A Comparative analysis of Early Stage Diabetes Prediction using Machine Learning and Deep learning Approach

This paper was downloaded from TechRxiv (<https://www.techrxiv.org>).

LICENSE

CC BY 4.0

SUBMISSION DATE / POSTED DATE

26-10-2021 / 01-11-2021

CITATION

Refat, Md. Abu Rumman; Amin, Md Al; Kaushal, Chetna; Yeasmin, Mst. Nilufa; Islam, Md Khairul (2021): A Comparative analysis of Early Stage Diabetes Prediction using Machine Learning and Deep learning Approach. TechRxiv. Preprint. <https://doi.org/10.36227/techrxiv.16870623.v1>

DOI

[10.36227/techrxiv.16870623.v1](https://doi.org/10.36227/techrxiv.16870623.v1)

A Comparative analysis of Early Stage Diabetes Prediction using Machine Learning and Deep learning Approach

Md Abu Rumman Refat

Information & Communication Technology Computer Science and Engineering Chitkara University Institute of Engineering and Technology, Chitkara University
Islamic University
Kushtia, Bangladesh.
refatiuice@gmail.com

Md. Al Amin

Prime University
Dhaka, Bangladesh
ahmedalamin2357@gmail.com

Chetna Kaushal

Chitkara University Institute of Engineering and Technology, Chitkara University
Punjab, India.
chetnakaushal3558@gmail.com

Mst. Nilufa Yeasmin

Information & Communication Technology
Islamic University
Kushtia, Bangladesh.
nilufayeasmin299@gmail.com

Md Khairul Islam

Information & Communication Technology
Islamic University
Kushtia, Bangladesh
mdkito51@gmail.com

Abstract—diabetes is a disease that affects how your body processes blood sugar and is often referred to as diabetes mellitus. Insulin insufficiency and ineffective insulin use coincide when the pancreas cannot produce enough insulin or the human body cannot use the insulin that is produced. Insulin is a hormone produced by the pancreas that helps in the transport of glucose from food into cells for use as energy. The common effect of uncontrolled diabetes is hyper-glycemia, or high blood sugar, which plus other health concerns, raises serious health issues, majorly towards the nerves and blood vessels. According to 2014 statistics, people aged 18 or older had diabetes and, according to 2019 statistics, diabetes alone caused 1.5 million deaths. However, because of the rapid growth of machine learning(ML) and deep learning (DL) classification algorithms. indifferent sectors, like health science, it is now remarkably easy to detect diabetes in its early stages. In this experiment, we have conducted a comparative analysis of several ML and DL techniques for early diabetes disease prediction. Additionally, we used a diabetes dataset from the UCI repository that has 17 attributes, including class, and evaluated the performance of all proposed machine learning and deep learning classification algorithms using a variety of performance metrics. According to our experiments, the XGBoost classifier outperformed the rest of the algorithms by approximately 100.0%, while the rest of the algorithms were over 90.0% accurate.

Index Terms—Diabetes prediction; Machine learning; Deep learning; Classification;

I. INTRODUCTION

Diabetes is a dangerous medical condition that results in blood sugar levels that are greater than normal. Diabetes develops when the body cannot effectively use its insulin, a hormone made by special cells in the pancreas called islets (eye-lets) [1]. It is the leading cause of heart disease, stroke, amputation, kidney failure, blindness, and premature death [2]. The pre-valance of diabetes is increased day by day because of several reasons like as too much salt eating , unhealthy

diet, overweight or obesity, immune system disorders, insulin resistance, excessive stress, genetic , etc. [3]. Diabetes affected approximately 463 million persons aged 20 to 79; scientists predict in 2045 it will rise to 700 million [4]. Diabetes caused 4.2 million deaths and type 1 diabetes affects more than 1.1 million children and adolescents, and it as well as 374 million people are at increased risk of developing type 2 diabetes. Type-1, type-2 most common types of diabetes, and gestational diabetes.

