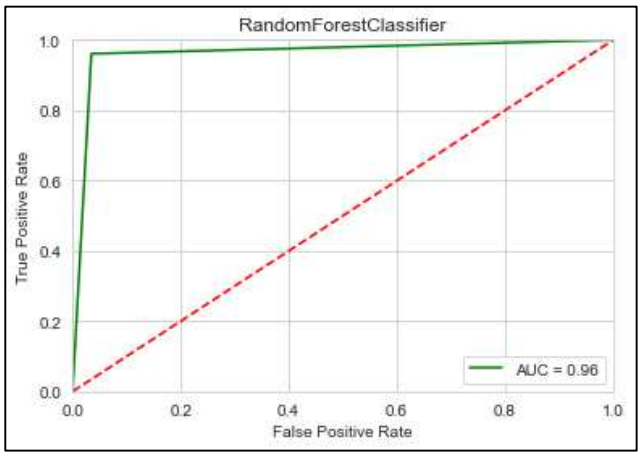

Fig. 2. RF Classifier Curve

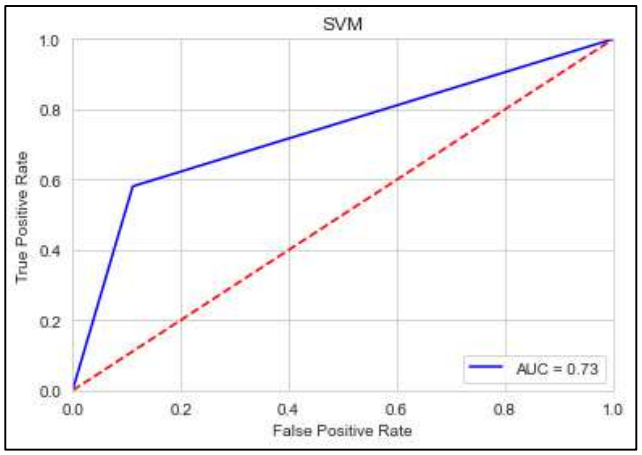

Fig. 3. SVM Classifier Curve

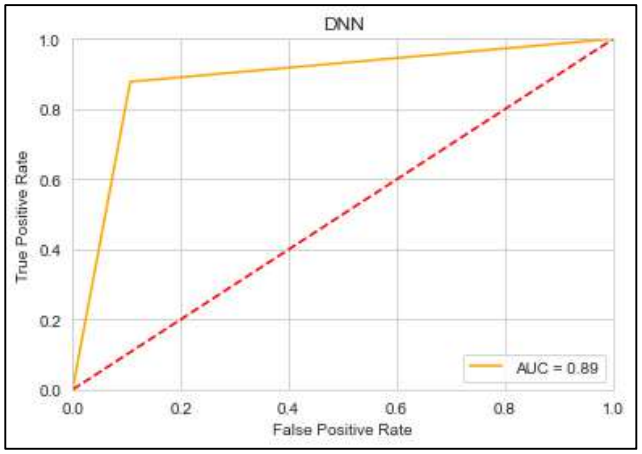

Fig. 4. DNN classifier Curve

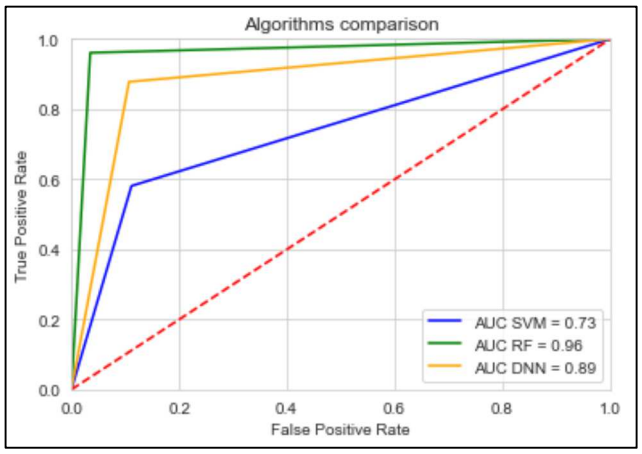

Fig. 5. Comparison of classifier accuracy

Figure 5 shows the performance comparison of the classification algorithms using the ROC curve. The higher the values of a curve, the larger area under the curve, and the less error the classifier makes. The RF classifier achieves an average accuracy of 96% better than DNN (89 %) and SVM (73%).

To conclude, we realize that the RF classifier performs better for the dataset than the DNN and SVM classifiers. It is important to consider multiple datasets to have better accuracy for the prediction of diabetes in patients.

V. Conclusion and Future Works

Predictive analytics in healthcare can change the researchers' vision and medical practitioners to obtain information from medical data and make decisions. Diabetes mellitus is a disease, which can cause many complications. The prediction and diagnosis of this disease is a hot topic that is worth studying by many researchers to provide predictive systems to assist people in predicting gestational diabetes and reduce the risks of complications that occur from diabetes. Techniques such as ML and data mining help diagnose diseases. Early diabetes detection is crucial in determining the patient's proper course of therapy.

In this paper, we conduct a literature review of the related works regarding diabetes disease and propose a hybrid method that involves three ML algorithms for diabetes awareness. The dataset we used to implement our approach is gathered from the hospital in Frankfurt, Germany. Our experiments show that the use of these three algorithms (DNN, SVM, and RF) provides better accuracy.

In our future work, we also plan to enhance the precision by applying different classifications algorithms and other attributes such as physical inactivity and smoking habit.

References

[1] Diabetes. Retrieved, September 2022, from https://www.who.int/health-topics/diabetes.

