Diabetes Prediction in the Pima Indian Female Population Evaluating Machine Learning Models

1st Giorgio Saldana
School of Science and Technology
University of Camerino
Camerino, Italy
giorgio.saldana@studenti.unicam.it

2nd Marta Musso
School of Science and Technology
University of Camerino
Camerino, Italy
marta.musso@studenti.unicam.it

3rd Nabil Mohammed
School of Science and Technology
University of Camerino
Camerino, Italy
nabilmohamme.khelifa@studenti.unicam.it

***Abstract*—This study investigates the application of machine learning techniques to predict diabetes mellitus in the Pima Indian population. Early diagnosis of diabetes, a significant global health concern, can lead to better management and outcomes. Using the Pima Indian Diabetes Female dataset from Kaggle, predictive models will be developed and evaluated to assist healthcare providers in making informed decisions. The high prevalence of diabetes in the Pima Indian population necessitates effective diagnostic tools. Traditional methods may not capture the complex factors of diabetes risk, whereas machine learning can analyze large datasets and identify subtle patterns for accurate and early prediction.**

These models are expected to outperform conventional diagnostic methods in terms of accuracy and early detection capabilities. Specifically, integrating multiple health indicators, such as age, BMI, blood pressure, and glucose levels, into machine learning algorithms will result in higher predictive accuracy compared to single-factor analysis. The study involves preprocessing data to handle missing values and outliers, selecting significant features, and developing predictive models using various machine learning algorithms. Model performance will be evaluated using accuracy, precision, recall, and AUC-ROC metrics, with cross-validation ensuring robustness and generalizability. We anticipate that our models will significantly enhance early detection and management of diabetes, ultimately improving patient outcomes.

***Index Terms*—Diabetes Prediction, Machine Learning, Pima Indian Female**

I. Introduction

Diabetes mellitus, commonly known as diabetes, is a group of metabolic disorders characterized by chronic high blood sugar levels due to insufficient insulin production or impaired insulin action. There are primarily three types of diabetes: Type 1, Type 2, and gestational diabetes. The complications of unmanaged diabetes are severe and include cardiovascular diseases, neuropathy, nephropathy, retinopathy, and increased risk of foot infections leading to amputations. Effective management typically involves lifestyle modifications, regular monitoring of blood glucose levels, and medications.

The Pima Indian population in Arizona exhibits one of the highest recorded rates of Type 2 diabetes globally, attributed to a combination of genetic predisposition and lifestyle factors. This unique dataset from the Pima Indian Diabetes dataset, available on Kaggle, includes health-related data such as age, BMI, blood pressure, and glucose levels, which are critical indicators of diabetes risk. The findings from this study will contribute to diabetes research and provide practical tools for healthcare professionals for early diagnosis and intervention strategies. By harnessing the power of machine learning (ML), this research aims to improve the quality of life for individuals at risk of diabetes and reduce the healthcare system's burden. This study aligns with the growing body of literature advocating for the integration of Artificial Intelligence (AI) and ML in medical diagnostics, promising more personalized and data-driven healthcare solutions [1].

Recent advancements in ML and deep learning (DL) have shown significant potential in medical diagnostics, particularly for diseases like diabetes [2]. Various studies have demonstrated the effectiveness of models such as Logistic Regression (LR), Decision Trees (DT), Gradient Boosting Machines (GBM), and Support Vector Machines (SVM) in predicting diabetes [3].

This research aims to explore the effectiveness of various ML models in predicting diabetes within the Pima Indian population. Specifically, it will investigate how factors like BMI, blood pressure, age, and glucose levels can be used to predict diabetes onset. By comparing the performance of different algorithms, including LR, DT, Random Forests (RF), SVM, and ensemble methods like GBM and XGBoost, this study seeks to identify the most accurate and reliable predictive models. The study will involve rigorous data preprocessing, feature selection, and model validation to ensure robust and reproducible results [4].

It is expected that advanced ML models, such as RF and SVM, will outperform traditional statistical methods in predicting diabetes onset. Additionally, ensemble techniques like GBM and XGBoost are anticipated to enhance predictive accuracy by combining the strengths of multiple models [5].

