

Improving Healthcare Prediction Of Diabetic Patients Using Tri-Ensemble Model With Stacking Classifier

¹ Butari Sravani, ² R.Sravani

² MTech., Assistant Professor, CSE Dept.

^{1,2} P.V.K.K INSTITUTE OF TECHNOLOGY, Sanapa Road, Alamur(Post), Anantapuramu (Dist.), A.P.

(Affiliated to JNTUA, Anantapuramu)

ABSTRACT

Diabetes is the most common disease in developing countries, which makes early diagnosis and professional medical treatment necessary to reduce its effects. The effective method for diagnosis of diabetes is the evaluation of specific indications associated with the condition. However, the prevailing obstacle in automated diabetes detection is the existence of data, which can significantly affect the efficiency of machine learning models. The study uses methodologies with KNN-IMPuter and without Knn-Imputer to control the missing values in the diabetes data set. The proposed model strives to improve the accuracy of prediction by using the stacking classifier that integrates the predictions of the random forest bag as a primary and decision -making tree with GBM light as a secondary estimate. The methodologies are evaluated according to their effectiveness in controlling missing data and creating reliable forecasts. The finding shows that the stacking classifier overcomes alternative techniques for the accuracy and resistance of predictions and serves as an effective tool for automated identification of diabetes. This method ensures that absent data does not significantly prevent the model's effectiveness, which provides a significant answer to timely identification of diabetes.

Index Terms - Diabetes detection, ensemble learning, missing values, KNN Imputer, healthcare.

I. INTRODUCTION

Health care experts are necessary in identifying and treating a number of medical diseases, including illnesses, injuries and physical or mental disabilities. This group includes, inter alia, doctors, dentists, nurses, optometrists, physiotherapists and pharmacists. The health care system is necessary to maintain physical and mental health. Early identification of the disease is necessary, especially in disorders such as diabetes mellitus (DM), widely referred to as diabetes. Diabetes is a condition characterized by insufficient insulin production or ineffective insulin control. Insulin modulates blood glucose and uncontrolled diabetes results in hyperglycaemia, which can significantly damage organs and systems, especially neurological and vascular systems [1]. In 2014, 8.5% of the global population aged over 18 years of age fell on diabetes, and in 2012 it was 2.2 million deaths worldwide, which was distinguished into one to one 6 million in 2016 [1] [2] [3]. By using the 2019 charges of mortality, it exceeded 1.5 million, and the growing prevalence of patients with diabetes has become a significant economic burden, while global expenditures are approaching \$ 825 billion per year for diabetes management. The forecasts suggest that by 2045 the global population of individuals with diabetes can reach 629 million [7].

Diabetes mellitus is classified into 4 types: KIND 1 (diabetes dependent on juvenile or insulin), type 2 (diabetes non-involin), gestational diabetes (GD) and poor glucose control (before diabetes, type 4). Type 1 diabetes is defined by the inability of the body to synthesize insulin, requires exogenous administration of insulin. Diabetes

2. Gestational diabetes is manifested during pregnancy, characterized by elevated blood sugar levels in women who were previously not diabetic. Type 4 or pre-diabetes is defined by blood glucose levels that exceed normal ranges, but do not reach Prague for diagnosis of diabetes 2. Risk factors for diabetes mellitus include blood glucose levels, increased level head, increased triglycerides, insufficient physical activity, advanced age, hypertension, obesity, family predisposition and pregnancy [8].

The increasing prevalence of diabetes mellitus in both developing and evolved nations is associated with a sedentary lifestyle, inadequate dietary practices, and many socio-economic factors such as stress and insufficient healthcare awareness. To tackle this escalating concern, technologies including meal recommendation systems, activity monitors, medication alert systems, and interactive chatbots are employed to enhance the management and treatment of diabetes. Data mining and machine learning (ML) are vital instruments in healthcare, facilitating expedited and precise diagnostics. data utilized for diabetes prediction frequently includes partial or absent values, which can impair the efficacy of machine learning models, hence necessitating the resolution of these deficiencies to enhance diagnostic accuracy and reliability [9][10].