[2] R. Saxena, S. K. Sharma, M. Gupta, and G. C. Sampada, “A Comprehensive Review of Various Diabetic Prediction Models: A Literature Survey,” Journal of Healthcare

Engineering, vol. 2022, pp. 1–15, Apr. 2022, doi: 10.1155/2022/8100697.

[3] A. C. Berger et al., "A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers," Cancer Cell, vol. 33, no. 4, pp. 690-705.e9, Apr. 2018, doi: 10.1016/j.ccell.2018.03.014.

[4] V. Jaiswal, A. Negi, and T. Pal, "A review on current advances in machine learning based diabetes prediction," Primary Care Diabetes, vol. 15, no. 3, pp. 435–443, Jun. 2021, doi: 10.1016/j.pcd.2021.02.005.

[5] N. H. Cho et al., "IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045," Diabetes Research and Clinical Practice, vol. 138, no. 1, pp. 271–281, Apr. 2018, doi: 10.1016/j.diabres.2018.02.023.

[6] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research," Computational and Structural Biotechnology Journal, vol. 15, pp. 104–116, 2017, doi: 10.1016/j.csbj.2016.12.005.

[7] R. Bekka, S. Kherbouche, and H. El Bouhissi, "Distraction Detection to Predict Vehicle Crashes: A Deep Learning Approach," Computación y Sistemas, vol. 26, no. 1, Mar. 2022, doi: 10.13053/cys-26-1-3871.

[8] P. Vijayakumar, R. G. Nelson, R. L. Hanson, W. C. Knowler, and M. Sinha. "HbA1c and the Prediction of Type 2 Diabetes in Children and Adults". Diabetes Care, vol. 40, no. 1, pp. 16–21, Jan. 2017, doi: 10.2337/dc16-1358.

[9] F. G. Woldemichael and S. Menaria, "Prediction of Diabetes Using Data Mining Techniques," 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI), May 2018, doi: 10.1109/icoei.2018.8553959.

[10] A. Mujumdar and V. Vaidehi, "Diabetes Prediction using Machine Learning Algorithms," Procedia Computer Science, vol. 165, pp. 292–299, 2019, doi: 10.1016/j.procs.2020.01.047.

[11] M. M. F. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, "Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques," Computer Vision and Machine Intelligence in Medical Image Analysis, pp. 113–125, Aug. 2019, doi: 10.1007/978-981-13-8798-2_12.

[12] S. Malik, S. Harous, and H. El-Sayed, "Comparative Analysis of Machine Learning Algorithms for Early Prediction of Diabetes Mellitus in Women". Springer Link, 2021. https://link.springer.com/chapter/10.1007%2F978-3-030-58861-8_7 (accessed Nov. 25, 2021).

[13] O. Daanouni, B. Cherradi, and A. Tmiri, "Diabetes Diseases Prediction Using Supervised Machine Learning and Neighbourhood Components Analysis". Proceedings of the 3rd International Conference on Networking, Information Systems & Security, Mar. 2020, doi: 10.1145/3386723.3387887.

[14] Pima Indians Diabetes Database. Retrieved October 2022, from https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.

[15] P. Karimi Darabi, M.J. Tarokh, "Type 2 Diabetes Prediction Using Machine Learning Algorithm". Jorjani Biomedicine Journal. 2020; 8(3): 4-18.

[16] N. K. Trivedi, V. Gautam, H. Sharma, A. Anand, and S. Agarwal, "Diabetes prediction using different machine learning techniques," 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), 2022.

[17] Talha Mahboob Alam, Muhammad Atif Iqbal, Yasir Ali, Abdul Wahab, Safdar Ijaz, Talha Imtiaz Baig, Ayaz Hussain, Muhammad Awais Malik, Muhammad Mehdi Raza, Salman Ibrar, Zunish Abbas, A model for early prediction of diabetes, Informatics in Medicine Unlocked, Volume 16, 2019, 100204.

[18] M. A. Sarwar, N. Kamal, W. Hamid, and M. A. Shah, "Prediction of Diabetes Using Machine Learning Algorithms in Healthcare," IEEE Xplore, Sep. 01, 2018. https://ieeexplore.ieee.org/abstract/document/8748992.

[19] Frankfurt Hospital. Retrieved October 2022, from https://www.kaggle.com/datasets/johndasilva/diabetes.

[20] N. Z. Abidin, A. Ritahani, and N. A., "Performance Analysis of Machine Learning Algorithms for Missing Value Imputation". International Journal of Advanced Computer Science and Applications, vol. 9, no. 6, 2018, doi: 10.14569/ijacsa.2018.090660.

[21] B. Malley, D. Ramazzotti, and J. T. Wu, "Data Pre-processing," *PubMed*, 2016. https://www.ncbi.nlm.nih.gov/books/NBK543629/.

[22] Mohanty, Sachi & Chatterjee, Jyotir & Jain, Sarika & Elngar, Ahmed & Gupta, Priya. (2020). Recommender System with Machine Learning and Artificial Intelligence. 10.1002/9781119711582.

[23] "Recommender System with Machine Learning and Artificial Intelligence: Practical Tools and Applications in Medical, Agricultural and Other Industries | Wiley," Wiley.com.

[24] H. El Bouhissi, M. Adel, A. Ketam, and A.-B. Salem, "Towards an Efficient Knowledge-based Recommendation System." Accessed: Nov. 30, 2022. [Online]. Available: https://ceur-ws.org/Vol-2853/paper1.pdf.

[25] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, Jul. 2009, doi: 10.1016/j.ipm.2009.03.002.