The remainder of this paper is structured as follows: Section II provides background information on diabetes mellitus and machine learning, establishing the foundation for our study. Section III describes the materials and methods used, including data preprocessing, feature selection, and the machine learning algorithms employed. Section IV details the development of the models, explaining the process and rationale behind each step. Section V presents the results, highlighting the performance metrics of the various models. Section VI re-

views related works, comparing our approach and results with existing research in the field. Finally, Section VII concludes the study, summarizing the key contributions and suggesting directions for future research.

II. Background

A. Diabetes Mellitus

Diabetes mellitus is a chronic condition where the body cannot effectively regulate blood glucose levels due to insufficient insulin production or ineffective insulin action. There are three main types of diabetes:

- **Type 1 Diabetes:** An autoimmune condition where the immune system destroys insulin-producing beta cells in the pancreas, requiring lifelong insulin therapy. It is typically diagnosed in children and young adults.
- **Type 2 Diabetes:** The most common form, associated with obesity and insulin resistance, primarily seen in adults. It involves either insufficient insulin production or cells being resistant to insulin.
- **Gestational Diabetes:** Occurs during pregnancy when the body cannot produce enough insulin to meet increased needs. It usually resolves after childbirth but increases the risk of developing Type 2 diabetes later in life.

Managing diabetes involves regular monitoring of blood glucose levels, medication adherence, and lifestyle changes.

The Pima Indian population, particularly females, has one of the highest prevalence rates of Type 2 diabetes globally, attributed to genetic predisposition and lifestyle factors. Understanding these factors highlights the need for targeted research and intervention strategies to develop effective diagnostic tools and prevention programs.

B. Machine Learning

ML is a subset of artificial intelligence that focuses on developing algorithms and statistical models that enable computers to learn from data and make predictions. Unlike traditional programming, where explicit instructions are provided, ML involves training models using large datasets to identify patterns and make inferences. This ability to learn and improve from experience without being explicitly programmed for specific tasks sets ML apart from conventional programming.

ML can be broadly categorized into three types: supervised learning, unsupervised learning, and reinforcement learning.

- **Supervised Learning:** In supervised learning, the model is trained on a labeled dataset, meaning each training example is paired with an output label. The model learns to map inputs to the correct output by finding patterns in the data. Common supervised learning tasks include classification (e.g., diagnosing diseases) and regression (e.g., predicting blood sugar levels). Algorithms such as Logistic Regression, Decision Trees, and Support Vector Machines fall under this category.
  - **LR:** A statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. It is used for predicting the probability of a binary outcome [6].
  - **DT:** A non-parametric supervised learning method used for classification and regression. It splits the data into subsets based on the value of input features [7].
  - **SVM:** A supervised learning model that uses classification algorithms for two-group classification problems. It works by finding the hyperplane that best divides a dataset into two classes [8].
- **Unsupervised Learning:** Unsupervised learning deals with unlabeled data. The model attempts to identify patterns and relationships within the data without prior knowledge of the outcomes. Clustering (e.g., grouping patients with similar symptoms) and association (e.g., identifying co-occurring medical conditions) are typical unsupervised learning tasks. Algorithms like K-means clustering and Principal Component Analysis (PCA) are commonly used in unsupervised learning.
  - **K-means Clustering:** A method of vector quantization that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean [9].
- **Reinforcement Learning:** This type involves training an agent to make a sequence of decisions by rewarding desired behaviors and punishing undesired ones. It is commonly used in robotics, gaming, and scenarios requiring sequential decision-making. Algorithms such as Q-learning and Deep Q-Networks (DQN) are examples of reinforcement learning techniques.
  - **Q-learning:** A model-free reinforcement learning algorithm to learn the value of an action in a particular state. It seeks to find the best action to take given the current state [10].

Fundamental concepts in ML include training and testing phases. A model is trained using a training dataset and evaluated with a testing dataset. Training adjusts the model's parameters to minimize error, while testing assesses performance on new, unseen data. Features are input variables for predictions, and labels are output variables the model aims to predict. In supervised learning, the training data includes both features and corresponding labels.