II. RELATED WORK

Diabetes Mellitus (DM) constitutes a significant global health issue, characterized by increasing prevalence rates and considerable implications for individuals and healthcare systems. Timely identification and efficient care are essential for alleviating its effects. Machine learning (ML) and data mining have become formidable instruments for automated diabetes prediction, improving precision and efficacy.

Numerous studies examined various machine learning techniques to predict diabetes. Deberneh and Kim [11] created a model of diabetes 2. A type using random vector machines (SVM), which has achieved increased accuracy. RuPapa et al. [12] Increased diagnosis of diabetes using Chi-cone and analysis of the main components (PCA) based on the selection of elements, which shows that the decrease in dimensions increases the efficiency of the model. Similarly, butt et al. [14] they examined decision -making trees, random forest and svm for classification of diabetes, which emphasized the importance of choosing an algorithm. Pethunachiyar [15] found that SVM based on cores that properly managed high -dimensional medical data files and exceeded other models in categorization of diabetes.

Deep learning methodologies have also made it easier to predict diabetes. Deg et al. [13] they used deep transmission and data enlargement to improve glucose levels in patients with 2 diabetes. Their methodology, which included pre -school models on extensive data sets followed by fine -tuning, increased accuracy and alleviated concerns about data imbalances. Madan et al. [17] Integrated Convolutional Neuron Network (CNN) with two-way long-term short-term (BI-LSTM) network that facilitates diabetes prediction in real time by analyzing spatial and time data formulas.

The learning of the file has shown effectiveness in predicting diabetes. Laila et al. [16] they used a file methodology that integrates decision -making trees and logistics regression, which significantly increases the identification of the risk of diabetes at an early stage. File approaches improve resistance and reduce prediction errors using many models. Optimization -based models have gained interest and represented hybrid architectures that include numerous neural networks to increase the accuracy of diabetes predictions.

Relieving data quality problems is essential for prediction of diabetes based on machine learning. Incomplete data and erroneous clinical records can prevent the model's accuracy. Imputations of KNN and various preliminary processing methods were used to solve incomplete data sets. The selection of functions and dimensions increases the efficiency of the model by removing foreign information. In addition, explaining AI (XAI) methodologies are examined to ensure the transparency of machine learning models and therefore cultivate confidence between doctors.

Regardless of these gains, problems persist, especially with data availability. For efficient predictive models, large, carefully annotated data sets with different patient features are essential. Scientists use generating synthetic data, data augmentation and transmission learning to resolve these restrictions. The ongoing breakthroughs in machine learning and deep learning offer the potential to increase diabetes prediction and facilitate the possibilities of early intervention, improving patient care.

III. MATERIALS AND METHODS

The recommended method seeks to create an advanced model of diabetes prediction using various algorithms and machine learning approaches. We will start a preliminary processing of a data set using KNN-IMPTEP [18] to solve missing data and compare the results with a version that excludes the imprint of Knn-Imputation. The data set will consist of authentic medical data, including variables such as patient demography, medical history and laboratory findings. We will implement many algorithms, including logistics regression, decision -making tree, random forest, stochastic gradient descent (SGD), Extratree, XGBOOST, Support Vector Machine (SVM) and Naive Bayes to evaluate their predictive efficiency. We will employ a voting classifier that integrates Extratree, XGBOOST and Random Forest, with a stacking classifier that combines bagging class with random forest as an estimate and decision tree with LightGBM as an estimate. The models will undergo the evaluation of the cross validation of K-time to guarantee robust overall performance and reliable predictions. The aim is to create a relatively accurate, scalable prediction device for diabetes.

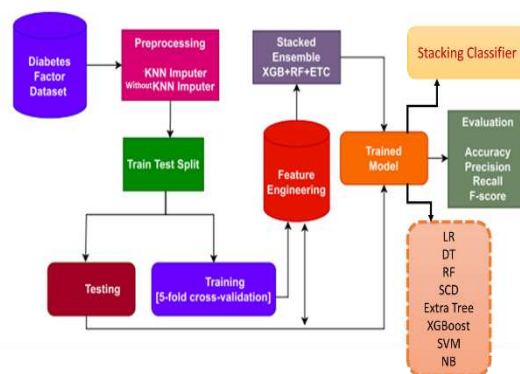


Fig.1 Proposed Architecture

The system design includes pre-work of diabetic factors data, divides them into training and test sets and implements engineering. It then uses the setting methodology of the file including several basic models (LR, DT [20], RF [21], etc.) to create a stacked model. The trained model is evaluated by means of measures such as accuracy, accuracy, download and f-score.