The prevention of diabetes is not so easy but there are several way to manage diabetes. The first and foremost way is to early diagnosis which can reduce the further risk. There are diverse way to early diagnosis in medical science but in this section ML also play a vital role to identify diabetes at early stage. Currently researchers have shown many positive results in the risk prediction of diabetes. Machine learning techniques like as Support Vector Machine, K-Nearest Neighbors, Random Forest, Decision Tree, Logistic Regression, XGBoost work better in early stage prediction . The rapid development of deep learning techniques and the availability of data, it is possible to predict diabetes related task: Diabetes diagnosis, management of glucose levels, and evaluation of complications related to diabetes. Diabetes detection is routinely aided by modern deep learning algorithms and frameworks [5].

Although accuracy is the way that can be used to find out the optimal accuracy and also help to find out the best algorithm, others performance metrics like precision, recall, f1-score, execution time and ROC value should be used to find the accurate one.

Early stage detection may help to the patients for proper and timely treatment for diabetic. In order to classify early-stage diabetes as either positive or negative, we implemented a classification algorithm on that dataset. Hence, we did a

comprehensive analysis is find out the best algorithm among XGBoost classifier, Decision tree, random Forest, Support Vector machine, Multi-layer Perceptrons, Logistic Regression. We used a benched-mark UCI repository dataset [6] that has 16 attributes and 520 data were assigned 416 data for training and 104 data for testing. Furthermore, we compare the performance using various evaluation metrics: precision, recall, f1-score, execution time and ROC value. The analysis show that XGBoost classifier is the best for predicting diabetes at initial stage which has about 99.99% training accuracy with 99.0% testing accuracy.

The major research objectives of this study include:

- 1) Publicly available dataset and their availability in the field of diabetic.
- 2) Performed a comprehensive analysis between ML and DL techniques.
- 3) Performance metrics evaluation that are employed to detect diabetes at early stage.
- 4) Future research direction, that needs to be addressed by the future researchers working in the area.

The remainder of the research paper is structured as follows: A short outline of the related literature is presented in section II. A brief overview of the research methodology is given in section III. Experimental Results are presented in IV section. Finally, the conclusion and future work are presented in V section.

II. LITERATURE REVIEW

We highlight a few researchers in this section who have made significant contributions to the study of diabetes mellitus prediction by extracting knowledge from public medical data using machine learning and deep learning approaches.

For an instant, Lin et al. [7] has examined Naive Bayes, SVM, and ANN classifiers using a diabetes dataset in their research. They carried out a based on weight-adjusted research in which The almost of votes concluded the probability of diabetes prediction. And finally, they concluded that combining models provides a higher level of classification accuracy than any single model. Kandhasamy et al. [8] proposed a predictive analytic model using J48 (C4.5), K-Nearest Neighbors classifier, Random Forest classifier, and SVM on the diabetes dataset. Before pre-processing the data, they premeditated that the J48 algorithm performed better than others by the accuracy of 73.82%, whereas KNN and Random Forest performed better accuracy after pre-processing. Agrawal et al. [9] developed a hybrid model for diabetes risk factor prediction using CNN, KNN, SVM, SVM+LDA, NB, SVM, ID3, C4.5, CART algorithms, with a dataset of 738 patients. They achieved the highest accuracy of 88.10 percent on the dataset by combining SVM and LDA.

Oh the other hand, Naiarun et al. [10] built a web application for diabetes prediction based on prediction accuracy. They then compared several machine learning and deep learning algorithms for prediction, including Random Forest, Decision Tree, Logistic Regression, Convolutional Neural Network, and Naive Bayes, as well as bagging and boosting. Finally, they

found Random Forest outperformed in both accuracy and ROC value, where the accuracy is 85.55% and the ROC value is 0.912. Kavakiotis et al. [11] evaluated three different machine learning algorithms using a 10-fold cross-validation evaluation approach, where Logistic regression, Naive Bayes, and SVM included. Where SVM outperformed other approaches with a performance accuracy of 84%. Additionally, Zheng et al. [12] used a variety of machine learning techniques to predict diabetes mellitus at an early stage, highlighting areas for improvement in filtering conditions. The various algorithms employed in the experiment included Random Forest, KNN, SVM, Naive Bayes, Decision tree, and Logistic regression, all of which were employed in the cross-validation procedure with 10-fold.