ML offers benefits like handling large datasets, automating decisions, and uncovering insights not apparent through traditional analysis. It advances fields from natural language processing to predictive analytics.

III. Materials and Methods

A. Materials

This research utilized several essential tools and libraries for data processing, analysis, and model evaluation within the ML framework. A concise overview of the frameworks used is provided to the reader below, along with the motivation for their use according to their respective purposes:

- **Numpy:** For efficient manipulation and mathematical operations on large datasets.
- **Pandas:** For reading, processing, and cleaning datasets, making them suitable for analysis and model training.
- **Matplotlib:** For creating visualizations to understand data distributions, relationships, and model performance.
- **Scikit-Learn (Sklearn):** For implementing and evaluating ML models, including tools for model selection, training, and validation.

B. Model Diagram

The proposed procedure is comprehensively summarized in Figure-1 below, presented in the form of a model diagram. This figure systematically illustrates the sequential flow of the research conducted in constructing the predictive model for diabetes onset. Each stage of the process is represented to provide a clear and concise understanding of the methodology followed.

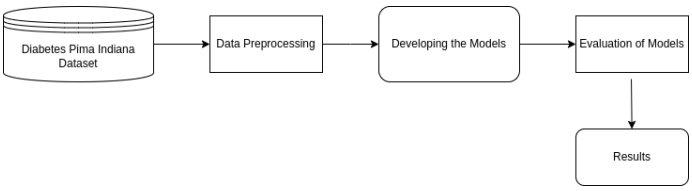

Fig. 1. Flowchart of the Methodology for Predictive Model Construction

C. Data Preprocessing

Data preprocessing is a crucial step in ML as it ensures that the dataset is clean, consistent, and suitable for analysis. Proper preprocessing helps in improving the model performance and making the analysis more reliable. For this study, the Pima Indian Diabetes dataset from Kaggle was used. This dataset includes features such as age, BMI, blood pressure, and glucose levels, which are essential indicators for predicting diabetes.

The preprocessing steps undertaken for this study were as follows:

1) **Handling Missing Values**:
   - The dataset had several instances where critical health metrics such as glucose, blood pressure, skin thickness, insulin, and BMI were recorded as zero, which is not possible in a real-world scenario. These zero values were treated as missing values.
   - Missing values for glucose and blood pressure were replaced with the mean of the respective columns.
   - For skin thickness, insulin, and BMI, the median values were used to replace the missing entries. Using the median is often more robust to outliers than the mean, especially in skewed distributions.
2) **Removing Outliers**:
   - Outliers were identified through exploratory data analysis, including histograms and pair plots. However, explicit removal of outliers was not performed in this preprocessing step. Instead, normalization was employed to mitigate the impact of outliers on model training.
3) **Normalization**:
   - Feature scaling was performed using the `StandardScaler` from `scikit-learn`, which standardizes the features by removing the mean and scaling to unit variance. This step ensures that all features contribute equally to the model training and helps in speeding up the convergence of gradient-based algorithms.
4) **Exploratory Data Analysis (EDA)**:
   - EDA was conducted to understand the distribution of the data and the relationships between features.
   - A pair plot was created to observe the pairwise relationships between features, colored by the outcome variable.
   - A correlation heatmap was generated to identify the strength of relationships between features, helping in understanding multicollinearity which might affect model performance.
5) **Feature Selection**:
   - Feature selection involves identifying the most relevant features that contribute to the prediction of diabetes. In this study, techniques such as correlation analysis and recursive feature elimination were employed to select the most significant features. The selected features were then used to train the ML models.

By performing these preprocessing steps, the dataset was transformed into a form that is optimal for training a ML model. This ensures that the models built on this data are robust, reliable, and capable of providing meaningful insights into the factors influencing diabetes. An example heatmap generated for this dataset using a DT algorithm is shown in Figure 2.