i) Dataset Collection:

Data set of diabetes includes basic clinical and physiological characteristics for predicting the advent of diabetes. The data set includes attributes including pregnancy, glucose, blood pressure, skin thickness, insulin, BMI, diabetes and age function, except for the target variable (0 indicating non -diabetic and 1 indicating diabetes). Data illustrate basic indicators affecting diagnosis of diabetes, facilitate comprehensive analysis and predictions of machine learning. This data set derived from healthcare records offers a reliable basis for evaluating various machine learning methods and increasing diagnostic accuracy of diagnostics of diabetes.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunc
0	6	148	72	35	0	33.6	0.6
1	1	85	66	29	0	26.6	0.1
2	8	183	64	0	0	23.3	0.6
3	1	89	66	23	94	28.1	0.1
4	0	137	40	35	168	43.1	2.2

Fig.2 Dataset Collection Table – Diabetes

ii) Pre-Processing:

Pre-processing converts raw data into a refined and organized format, guaranteeing quality for analysis. The process encompasses data cleansing, exploratory data analysis (EDA), and feature selection to enhance model efficacy.

a) Data Processing: data processing commences with the elimination of duplicate records to eradicate redundancy and uphold data integrity. Subsequently, irrelevant or null entries are eliminated during the cleaning phase to diminish noise and improve the dataset's quality. Categorical features, when present, are transformed into numerical representations by label encoding, facilitating compatibility with machine learning methods. Those methods mutually enhance the dataset, rendering it appropriate for efficient analysis and predictive modeling. Pristine, duplicate-free data constitutes the cornerstone for precise machine learning results.

b) EDA: EDA include comprehending data distributions, correlations, and trends. A correlation matrix is employed to discern interdependencies among features, hence emphasizing strongly associated attributes. Visualizations and sample outcome analysis give insights into the distribution of the goal variable (diabetic versus non-diabetic). Exploratory data analysis facilitates the identification of anomalies, absent values, and trends within the dataset, providing an in-depth comprehension of its structure. These insights are essential for informed decision-making in feature engineering and selection to achieve optimal model performance.

c) Feature Selection: feature selection determines the most pertinent features that contribute to the forecasting task. Methods such as correlation analysis, mutual information, or statistical tests eliminate redundant or less informative aspects. This diminishes dimensionality, improves model interpretability, and mitigates overfitting. The model emphasizes essential factors of diabetes by selecting significant features such as Glucose, BMI, and Insulin. Feature selection optimizes the dataset, enhancing computational efficiency and augmenting the accuracy and generalizability of machine learning models.

iii) Training & Testing:

The training phase entails equipping the model with a labeled dataset, enabling it to discern patterns and relationships within the data. During evaluation, the trained model is assessed on an independent dataset to determine its generalization capability and performance. The testing phase gives insights into the model's

predictive accuracy on novel data, thereby confirming its resilience and reliability. This approach is essential for assessing the model's efficacy in practical applications.

iv) Algorithms:

LR: Logistic Regression is employed to model the likelihood of diabetes based on the dataset's attributes. It employs a logistic function to predict the result, yielding a binary classification (diabetic or non-diabetic). Both WithKNN-Imputer and WithoutKNN-Imputer methodologies are employed to address missing values and enhance model accuracy through the imputation of absent data prior to training.

The simple linear regression line,

$$\hat{y} = a + bx \quad (1)$$

It can be read as follows: \hat{Y} represents the estimated value of Y , indicates the penetration indicating where the regression line intersects the y axes, and B predicts the change in y for each unit change in x .

DT: The decision Tree approach constructs a model by partitioning the data at each node according to the most informative attribute. It is utilized for the classification of diabetic cases, providing a comprehensible decision-making procedure. The WithKNN-Imputer and WithoutKNN-Imputer approaches are employed to address missing data, facilitating robust splits during tree construction, hence enhancing model performance [20].