Following that, Ahmed et al. [13] developed a J48 Decision Tree algorithm based model for the treatment of type 2 diabetes patients. In addition to using 7 specific patient attributes, they also examined seven overall control attributes, including age, gender, renal problem, smoking, hypertension, cardiac problem, and diabetes. The accuracy rate of 70.80% and ROC value of 0.624 were finally surpassed by their results. Oleiwi et al. [14] suggested a novel model of classification based-on an early stage diabetes prediction dataset [6] to ensure the use of significant features and provide results that were close to clinical outcomes using machine learning algorithms. They trained three machine learning models, including a Random Forest (RF), a Multi-layer Perceptron (MLP) classifier, and also a Radial Basis Function Network (RBF), to find the best classifier for diabetes mellitus prediction. Meanwhile, they showed that the RBF algorithm outperformed with a 98.80 % accuracy. In another research, Bukhari et al. [15] used an Artificial Neural Networks (ANN) model with a different number of neurons in the hidden layers, ranging from 5 to 50, to predict female diabetes using the PIMA Indians diabetes dataset [16]. They considered eight features to train the model and outperformed 93% accuracy on the validation set.

Most related works improved performance through the use of many machine learning and deep learning classification methods. Most studies used accuracy, precision, recall, F-score, ROC score, and execution time as evaluation metrics, to compare and determine the best optimal approach. To conclude our research, we present a comparison of real diagnostic medical data using prominent machine learning classification algorithms for early stage diabetes mellitus prediction based on several risk factors.

III. METHODOLOGY

Our proposed technique comprises four components, as illustrated in the Figure 1, which work together to accomplish our research objective. To begin, we collected the diabetes dataset and then pre-processed it. After pre-processing, the dataset was divided into train and test sets using a tenfold cross-validation procedure. Then, the suggested algorithms are used on the training set to provide early stage diabetes mellitus prediction, and finally, the comparative analysis of

Fig. 1. Graphical representation of proposed methodology.

the performance is evaluated on the test set using evaluation metrics. These phases will be briefly discussed in this section.

A. Dataset and Attributes

Here, we used the UCI repository diabetes dataset [6] to investigate the efficacy of Machine learning and Deep learning algorithms at earlier stages of diabetes diagnosis. This dataset was collected using a direct questionnaire distributed to 520 patients at the hospital that name is Sylhet Diabetes Hospital in Bangladesh who had recently been diagnosed with diabetes or having diabetes-related symptoms. It comprises 16 attributes with 320 positive and 200 negative instances, with positive and negative indicators used to determine whether a patient is at risk of diabetes or not. The attributes and their corresponding possible values are as follow:

- Age: 20 to 65 years
- Sex: 1 represents Male and 0 represents Female
- Polyuria: 1 means Yes and 0 means No
- Sudden Weight loss : 1 and 0 represent Yes and No respectively
- Weakness : 1 represents Yes and 0 represents No
- Polyphagia : 1 and 0 means Yes and No
- Genital thrush : 1 equal to Yes and 0 equal to No
- Visual blurring : 1 and 0 means Yes and NO
- Itching : 1 and 0 represents Yes and No respectively
- Irritability : 1 means Yes and 0 means No
- Delayed Healing: 1 denotes yes and 0 denotes No
- Partial Paresis : 1 and 0 denotes Yes and No
- Muscle Stiffness : Yes and No represents 1 and 0 respectively
- Alopecia: 1 means Yes, 0 means No
- Obesity : 1 means Yes and 0 means No
- Class : 1 represents Positive and 0 represents Negative

B. Pre-Processing

The data pre-processing used to achieve our research goal by addressing missing values in the dataset of diabetes. For example, the nominal values of the features are inappropriate for machine learning and deep learning algorithms for diabetes prediction. We convert nominal attribute values to numeric values, such as 1 for male and 0 for female in the sex group, and 1 for Yes and 0 for No in the remaining attributes group, where 1 indicates a positive value and 0 indicates a negative value in the class group.