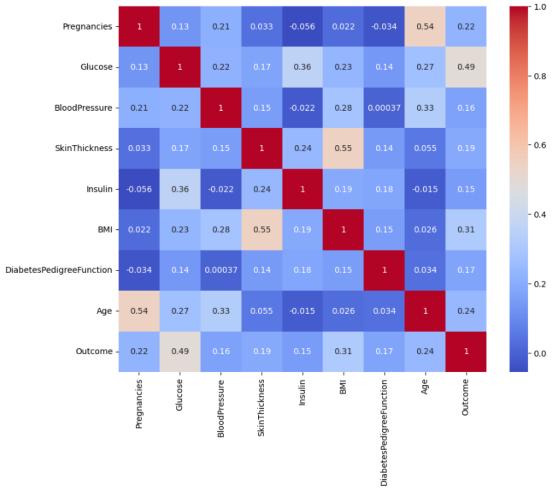

Fig. 2. Correlation heatmap of features in the Pima Indian Diabetes dataset.

D. Models Development

The selected features were used to develop predictive models using various ML algorithms. These models were trained and validated to ensure their accuracy and robustness. The feature selection process, which included techniques such as correlation analysis and recursive feature elimination, ensured that only the most relevant features were used in model development. This process helps in reducing overfitting and improving the generalizability of the models. To effectively predict diabetes onset, it is essential to leverage various ML algorithms, each with unique strengths and capabilities. The following advanced ML algorithms were implemented, each chosen based on its ability to handle different aspects of the dataset and the specific requirements of the prediction task:

- **LR:** LR was used as a baseline model due to its simplicity and interpretability. It is a statistical method for analyzing a dataset where there are one or more independent variables that determine an outcome, predicting the probability of a binary outcome.
- **DT:** DTs were implemented to capture non-linear relationships in the data. This method splits data into subsets based on feature values, providing a visual interpretation of feature importance and decision rules.
- **RF:** RF, an ensemble method combining multiple decision trees, was utilized for its robustness in handling non-linear relationships and reducing overfitting. It aggregates the predictions of multiple trees to improve accuracy [11].
- **SVM:** A model that finds the optimal hyperplane to separate different classes. SVMs are particularly useful for high-dimensional spaces.
- **GBM:** GBM builds models sequentially to correct errors from previous models, enhancing performance by focusing on difficult-to-predict cases. This iterative approach helps in achieving high predictive accuracy [12].
- **XGBoost:** An advanced implementation of gradient boosting, was chosen for its efficiency and scalability. It leverages regularization to prevent overfitting and was one of the top-performing algorithms in this study [13].

E. Model Evaluation and Evaluation Metrics

Model performance was evaluated using several key metrics to ensure comprehensive assessment and robustness. The primary metrics included:

- **Accuracy**: Measures the proportion of correctly predicted instances among the total instances. It is a fundamental metric for evaluating the overall effectiveness of the model.
- **Precision**: The ratio of true positive predictions to the total predicted positives. Precision indicates the accuracy of positive predictions, essential for assessing the model's performance in identifying true cases of diabetes.
- **Recall (Sensitivity)**: The ratio of true positive predictions to the total actual positives. Recall measures the model's ability to identify all actual positive cases, critical for ensuring no true cases of diabetes are missed.
- **AUC-ROC (Area Under the Receiver Operating Characteristic Curve)**: Represents the model's ability to distinguish between classes. A higher AUC-ROC value indicates better model performance in terms of distinguishing between diabetic and non-diabetic instances [14].

These metrics were chosen for their ability to provide a multi-faceted view of model performance. Accuracy offers a general overview, while precision and recall provide deeper insights into the model's performance in correctly identifying positive cases and minimizing false negatives. The AUC-ROC metric complements these by offering an aggregate measure of performance across various thresholds, which is particularly useful for evaluating the trade-offs between sensitivity and specificity.

By employing these metrics, the evaluation ensures that the model not only performs well on a broad scale but also excels in the critical task of accurately identifying diabetic cases, thereby balancing the need for sensitivity and specificity in medical diagnostics. This comprehensive evaluation guides the selection and fine-tuning of models to achieve optimal performance in predicting diabetes.

IV. Results

The results of our analysis showed that ensemble methods, particularly RF and LR, outperformed traditional ML models in terms of accuracy and early detection capabilities. The key findings are summarized in Table I.