Upon determining the cost of each occurrence and its associated probability, compute the anticipated fee of each outcome utilizing the subsequent formula:

Expected value (EV) = (first potential result x probability of result) + (2. Potential result x probability of result) - cost.

RF: A random forest is a set of files that use several decision -making trees for categorization purposes. Each tree increases decision -making on the final categorization and makes it more accurate and resistant. It is used to classify cases of diabetes and processes missing data efficiently using techniques without and without KNN-Imputer to improve the reliability of prediction and prevent excessive expulsion [21].

SGD: Stochastic Gradient Descent is a gradient-based optimization method employed to minimize the loss function and enhance model performance for extensive datasets. It is utilized for binary classification in diabetes detection. The WithKNN-Imputer and WithoutKNN-Imputer approaches address missing values to enhance the learning process and guarantee exact model training despite data deficiencies.

ExtraTree: ExtraTree is an ensemble approach that constructs several decision trees using random feature choice and partitioning. It is applied to enhance classification precision in diabetes prediction. Missing data is controlled using WithKNN-Imputer and WithoutKNN-Imputer approaches to improve the decision-making process, guaranteeing that missing input does not compromise model performance.

XGBoost: XGBoost is a gradient boosting approach that enhances classification performance by sequentially amalgamating weak learners into a more robust model. It is employed to identify diabetes through feature patterns. The WithKNN-Imputer and WithoutKNN-Imputer approaches are employed to address missing data, enhancing the stability and accuracy of the final boosted version.

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i) \quad (3)$$

Where \hat{y}_i denotes the final estimated value for the i th data point, K indicates the number of trees in the file and $f_k(x_i)$ indicates the prognosis of the tree for the i th data point.

SVM: The vector machine support constructs a hyperplane that optimizes the range between classes. It is used to categorize individuals as diabetic or non-diabetic. Within-Imputer and without KNN-Imputer, it is close to missing data, allowing the algorithm to create a significant boundary of decision-making despite the gaps of the data set, thus increasing the correctness of the model.

Equation for linear hyperplane can be written as:

$$W^T x + b = 0 \tag{4}$$

Where:

- W is a normal vector for hyperplan, which indicates the direction perpendicular to it.
- b is the term offset or distortion, which shows at the distance of hyperplan from origin along the normal vector W .

NB: Naive Bayes employs probabilistic methods to classify diabetes by determining the probability of each result based on the features. It is especially proficient in managing extensive datasets. The WithKNN-Imputer and WithoutKNN-Imputer approaches are utilized to address missing values, enabling the model to generate informed predictions despite partial data.

In generalized notation we write:

$$P(A, B|A) = P(A) * P(B|A) \tag{5}$$

The statement reads, “the probability of A and B given A is equal to the probability of A multiplied by the chance of B given A is known or has occurred.” This is referred to as conditional probability, or more accurately, joint conditional probability, as it quantifies the likelihood based on a preceding event or condition.

VC: The voting Classifier amalgamates the predictions of ExtraTree, XGBoost, and Random forest models, use a majority vote to classify diabetes. This ensemble technique enhances classification precision. Missing data is addressed with WithKNN-Imputer and WithoutKNN-Imputer approaches to guarantee that every version inside the ensemble possesses complete data, resulting in more dependable predictions [19].

SC: The Stacking Classifier employs an ensemble methodology that integrates BaggingClassifier with Random forest as an estimator, in conjunction with decision Tree and LightGBM for stacking functions. It utilizes many foundation learners to decorate performance in diabetes classification. The WithKNN-Imputer and WithoutKNN-Imputer strategies tackle missing data in all models within the ensemble, thereby enhancing the overall system's robustness and predictive accuracy.