C. Classifier Algorithms

1) Machine Learning Classifier:

a) **XGBoost**: XGBoost (XGB) is a convenient and effective deployment of the Gradient Boosted Trees algorithm, and is a believable distributed machine learning platform to scale algorithms for tree boosting. Under a distributed setting for a rapid parallel tree layout, the classifier is well configured as well as fault-tolerant. It combines a single node with tens of millions of samples and billions of distributed software samples that are scale beyond [17].

$$y = \sum_{i=1}^n (w_i \cdot x_i) \quad (1)$$

b) **Random Forest**: In regression and classification problems, a random forest (RF) is a classifier made up of a collection of tree-structural classifiers, all of which receive random independent vectors that are identically distributed and each tree vote for the most popular class [18]. This classifier is a simple, flexible machine learning algorithm that reliably produces excellent results most of the time even without tuning hyper parameters.

c) **Decision Tree**: A decision tree (DT) is a supervised learning technique that is mostly used for classification tasks and is also helpful in discovering relevant features and patterns in large databases, thus serving as a powerful means of discrimination and predictive modeling [19]. Each node in a Decision tree, or Decision node and Leaf node, consists of a test for an attribute, which creates one of the two branches in the tree that descends from that node.

d) **Support Vector Machine**: When it comes to supervised machine learning models, Support Vector Machine (SVM) is useful because it allows us to take into account more complex relationships between our data points while still using a classification algorithm. The algorithm is specifically geared towards text classification problems, where only a handful of tagged samples is required, and is making progress in several biological fields [20].

e) **K-Nearest-Neighbours**: A supervised machine learning algorithm that is easy to implement and can be used to solve both classification and regression problems is called the k-nearest neighbors (KNN) algorithm. The K-NN algorithm classifies new cases and data based on the similarity with previously studied cases and data, with the new case first entering the category that is most similar to the previous

categories. This classification approach should be utilized whenever there is hardly any previous experience with data distribution [21].

f) **Logistic Regression:** A Logistic Regression (LR), also known as a Logit Model, is a statistical method that models binary response variables, and it is a type of logistic regression. Both logistic regression and logistic regression incorporate linear regression into their models because it predicts the likelihood of one outcome class compared to another (for example, favorable treatment or unpleasant treatment) [22]. The simple logistic model can be defined as [23]

$$\text{logit}(Y) = \text{naturallog}(\text{odds}) = \ln \frac{\pi}{\pi - 1} = \alpha + \beta X \quad (2)$$

where the regression co-efficient β is the logit.

2) Deep Learning Classifier:

a) **Multi-layer Perceptron:** Multi-layer Perceptron (MLP) classifier is a type of feed-forward artificial neural network (FFANN) that consists of more than two layers; the first layer and last layer are called input layer and output layer respectively. There has also a higher one layer in the middle of the input and output layer called the hidden layer. The time complexity depends on the number of layers; if the number of the layer will increase then time complexity will also increase. For instance, every neuron accepts input like as $x_1, x_2, x_3, \dots, x_n$, and the bias and wight are denoted by (b) and (w) ; Input multiplied by wight and produced output which may be denoted y based on the activation function(ϕ) [24]

$$y = \phi\left(\sum_{i=1}^n ((w_i \cdot x_i) + b)\right) \quad (3)$$

b) **Artificial Neural Network:** Artificial neural networks (ANNs) are designed to mimic the way the human brain processes and analyzes information. Their models, created based on the brain's methods, can be used to simulate complex patterns and prediction problems. An overall architecture for building neural networks has three main components: input, hidden, and output layers. The frequency stream in feed-forward networks is from internal to external components, going in a straight line [25]. We can define ANN as look like [26]

$$z = f(b + x \cdot w) = f\left(b + \sum_{i=1}^n (x_i \cdot w_i)\right) \quad (4)$$

where $x \in (1 \times n)$, $w \in (n \times 1)$, $b \in (1 \times 1)$ as well as $z \in (1 \times 1)$

In the above equation, b is bias, w denotes weights, the input and output node denotes x and z respectively.