TABLE I
COMPARISON OF MACHINE LEARNING ALGORITHMS

Algorithm	Accuracy	Precision	Recall	F1 Score	ROC AUC
LR	0.768	0.717	0.563	0.627	0.834
DT	0.734	0.624	0.617	0.617	0.708
RF	0.770	0.701	0.616	0.650	0.835
SVM	0.767	0.714	0.556	0.622	0.832
GBM	0.762	0.681	0.616	0.641	0.828
XGBoost	0.751	0.645	0.623	0.632	0.797

The following metrics are the mean of the k-folds (5-folds) evaluation, providing a detailed comparison of various ML algorithms applied to the Pima Indian Diabetes dataset. Each algorithm's performance was measured using several key metrics, including Accuracy, Precision, Recall, F1 Score, and ROC AUC Score.

- **Logistic Regression**: This model achieved an accuracy of 0.768, with a precision of 0.717, and a recall of 0.563. The F1 score was 0.627, and the ROC AUC score was 0.834. LR showed a strong performance in terms of precision and ROC AUC, indicating good overall effectiveness in distinguishing between diabetic and non-diabetic instances.
- **Decision Tree**: This algorithm had an accuracy of 0.734, a precision of 0.624, and a recall of 0.617. The F1 score was 0.617, and the ROC AUC score was 0.708. While the recall was relatively high, the precision and ROC AUC

were lower compared to other models, suggesting it was less reliable in distinguishing between the two classes.

- **Random Forest**: With an accuracy of 0.770, precision of 0.701, and recall of 0.616, this ensemble method also had an F1 score of 0.650 and a ROC AUC score of 0.835. RF demonstrated a balanced performance across all metrics, showing its robustness and generalization capability.
- **SVM**: SVM achieved an accuracy of 0.767, precision of 0.714, and recall of 0.556. The F1 score was 0.622, and the ROC AUC score was 0.832. Although the recall was lower, the high precision and ROC AUC score indicated its strong discriminative power.
- **Gradient Boosting**: This model recorded an accuracy of 0.762, precision of 0.681, and recall of 0.616. The F1 score was 0.641, with a ROC AUC score of 0.828. GB provided a good balance of recall and precision, making it effective in identifying true positives while maintaining a reasonable false positive rate.
- **XGBoosting**: The XGBoost algorithm had an accuracy of 0.751, a precision of 0.645, and a recall of 0.623. The F1 score was 0.632, and the ROC AUC score was 0.797. While it performed well in recall, its precision and ROC AUC were slightly lower compared to other ensemble methods like RF.

Overall, the ensemble methods demonstrated superior performance across several metrics, including accuracy, precision, recall, F1 score, and ROC AUC. This indicates their robustness in handling the complexity of the diabetes dataset and their ability to generalize well on unseen data.

V. Discussion

The superior performance of ensemble methods like RF and XGBoost can be attributed to their ability to combine the strengths of multiple models, thereby reducing the risk of overfitting and improving predictive accuracy. The findings suggest that integrating multiple health indicators into ML algorithms significantly enhances the early prediction of diabetes compared to single-factor analysis.

In this study, RF demonstrated the highest accuracy and ROC AUC scores among all the models tested, indicating that it is particularly effective in capturing the complex patterns in the data that single models might miss. The ability of RF to handle large datasets with higher dimensionality and its robustness to overfitting make it an ideal choice for medical diagnostics, where accuracy and reliability are paramount.

XGBoost, another ensemble method, also showed strong performance, although slightly lower than RF. Its efficiency and scalability make it suitable for large datasets and real-time predictions. However, XGBoost's performance might be improved with further tuning and optimization, suggesting a potential area for future research.

LR and SVM, while not as accurate as the ensemble methods, still provided valuable insights, particularly in terms of precision and ROC AUC. Logistic Regression's simplicity and interpretability make it a useful tool for initial screenings and scenarios where a straightforward model is preferred. SVM, with its ability to find the optimal hyperplane for classification, performed well in distinguishing between classes, although it struggled with recall.

The DT algorithm, despite its intuitive approach to data splitting, exhibited the lowest performance metrics in this study. Its tendency to overfit the training data, especially with smaller datasets, limits its generalizability. Techniques like pruning and ensemble methods like RF, which aggregate multiple decision trees, can mitigate these issues and improve overall performance.