IV. RESULTS AND DISCUSSION

Accuracy: The test accuracy is its ability to distinguish between patients and healthy cases. If you want to assess the test accuracy, calculate the ratio of real positives and real negatives in all evaluated cases. This can be mathematically expressed as:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{6}$$

Precision: The accuracy measures the percentage of positive cases or samples precisely classified. Accuracy is calculated using the formula:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{7}$$

Recall: The evaluation of machine learning evaluates the ability of the model to identify all appropriate cases of the class. It shows the efficiency of the model in the encapsulation of class instances by evaluating accurately predicted observations of the total number of positives.

$$"Recall = \frac{TP}{TP + FN} (8)"$$

F1-Score: The accuracy of the machine learning model is evaluated using the F1 score. Integration of the accuracy of the model and metrics of evocation. The metric of accuracy quantifies the frequency of the actual predictions made by the model throughout the data file.

$$"F1 Score = 2 * \frac{Recall * Precision}{Recall + Precision} * 100(9)"$$

Tables (1 and 2) evaluate power metrics-for each algorithm of performance metric F1, accuracy, induction and accuracy. The stacking classifier routinely overcomes all other algorithms across all metrics. The tables provide comparative exploration of metrics for alternative methods.

Table.1 Performance Evaluation Metrics of classification

ML Model	Accuracy	f1_score	Recall	Precision
LogisticRegression	0.773	0.775	0.773	0.780
DecisionTree	0.740	0.739	0.740	0.739
RandomForest	0.760	0.756	0.760	0.758
SGD	0.669	0.663	0.669	0.707
ExtraTree	0.662	0.657	0.662	0.658
XGBoost	0.708	0.704	0.708	0.704
SVM	0.649	0.781	0.649	0.994
Naive Bayes	0.727	0.724	0.727	0.724
Voting Classifier	0.727	0.723	0.727	0.725
Stacking Classifier	0.974	0.974	0.974	0.975

Graph.1 Comparison Graphs of Classification

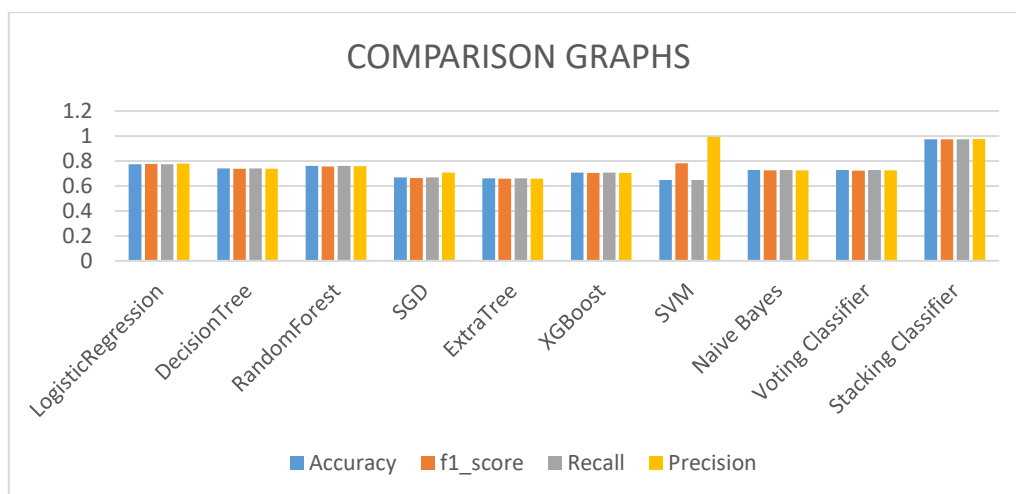
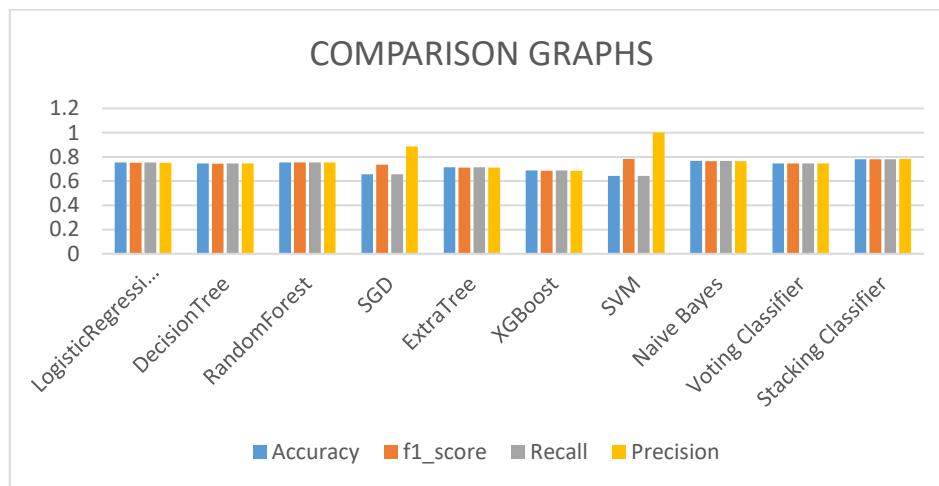