c) **Long Short-Term Memory(LSTM):** LSTM networks are a category of recurrent neural network that was specifically developed to model sequences and their long-range dependencies and to accurately model sequences as a result. This model incorporates a variety of deep learning algorithms. The introduction to LSTMs has its difficulties, as well as bidirectional and sequence-to-sequence, two concepts closely

related to LSTMs, which means that there is no relationship between the length of the input and the capacity of the system and then each time step, the execution time per weight is constant $O(1)$ [27]. The input sequence $x = (x_1, \dots, x_T)$ and output sequence $y = (y_1, \dots, y_T)$ of an LSTM network by using these formulas to find the network unit activation starting from $t = 1$ to T : [27]

$$\begin{cases} i_t = \sigma(W_{ix}x_t + W_{im}m_{t-1} + W_{ic}c_{t-1} + b_i) \\ f_t = \sigma(W_{fx}x_t + W_{fm}m_{t-1} + W_{fc}c_{t-1} + b_f) \\ o_t = \sigma(W_{ox}x_t + W_{om}m_{t-1} + W_{oc}c_{t-1} + b_o) \\ c_t = f_t \odot c_{t-1} + i_t \odot g(W_{cx}x_t + W_{cm}m_{t-1} + b_c) \\ m_t = o_t \odot h(c_t) \\ y_t = \phi(W_{ym}m_t + b_y) \end{cases} \quad (5)$$

weight matrices in which the W terms stand for W_{ix} , W_{fx} , W_{ox} is weight matrices have the geometric shape of a triangle with each side being the same length, and the b component is non-symmetrical. Sigma sigmoid function is represented by the σ , and i, f, o and c are input gates, for example gate, exit gate and cell activation vectors, both of which have the same size as the vector m cell activation, product of the vector with elements wise denoted $\odot$, g and h are the functions for cell input and cell output, and ϕ is the function to trigger the network output.

D. Performance Evaluation Metrics

After cross-validating our proposed approaches, we'll need some tools to evaluate the performance of our proposed approach's. In this paper, we evaluated the performance of our experiments using a set of commonly used evaluation metrics for classification problems. Precision, recall, f1-score, ROC-score, and accuracy are the metrics used to calculate the predictive ability of models [28].

Precision. The number of true positives divided by the sum of true positives and false positives is how precision is calculated:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (6)$$

Recall. The number of true positives divided by the sum of true positives and false negatives is known as recall, and it is calculated as follows:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (7)$$

F1-score. The geometric average of precision and recall is defined as the f1-score. Mathematically:

$$F1 - \text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (8)$$

Accuracy. The number of right predictions divided by the total number of predictions stated below is used to measure accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (9)$$

TABLE I
COMPARED TO NINE MODELS TO REVEAL WHICH ONE IS THE MOST ACCURATE.

Classification Model	Train Accuracy	Test Accuracy	Train Loss	Test Loss	Precision	Recall	F1-Score	Roc-Auc Score	Execution Time
XGB	1.000	1.000	0.000	0.000	1.000	1.000	1.000	1.000	46.074
RF	0.969	0.923	1.079	2.657	1.000	0.882	0.938	0.941	95.798
DT	0.966	0.923	1.162	2.657	0.984	0.897	0.938	0.935	31.116
KNN	0.954	0.846	1.578	5.314	1.000	0.765	0.867	0.882	43.837
SVM	0.942	0.913	1.993	2.989	0.968	0.897	0.931	0.921	51.088
LR	0.947	0.894	1.827	3.653	0.983	0.853	0.913	0.913	65.865
ANN	0.950	0.885	1.744	3.985	0.983	0.838	0.905	0.905	47.197
MLP	0.913	0.885	2.989	3.985	0.924	0.897	0.910	0.879	22.376
LSTM	0.935	0.923	1.744	3.985	0.983	0.838	0.905	0.905	146.61

ROC-AUC Score. An ROC curve, formed by expressing the relationship between the true positive rate (sensitivity or recall) and the false positive rate (false positive rate), is called the receiver operating characteristic (ROC) curve (1-specificity). The ROC-AUC measure is commonly used to evaluate the accuracy of models that assign positive and negative class labels to binary classification problems.

IV. EXPERIMENTAL RESULT

Fig. 2. Bar plot that illustrated to visualize the Evaluation metrics of different models. Here, Test accuracy, F1-score and Roc-score are mapped in order to represent the comparison of different models.

