ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 11 Issue 6 June 2024

Comprehensive Diabetes Prediction: Integrating Early Stage and Overall Risk Assessment Using Ada Boost, Random Forest, and Gradient Boosting Models

[1] Keshav Agarwal, [2] Dr. Abhijit Saha

[1] [2] Dept. of Computing Technologies, SRM Institute of Science and Technology, KTR, Chennai
Corresponding Author Email: [1] ka7960@srmist.edu.in, [2] abhijits1@srmist.edu.in

Abstract*— The Comprehensive Diabetes Prediction system is a pioneering advancement in healthcare analytics, designed to integrate early-stage and overall risk assessment for diabetes through the utilization of AdaBoost, Random Forest, and Gradient Boosting models. The goal of this study is to improve the accuracy and timeliness of diabetes prediction by leveraging the amount of information included in several datasets that include patient demographics, medical history, lifestyle factors, clinical measurements, and blood sugar levels. Using painstaking data pre-processing, feature selection, and model generation procedures, the system achieves promising prediction performance by employing ensemble learning methodologies and optimisation tactics. The models' robustness is assessed using evaluation metrics like as accuracy, precision, recall, and F1-score. The implementation of the Comprehensive Diabetes Prediction system marks a shift in proactive interventions and personalised treatment techniques, with the potential to considerably enhance diabetes management outcomes. Future endeavours will focus on further refining predictive models, exploring additional data modalities, and validating the system in real-world clinical settings. Through sustained innovation and collaboration with healthcare professionals, this research endeavours to unlock the full potential of predictive analytics in advancing the field of diabetes care, ultimately leading to more effective disease management and improved patient outcomes.

***Index Terms*—chronic disease, diabetes prediction, ensemble learning, machine learning.**

I. INTRODUCTION

Diabetes, a global health concern, necessitates precise prediction methodologies. Machine learning (ML) is a promising technology that uses various data to forecast diabetes onset. Ensemble learning, notably Random Forest and Gradient Boosting, enhances predictive accuracy. Our research integrates these techniques with Artificial Bee Colony optimization to optimize models further. By [2] combining early stage and overall risk assessment, our approach improves diabetes detection. We discuss the rationale, methodology, and experimental results, highlighting its efficacy in enhancing disease management and patient outcomes.

A. Problem Statement:

The global surge in diabetes cases highlights the urgency for precise prediction methods, essential for proactive interventions and personalized treatment. Despite abundant data, accurately identifying individuals at risk, especially in early stages, remains challenging. Current models often lack accuracy, interpretability, and struggle with complex data. Our study, "Comprehensive Diabetes Prediction," addresses these challenges by integrating AdaBoost, Random Forest, and Gradient Boosting techniques. This holistic approach aims to improve predictive performance, providing reliable tools for healthcare practitioners. By advancing predictive analytics, our research contributes to early diabetes detection and management, ultimately [5] enhancing patient outcomes and alleviating healthcare system burdens.

B. Objective:

The diabetes epidemic demands precise prediction methods for early interventions. Current models often lack accuracy and interpretability [10], hampering effective risk identification. Our approach integrates advanced machine learning to improve diabetes management globally.

Objectives are listed below-

a. Investigate the effectiveness of ensemble learning algorithms, including Random Forest and AdaBoost in early-stage diabetes prediction.
b. Use Gradient Boosting to improve machine learning performance by optimizing hyperparameters and feature selection.
c. Evaluate the developed models using diverse datasets containing demographic [6], clinical, and lifestyle variables to ensure robustness and generalizability.

C. Research Methodology:

The experiment's primary purpose is to discriminate between the various outputs generated. The framework utilised is a deep learning classifier: [14]

Stage 1: Examine the literature where this research will be

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 11 Issue 6 June 2024

conducted, namely, deep learning and Speech Emotion Recognition Systems.

Stage 2: Then, using the literature review, identify the problem.

Stage 3: Then, we will proclaim our work's goal.

Stage 4: Acquire the data from open source.

Stage 5: The system's design and evolution under it. The suggested approach will then be examined to see if it generates the desired results.

Stage 6: Calculate the accuracy and compare it to other systems to enhance or develop the plan in the future