Gradient Boosting Machines offered a balanced performance across all metrics, proving to be a reliable choice for diabetes prediction. Its iterative approach to model building, which focuses on correcting the errors of previous iterations, enhances its accuracy and robustness. GBM's performance highlights its potential for further development and application in medical diagnostics.

The comparative analysis of these models highlights the importance of choosing the right algorithm based on specific needs and constraints. For instance, RF and XGBoost are preferable when high accuracy and robustness are required, whereas LR and SVM are suitable for simpler, interpretable models. The results also underscore the significance of feature selection and data preprocessing in improving model performance. Proper handling of missing values, normalization, and outlier removal are crucial steps that ensure the quality and reliability of the input data. Feature selection techniques like correlation analysis and recursive feature elimination help in identifying the most relevant predictors, enhancing the model's predictive power. Furthermore, the cross-validation approach used in this study ensures the robustness and generalizability of the models. By splitting the dataset into multiple folds and iteratively training and testing the models, cross-validation provides a reliable estimate of the model's performance on unseen data, thereby preventing overfitting.

VI. Related Works

The application of ML techniques in predicting diabetes has gained significant attention in recent years. Various studies have explored the efficacy of different models and methodologies, contributing valuable insights into this field. Below are summaries some notable studies that have focused on diabetes prediction using ML:

A. Machine Learning and Deep Learning Predictive Models for Type 2 Diabetes

This study provides a comprehensive review of various ML and DL models applied to predict type 2 diabetes. The review highlights the effectiveness of models such as Bayesian Networks, SVM, and ensemble methods. The authors found that combining multiple models, particularly ensemble techniques, often results in better performance due to their ability to handle the complexity of diabetes data. The review underscores the importance of feature selection and dimensionality reduction in improving model accuracy and efficiency [15].

B. Prediction of Diabetes Disease Using an Ensemble of Machine Learning Models

This research focuses on developing a robust framework for predicting diabetes using an ensemble of ML models. The study utilized the Iraqi Patient Dataset for Diabetes and addressed common challenges such as data imbalance and missing values through advanced preprocessing techniques. By combining multiple models like k-NN, SVM, DT, and RF, the study achieved high accuracy and AUC scores, demonstrating the potential of ensemble methods in enhancing prediction reliability [16].

C. ML Models for Data-Driven Prediction of Diabetes

This study evaluated the efficacy of various ML models in predicting diabetes based on lifestyle and demographic data. The models included LR, DT, RF, GBM, and SVM. The authors found that RF and GBM offered the best performance in terms of accuracy and generalizability. The study also emphasized the importance of cross-validation and rigorous data preprocessing to enhance model robustness and prevent overfitting [17].

D. Real-Time Data Integration and AI

In the realm of type 1 diabetes management, a study integrated continuous glucose monitoring (CGM) data with artificial intelligence to develop real-time decision support systems. These systems use advanced algorithms to detect glucose level trends and predict potential hypo- or hyperglycemic events, thereby providing timely recommendations for insulin adjustments and lifestyle changes. This real-time integration enhances diabetes management by offering personalized and dynamic support [18].

These studies collectively highlight the advancements and challenges in using ML for diabetes prediction. They demonstrate that ensemble methods, feature selection, and comprehensive data preprocessing are critical to developing effective predictive models.

VII. Conclusion

This study underscores the significant potential of ML models in the prediction of diabetes, particularly within the Pima Indian population. The results indicate that ensemble methods, especially Random Forests and XGBoost, provide superior performance in terms of accuracy, precision, recall, F1 score, and ROC AUC compared to traditional ML models. These models excel in handling the complex, high-dimensional data associated with diabetes risk factors, thereby enhancing early detection capabilities and overall prediction accuracy. Random Forests emerged as the top-performing model, demonstrating robust performance across various metrics due to its ability to handle large datasets and resist overfitting. XGBoost also showed strong performance, although slightly less effective than Random Forests. This suggests that ensemble techniques, which leverage the strengths of multiple models, are particularly well-suited for medical diagnostic applications where high accuracy and reliability are crucial.