Deep learning and machine learning algorithms are deployed to develop a computer assisted diabetes diagnosis systems using the dataset. In this study, we have introduced six machine learning models such as — XGBoost, Random Forest, Decision Tree, Support Vector Machine, Logistic Regression, and K-Nearest-Neighbors. We have also performed Deep Learning base classification techniques including LSTM, ANN, and MLP. Before applying classification algorithms we have performed pre-processing of each data points in the dataset. Table 1 shows Train Accuracy, Test Accuracy, Train Loss, Test Loss, Precision, Recall, F1-Score, Roc-Auc Score, and Execution Time for all Deep Learning and Machine Learning models so that you can quickly choose the best model while taking into account all scores and losses. Testing accuracy is always a measure of your model's capability to predict the results on new data, whereas training accuracy

Fig. 3. Bar plot illustrated to visualize the loss of different models. Both train and test loss represented to compare different models error.

represents your model's ability to predict the results on seen or training data to develop the model. Loss is the consequence of an incorrect prediction. A model's prediction that ends up being incorrect is considered a loss. If the model's prediction is correct, the damage will be lower; if the prediction is incorrect, there is. The training loss, or loss during training, is known as the training loss, and the test loss, or loss on the test dataset or the unseen dataset, is known as the test loss. In Figure:2, F-1 Score, which shows 3 different studies, you can see F-1 Score, Roc-Score, and Test-Accuracy all listed for 9 different models. And underfitting and overfitting are described in Fig:3, represented by train loss and test loss. XGBoost is the most accurate model for this classification, and it has an exceptional prediction ability, with 100.0 % train and test accuracy, 0.00 train and test loss, 1.00 precision, recall, F1-score, and Roc-Auc accuracy, with an execution time of 46.074. It should be noted that the random forest algorithm is also effective though its test accuracy is 92.3%. However, deep learning models in this study have been less effective when compared to machine learning models. Three different deep learning models (with accuracies ranging from 88.5% to 92.3%) all received satisfactory performance from an artificial neural network. While both models are capable of achieving test accuracy of 0.923, machine learning models Random Forest and long short-term memory-based deep learning can be utilized to

great effect. But, according to this study, other models of deep learning, such as machine learning models, haven't been effective. Although we have achieved an impressive result to classify diabetes patients, we have a huge drawback in our study. There are very few data points available in the dataset. However, our trained model is robust enough to perform on a wide range of similar type unseen data.

V. CONCLUSION AND FUTURE WORK

Despite the lack of conclusive evidence in the scientific literature, age and diabetes do not have a direct relationship. Diabetes must be detected early to ensure optimal treatment, and machine learning and deep learning have revolutionized research into risk prediction for early stage diabetes. In this paper, we investigated diabetes prediction at an early stage using various machine learning and deep learning classification algorithms, along with a variety of diabetes risk factors. The diabetes dataset was evaluated using nine different classification algorithms, including XGB, RF, DT, KNN, SVM, ANN, MLP, and LSTM. In our experiment, XGBoost outperformed near 100.0% and was significantly superior to other machine learning and deep learning approaches for detecting early stage diabetes. The findings of the study may help health care providers in detecting diabetes earlier and making more informed clinical decisions about diabetes management, potentially saving lives. Even though, we can accurately predict diabetes, our research has limitations. The primary limitation of the study is its small sample size, which made it incredibly difficult to establish statistical significance for any of the outcomes.

In future, we intend to collect additional data from around the world to create a more comprehensive dataset that will assist in the prediction of more precise and accurate disease classification. Following, we will focus on identifying additional factors in the dataset that could lead to diabetes detection at an early stage. In addition, our approach can be applied to generalize for any disease prediction, not just diabetes.