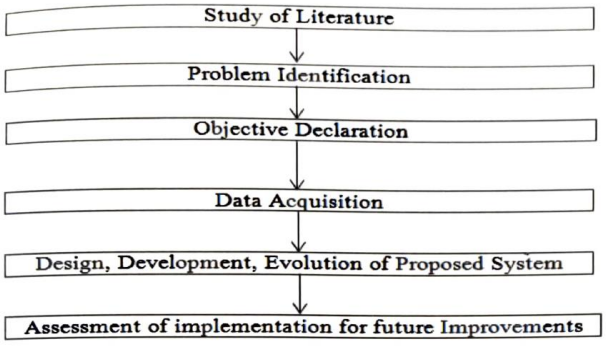

Figure 1. Research Methodology

D. Motivation:

The motivation behind our research stems from the urgent need to address the escalating global burden of diabetes. With diabetes affecting millions worldwide and its prevalence projected to rise, there is a critical demand for effective predictive tools that can enable timely interventions and personalized treatment strategies. Current predictive models often fall short in accurately identifying individuals at risk[15], particularly in the early stages of the disease, leading to missed opportunities for intervention and suboptimal outcomes. By leveraging advanced machine learning techniques, ensemble learning methods, and optimization strategies, we aim to develop robust and interpretable models for comprehensive diabetes prediction. Our goal is to provide healthcare practitioners with reliable tools that can accurately assess an individual's risk of diabetes onset [4], paving the way for proactive management and improved patient outcomes. Through our research, we aspire to add to the global initiatives to fight diabetes and enhance public health outcomes broadly.

E. Research gaps:

a. The incorporation of many data sources into diabetes prediction models is an area of unmet research need. A large number of existing models primarily rely on traditional risk variables such as clinical and demographic data, even though multiple studies emphasise the significance of including new data types such as genetics, lifestyle, and environmental factors. Therefore, there is an urgent need to create models that efficiently make use of a wider variety of data sources in order to improve prediction accuracy and comprehensiveness.

b. Regarding the validity of evaluation methods, there is a substantial discrepancy in the context of diabetes prediction model validation. Extensive validation, especially external validation on distinct datasets, is frequently absent from published models. This restriction makes it more difficult for these models to be applied and generalized in practical contexts. Therefore, stronger validation techniques are desperately needed to guarantee the accuracy and consistency of diabetes prediction algorithms.

c. The interpretability of intricate diabetes prediction models is lacking. While models that are transparent, such as those that use the SHAP framework, can boost trust, many of the models in use today are opaque, which makes it more difficult for them to be easily incorporated into clinical decision-making.

d. The scalability and accessibility of diabetes prediction technology are areas lacking in research. Healthcare adoption depends on characteristics like user-friendly interfaces and connection with electronic health records, but many of the systems in use today lack them, which restricts their general use.

F. Existing Base Paper:

The base paper used is "Diabetes Prediction using Machine Learning Algorithms with Feature Selection and Dimensionality Reduction". (https://ieeexplore.ieee.org/document/9441935)

G. Drawbacks of the Existing Base Paper:

a. The contextual knowledge of the problem is limited because the base study does not specifically address the research gaps in the diabetes prediction domain.

b. The primary emphasis of the base work is on the technical use of dimensionality reduction and feature selection approaches, with little attention paid to addressing the broader issues surrounding diabetes prediction.

c. A thorough explanation of how to combine various data sources and validate the models for broader applicability is absent from the main study.

d. The base study does not fully address the models' interpretability or the possibility of scalable, user-friendly installations.

H. Novelty:

a. A more comprehensive approach to risk assessment is provided by the integration of many ensemble learning algorithms, such as AdaBoost, Random Forest, and Gradient Boosting, to capture both early-stage and total diabetes risk variables.

b. The focus is on filling in the research gaps that have

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 11 Issue 6 June 2024

been found by utilizing a variety of data sources, strengthening validation techniques, expanding model interpretability, and making sure the system is scalable and accessible for practical use.

c. Comprehensive focus on not only predicting the onset of diabetes but also providing insights for timely interventions, contributing to enhanced diabetes management and better patient outcomes.

I. Rationale for Chosen Models:

a. By merging weak learners into a powerful classifier, AdaBoost is a resilient and adaptable ensemble learning algorithm that has proven successful in binary classification problems, such as early-stage diabetes prediction.

b. Random Forest, an ensemble of decision trees, is well-known for its capacity to handle complicated, non-linear relationships in data, making it ideal for capturing the nuanced patterns associated with diabetes risk factors.

c. Especially in the later stages of the disease, Gradient Boosting is a potent ensemble strategy that can capture subtle correlations between features and targets, hence enhancing the overall forecast accuracy for diabetes.

d. With each of these three learning models having its own advantages, it is anticipated that the combination of them will offer a complete and precise diabetes prediction system, overcoming the shortcomings of the individual algorithms and improving the overall efficiency.

II. DOMAIN

A. Data Science:

The cutting edge of contemporary analytics is represented by data science, an enthralling fusion of art and science that is transforming companies through skillful data manipulation to glean actionable insights and facilitate well-informed decision-making. At its core, data science marries the disciplines of mathematics, statistics, computer science, and domain expertise, synergistically harnessing the power of data to unlock hidden patterns, unearth valuable correlations, and predict future trends with unprecedented accuracy. Through the lens of data science, the vast sea of raw data becomes a canvas upon which algorithms paint intricate portraits of reality, illuminating the obscured pathways to innovation and discovery.