Table.2 Performance Evaluation Metrics for Without KNN Imputator

ML Model	Accuracy	f1_score	Recall	Precision
LogisticRegression	0.753	0.752	0.753	0.752
DecisionTree	0.747	0.743	0.747	0.745
RandomForest	0.753	0.753	0.753	0.753
SGD	0.656	0.736	0.656	0.885
ExtraTree	0.714	0.711	0.714	0.711
XGBoost	0.688	0.684	0.688	0.685
SVM	0.643	0.783	0.643	1.000
Naive Bayes	0.766	0.765	0.766	0.764
Voting Classifier	0.747	0.745	0.747	0.745
Stacking Classifier	0.779	0.781	0.779	0.784

Graph.2 Comparison Graphs for Without KNN Imputator



In Graphs (1 & 2), accuracy is depicted in blue, F1-score in orange, recall in grey, and precision in light yellow. Compared to the other models, the Stacking Classifier has greater performance in both methodologies, attaining the highest values. The graphs above visually represent these findings.

V. CONCLUSION

Recently, there has been a significant increase in the incidence of diabetes, affecting millions globally. Prompt therapies are essential for alleviating the complex problems linked to diabetes. Among the machine learning algorithms assessed for diabetes prediction, the Stacking Classifier, which integrates BaggingClassifier with Random forest as the estimator and decision Tree with LightGBM as the stacking estimator, emerged as the most effective model. This algorithm attained a remarkable accuracy of 97.4% with the WithKNN-Imputer method, showcasing its exceptional proficiency in managing missing data and providing very precise predictions. Conversely, the application of the WithoutKNN-Imputer approach resulted in a diminished accuracy of 77.9% for the model. The notable enhancement in performance tested by the WithKNN-Imputer technique highlights the critical role of imputing missing data in optimizing prediction accuracy. The Stacking Classifier's outstanding

performance underscores its capability to deliver dependable and efficient diabetes risk predictions, hence enhancing early detection and management of diabetes-related problems.

Future study intends to include deep learning models into diabetes prediction to get enhanced accuracy and resilience. Utilizing sophisticated neural network designs, the model is anticipated to manage larger and more intricate datasets efficiently, hence improving performance and adaptability. This approach promises to enhance prediction capabilities, tackling the complexities of high-dimensional data and fostering more sturdy and accurate diabetes detection systems in future applications.