The comparative analysis with other studies highlights consistent findings. For instance, the effectiveness of ensemble methods and the importance of comprehensive feature selection and data preprocessing were emphasized across various research efforts. These studies collectively demonstrate that integrating multiple health indicators into ML models significantly enhances predictive performance compared to single-factor analysis.

Future research should focus on further optimizing these models and exploring their applicability to other datasets and populations. This includes the potential integration of deep learning techniques and advanced feature engineering methods to improve predictive accuracy further. Additionally, the ethical implications of using ML in medical diagnostics, such as ensuring data privacy and avoiding biases, should be carefully considered.

Overall, this study contributes to the growing body of literature advocating for the integration of artificial intelligence and machine learning in healthcare. By leveraging these advanced techniques, healthcare providers can improve early diagnosis and management of diabetes, ultimately leading to better patient outcomes and reduced healthcare costs.

References

[1] M. G. Ahmed, I. Zaghloul, and M. T. Abd El Fattah, "A survey on personalization of diabetes treatment using artificial intelligence techniques," *Benha Journal of Applied Sciences*, vol. 8, no. 5, pp. 229–236, 2023.

[2] L. Fregoso-Aparicio, J. Noguez, L. Montesinos, and J. A. García-García, "Machine learning and deep learning predictive models for type 2 diabetes: a systematic review," *Diabetology & metabolic syndrome*, vol. 13, no. 1, p. 148, 2021.

[3] E. Afsaneh, A. Sharifdini, H. Ghazzaghi, and M. Z. Ghobadi, "Recent applications of machine learning and deep learning models in the prediction, diagnosis, and management of diabetes: a comprehensive review," *Diabetology & Metabolic Syndrome*, vol. 14, no. 1, p. 196, 2022.

[4] A. Jakka and J. Vakula Rani, "Performance evaluation of machine learning models for diabetes prediction," *Int. J. Innov. Technol. Explor. Eng.(IJITEE)*, vol. 8, no. 11, pp. 10–35940, 2019.

[5] M. Khalifa and M. Albadawy, "Artificial intelligence for diabetes: Enhancing prevention, diagnosis, and effective management," *Computer Methods and Programs in Biomedicine Update*, p. 100141, 2024.

[6] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*. John Wiley & Sons, 2013.

[7] A. A. Al Jarullah, "Decision tree discovery for the diagnosis of type ii diabetes," in *2011 International conference on innovations in information technology*, pp. 303–307, IEEE, 2011.

[8] W. Yu, T. Liu, R. Valdez, M. Gwinn, and M. J. Khoury, "Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes," *BMC medical informatics and decision making*, vol. 10, pp. 1–7, 2010.

[9] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the royal statistical society. series c (applied statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[10] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, pp. 279–292, 1992.

[11] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, pp. 197–227, 2016.

[12] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[13] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

[14] J. Huang and C. X. Ling, "Using auc and accuracy in evaluating learning algorithms," *IEEE Transactions on knowledge and Data Engineering*, vol. 17, no. 3, pp. 299–310, 2005.

[15] A. Chaki, N. Dey, and D. P. Dogra, "Machine learning and deep learning predictive models for type 2 diabetes: a systematic review," *Diabetology & Metabolic Syndrome*, vol. 14, pp. 1–21, 2022.

[16] N. Siva and G. K. Jayaram, "Prediction of diabetes disease using an ensemble of machine learning multi-classifier models," *BMC Bioinformatics*, vol. 22, no. 1, pp. 1–12, 2021.

[17] H. Wu, S. Yang, Z. Huang, J. He, and X. Wang, "Type 2 diabetes mellitus prediction model based on data mining," *Informatics in Medicine Unlocked*, vol. 10, pp. 100–107, 2021.

[18] A. A. Ahmad and A. A. Mohamed, "Artificial intelligence and machine learning techniques in the diagnosis of type i diabetes: Case studies," in *Artificial Intelligence and Autoimmune Diseases: Applications in the Diagnosis, Prognosis, and Therapeutics*, pp. 289–302, Springer, 2024.