B. Sub-Domain:

a) Artificial Inttelligience (Ai):

Artificial intelligence (AI) is a branch of computer science that focuses on developing intelligent systems capable of replicating human cognitive abilities such as learning, problem solving, and taking decisions. Machine learning, natural language processing, and computer vision are key components of AI, which enable activities like data analysis, language comprehension, and image recognition. AI has applications in healthcare, finance, transportation, and other fields, including medical diagnosis, fraud detection, and self-driving vehicles. Recent advancements in deep learning have propelled AI to new heights, enabling breakthroughs in areas like image and speech recognition.

b) Machine Learning (Ml):

A subset of artificial intelligence (AI), machine learning (ML) enables computers to learn from data and forecast without the need for explicit programming. ML encompasses supervised learning, where algorithms learn from labelled data for tasks like classification and regression, unsupervised learning for discovering hidden patterns, and reinforcement learning for optimizing strategies through interaction with an environment. ML finds applications across healthcare, finance, e-commerce, and cybersecurity, driving advancements such as deep learning for tasks like image and speech recognition. Despite its transformative potential, ML raises ethical concerns regarding bias and privacy, emphasizing the importance of responsible development and deployment to harness its benefits for society.

C. Limitations:

a) **Data Quality**: Model effectiveness hinges on high-quality, representative data. Incomplete, noisy, or biased datasets can lead to inaccurate predictions.

b) **Model Interpretability**: Ensemble methods like Random Forest and Gradient Boosting offer high accuracy but may lack interpretability compared to simpler models. Understanding feature contributions may be challenging, impacting trust.

c) **Overfitting**: Ensemble models, especially Gradient Boosting, are prone to overfitting, memorizing training data instead of learning patterns, leading to unreliable predictions.

d) **Computational Complexity**: Training and tuning ensemble models require significant computational resources, especially with large datasets and hyperparameter optimization, posing challenges in real-time or resource-constrained settings.

e) **Integration Challenges**: Integrating multiple models into a cohesive system may pose technical hurdles, requiring effort to ensure seamless communication and maintenance.

f) **Limited Scope**: While our approach focuses on diabetes prediction, it may not address all aspects of management and prevention, such as lifestyle interventions or genetic factors.

D. Workflow:

a) **Data Acquisition and Preprocessing**: Gather diverse datasets on diabetes risk factors, addressing anomalies like missing values and outliers. Standardize attributes and encode variables for consistency.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 11 Issue 6 June 2024

b) **Feature Engineering and Selection:** Extract informative attributes and derive new variables using techniques like correlation analysis and dimensionality reduction.

c) **Model Development:** To identify diabetes risk variables, use machine learning models such as AdaBoost, Random Forest (RF), & Gradient Boosting (GB).

d) **Ensemble Model Integration**: Combine predictions from AB, RF, and GB models for improved accuracy using techniques like model averaging or stacking.

e) **User Input and Interaction:** Design a user-friendly interface for data input and interaction, allowing users to provide feedback.

f) **Model Assessing**: Make that the model is robust through cross-validation by evaluating its performance using measures like accuracy, precision, and recalls.

g) **Final Answer Delivery**: Present predictions clearly, highlighting key risk factors and providing actionable insights for disease management.

h) **Continuous Monitoring and Improvement**: Implement mechanisms for ongoing model performance monitoring and regular updates based on new data and user feedback.

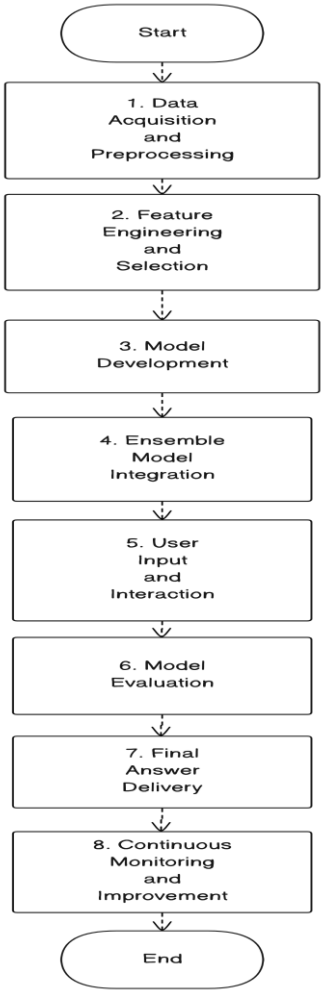

Figure 2. Workflow for Comprehensive Diabetes Prediction

III. LITERATURE REVIEW

A. Diabetes Prediction using Machine Learning Algorithms (2019)

a. Authored by Aishwarya Mujumdar and Dr. Vaidehi.
b. Proposal of a diabetes prediction model incorporating external factors along with traditional indicators like Glucose and BMI.
c. Utilization of Big Data Analytics to study large datasets and predict outcomes.
d. Emphasis on the significance of accurate classification for appropriate treatment.
e. Implementation of ensemble learning techniques like Random Forest and Gradient Boosting for enhanced prediction.