REFERENCES

- [1] "Learning about diabetes and its type." [Online]. Available: <https://www.diabetesresearch.org/what-is-diabetes>
- [2] D. Gahlan, R. Rajput, and V. Singh, "Metabolic syndrome in north indian type 2 diabetes mellitus patients: A comparison of four different diagnostic criteria of metabolic syndrome." *Diabetes & metabolic syndrome*, vol. 13, no. 1, pp. 356–362, 2018.
- [3] "Causes of diabetes that must be known." [Online]. Available: <https://markethealthbeauty.com/causes-of-diabetes/>
- [4] "Learning of diabetes facts figures." [Online]. Available: <https://www.idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html>
- [5] T. Zhu, K. Li, P. Herrero, and P. Georgiou, "Deep learning for diabetes: a systematic review," *IEEE Journal of Biomedical and Health Informatics*, 2020.
- [6] M. M. F. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, "Likelihood prediction of diabetes at early stage using data mining techniques," in *Computer Vision and Machine Intelligence in Medical Image Analysis*, vol. 992. Springer, Singapore, 2020, pp. 113–125.
- [7] L. Li, "Diagnosis of diabetes using a weight-adjusted voting approach," in *2014 IEEE International Conference on Bioinformatics and Bioengineering*, 2014, pp. 320–324.
- [8] J. P. Kandhasamy and S. Balamurali, "Performance analysis of classifier models to predict diabetes mellitus," *Procedia Computer Science*, vol. 47, pp. 45–51, 2015.
- [9] P. Agrawal and A. kumar Dewangan, "A brief survey on the techniques used for the diagnosis of diabetes-mellitus." *International Research Journal of Engineering and Technology*, vol. 02(03), 2015.
- [10] N. Naiarun and R. Moungrmai, "Comparison of classifiers for the risk of diabetes prediction," *Procedia Computer Science*, vol. 69, pp. 132–142, 2015, the 7th International Conference on Advances in Information Technology.
- [11] 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.
- [12] T. Zheng, W. Xie, L. Xu, X. He, Y. Zhang, M. You, G. Yang, and Y. Chen, "A machine learning-based framework to identify type 2 diabetes through electronic health records," *International Journal of Medical Informatics*, vol. 97, 2016.
- [13] T. Ahmed, "Developing a predicted model for diabetes type 2 treatment plans by using data mining," *Journal of Theoretical and Applied Information Technology*, vol. 90, pp. 181–187, 2016.
- [14] A. Oleiwi, L. Shi, Y. Tao, and L. Wei, "A comparative analysis and risk prediction of diabetes at early stage using machine learning approach," *International Journal of Future Generation Communication and Networking*, pp. 4151–4163, 2020.
- [15] M. M. Bukhari, B. F. Alkhamees, S. Hussain, A. Gumaie, A. Assiri, and S. S. Ullah, "An improved artificial neural network model for effective diabetes prediction," *Complex.*, vol. 2021, pp. 5 525 271:1–5 525 271:10, 2021.
- [16] "Pima. university of california, irvine, school of information and computer sciences," Oct 6, 2020. [Online]. Available: <https://www.kaggle.com/uciml/pima-indians-diabetes-database>
- [17] T. Chen and C. Guestrin, "Xgboost: Reliable large-scale tree boosting system," in *Proceedings of the 22nd SIGKDD Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA*, 2015, pp. 13–17.
- [18] E. Goel, E. Abhilasha, E. Goel, and E. Abhilasha, "Random forest: A review," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 7, no. 1, 2017.
- [19] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, "An introduction to decision tree modeling," *Journal of Chemometrics: A Journal of the Chemometrics Society*, vol. 18, no. 6, pp. 275–285, 2004.
- [20] W. S. Noble, "What is a support vector machine?" *Nature biotechnology*, vol. 24, no. 12, pp. 1565–1567, 2006.
- [21] L. E. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009.
- [22] J. M. Hilbe, *Logistic regression models*. Chapman and hall/CRC, 2009.
- [23] C.-Y. J. Peng, K. L. Lee, and G. M. Ingersoll, "An introduction to logistic regression analysis and reporting," *The journal of educational research*, vol. 96, no. 1, pp. 3–14, 2002.
- [24] A. Odeh, I. Keshta, and E. Abdelfattah, "Efficient detection of phishing websites using multilayer perceptron," 2020.
- [25] A. Abraham, "Artificial neural networks," *Handbook of measuring system design*, 2005.
- [26] "Neural network model introduction."
- [27] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," 2014.
- [28] D. M. W. Powers, "Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.