REFERENCES

- [1] Diabetes Gojka. (Jul.2019). Diabetes: World Health Organization (WHO). Accessed: May 25, 2023.
- [2] S. El-Sappagh, F. Ali, S. El-Masri, K. Kim, A. Ali, and K.-S. Kwak, "Mobile health technologies for diabetes mellitus: Current state and future challenges," *IEEE Access*, vol. 7, pp. 21917–21947, 2019.
- [3] L. Mertz, "Automated insulin delivery: Taking the guesswork out of diabetes management," *IEEE Pulse*, vol. 9, no. 1, pp. 8–9, Jan. 2018.
- [4] H. A. Klein and A. R. Meininger, "Self management of medication and diabetes: Cognitive control," *IEEE Trans. Syst., Man, Cybern., A, Syst. Hum.*, vol. 34, no. 6, pp. 718–725, Nov. 2004. [5] WHO. (Apr. 2023). Diabetes: World Health Organization (WHO). Accessed: May 25, 2023.
- [6] A. A. Al Jarullah, "Decision tree discovery for the diagnosis of type II diabetes," in *Proc. Int. Conf. Innov. Inf. Technol.*, Apr. 2011, pp. 303–307.
- [7] G.D.Kalyankar, S.R.Poojara, and N.V.Dharwadkar, "Predictive analysis of diabetic patient data using machine learning and Hadoop," in *Proc. Int. Conf. I-SMAC*, Feb. 2017, pp. 619–624.
- [8] B.S.Ahamed, M.S.Arya, and A.O.V.Nancy, "Diabetes mellitus disease prediction using machine learning classifiers with oversampling and feature augmentation," *Adv. Hum.-Comput. Interact.*, vol. 2022, pp. 1–14, Sep. 2022.
- [9] S. Perveen, M. Shahbaz, A. Guergachi, and K. Keshavjee, "Performance analysis of data mining classification techniques to predict diabetes," *Proc. Comput. Sci.*, vol. 82, pp. 115–121, Jan. 2016.
- [10] I. Kavakiotis, O. Tsave, and A. Salifoglou, "Machine learning and data mining methods in diabetes research," *Comput. Struct. Biotechnol. J.*, vol. 15, no. 9, pp. 104–116, 2017.
- [11] H. M. Deberneh and I. Kim, "Prediction of type 2 diabetes based on machine learning algorithm," *Int. J. Environ. Res. Public Health*, vol. 18, no. 6, p. 3317, Mar. 2021.
- [12] V. Rupapara, F. Rustam, A. Ishaq, E. Lee, and I. Ashraf, "Chi-square and PCA based feature selection for diabetes detection with ensemble classifier," *Intell. Autom. Soft Comput.*, vol. 36, no. 2, pp. 1931–1949, 2023.
- [13] Y. Deng, L. Lu, L. Aponte, A. M. Angelidi, V. Novak, G. E. Karniadakis, and C. S. Mantzoros, "Deep transfer learning and data augmentation improve glucose levels prediction in type 2 diabetes patients," *NPJ Digit. Med.*, vol. 4, no. 1, p. 109, Jul. 2021.
- [14] U. M. Butt, S. Letchmunan, M. Ali, F. H. Hassan, A. Baqir, and H. H. R. Sherazi, "Machine learning based diabetes classification and prediction for healthcare applications," *J. Healthcare Eng.*, vol. 2021, pp. 1–17, Sep. 2021.

- [15] G. A. Pethunachiyar, “Classification of diabetes patients using kernel based support vector machines,” in Proc. Int. Conf. Comput. Commun. Informat. (ICCCI), Jan. 2020, pp. 1–4.
- [16] U. E. Laila, K. Mahboob, A. W. Khan, F. Khan, and W. Taekeun, “An ensemble approach to predict early-stage diabetes risk using machine learning: An empirical study,” *Sensors*, vol. 22, no. 14, p. 5247, Jul. 2022.
- [17] P. Madan, V. Singh, V. Chaudhari, Y. Albagory, A. Dumka, R. Singh, A. Gehlot, M. Rashid, S. S. Alshamrani, and A. S. AlGhamdi, “An optimization-based diabetes prediction model using CNN and bi directional LSTM in real-time environment,” *Appl. Sci.*, vol. 12, no. 8, p. 3989, Apr. 2022.
- [18] A. Juna, M. Umer, S. Sadiq, H. Karamti, A. A. Eshmawi, A. Mohamed, and I. Ashraf, “Water quality prediction using KNN imputer and multilayer perceptron,” *Water*, vol. 14, no. 17, p. 2592, Aug. 2022.
- [19] Y. Zhang, H. Zhang, J. Cai, and B. Yang, “A weighted voting classifier based on differential evolution,” *Abstract Appl. Anal.*, vol. 2014, pp. 1–6, Jan. 2014.
- [20] M. Brijain, R. Patel, M. R. Kushik, and K. Rana, “A survey on decision tree algorithm for classification,” *Int. J. Eng. Develop. Res.*, vol. 2, no. 1, pp. 1–5, 2014.
- [21] B. Gregorutti, B. Michel, and P. Saint-Pierre, “Correlation and variable importance in random forests,” *Statist. Comput.*, vol. 27, no. 3, pp. 659–678, May 2017.