B. A model for early prediction of diabetes (2019)

a. Authored by Muhammad Atif Iqbala and Talha Mahboob Alama.
b. Prediction of diabetes using significant attributes and characterization of their relationships.
c. Selection of attributes via principal component analysis and Apriori method.
d. Using K-means clustering, Random Forest (RF), and Artificial Neural Networks (ANN) for prediction.
e. Identification of associations between Type 2 diabetes and body mass index (BMI) and glucose level through feature engineering.

C. Diabetes Prediction Using Machine Learning (2020)

a. Authored by KM Jyoti Rani.
b. Exploration of the chronic nature and global impact of diabetes mellitus.
c. An overview of machine learning methods for diabetes early diagnosis.
d. Description of how diabetes disrupts the body's glucose regulation system.
e. Discussion on the significance of model interpretability and feature importance analysis for understanding prediction outcomes.

D. Application of Artificial Intelligence in Diabetes Education and Management (2020)

a. Authored by Juan Li and Jin Huang.
b. Review of AI techniques for diabetes education and management, including Natural Language Processing (NLP) for patient education materials.
c. Put an emphasis on interventions for lifetime education and individualized patient management.
d. Using AI to turn genomic and medical records into relevant insights.
e. AI decision support tools for diabetes management can use Reinforcement Learning to provide personalized treatment strategies.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 11 Issue 6 June 2024

E. Research on Diabetes Prediction Method Based on Machine Learning (2020)

a. Authored by Jingyu Xue.
b. Introduction to diabetes mellitus and its clinical forms (Type 1 and Type 2).
c. Use supervised machine learning methods such as SVM and naive Bayes classification algorithms for prediction.
d. Training based on data from diabetic and probable diabetic individuals ranging 16–90, with feature extraction using techniques like Principal Component Analysis (PCA).
e. Comparison of classification and recognition accuracy, highlighting SVM's superior performance in handling non-linear data distributions.

IV. PROPOSED SYSTEM

The machine learning-based system proposed for early-stage diabetes prediction follows a systematic approach. Comprehensive data collection covers various risk factors, from demographics to medical history and clinical measures. Pre-processing ensures data quality, handling missing values and outliers. Feature selection enhances model efficiency and interpretability. Model development involves selecting an appropriate classification algorithm, training it rigorously, and optimizing its performance through hyperparameter tuning and cross-validation. Evaluation on a separate testing dataset includes various performance metrics [3], with interpretability techniques providing insights into the model's decisions. Deployment in a user-friendly interface ensures accessibility, with continuous monitoring for performance tracking and updates based on new data. By facilitating early identification of diabetes risk, the system aims to enable timely interventions and improve health outcomes.

A. Proposed Model Architecture:

The model architecture comprises interconnected components for early-stage diabetes prediction, data preprocessing, feature extraction, model training, and evaluation. Comprehensive data collection precedes pre-processing, which standardizes and cleanses data. Feature extraction identifies relevant predictors, utilized in training various classification algorithms. Hyperparameter tuning optimizes model performance, evaluated using diverse metrics. Interpretability techniques enhance transparency in decision-making [12]. This architecture offers a scalable framework leveraging advanced machine learning for accurate diabetes prediction, aiming to enhance healthcare outcomes.

a) Real-Time Testing Model:

Trained AdaBoost, RF, and GB models are deployed for real-time predictions on new data in a production environment. Continuous monitoring of patient data streams enables real-time risk assessment based on demographics, medical history, and lifestyle factors. Performance metrics like accuracy and precision are tracked to evaluate model effectiveness. Feedback from model predictions informs iterative improvements, ensuring accuracy [14] and reliability over time. Scalability and reliability are prioritized to handle varying workloads seamlessly. This real-time testing approach ensures optimal model performance under dynamic conditions, enhancing the efficiency of diabetes risk prediction.

b) Validation And Training Model:

Data collection, pre-processing, and choosing features are all part of the training and validation process before the dataset is split. Cross-validation and hyperparameter tweaking are used to train machine learning algorithms such as logistic regression or decision trees on the training data set. Model evaluation metrics like as preciseness and accuracy are computed on the testing set, and approaches such as SHAP [11] values improve interpretability. The deployed model is accessible via a user-friendly interface, with continuous monitoring and updates to ensure relevance and user privacy. Ethical considerations and user education are integrated into the system for comprehensive support.

B. Proposed Workflow of Methodology:

a) **Data Acquisition**: Gather diverse datasets containing pertinent diabetes risk factors, including patient demographics, medical history, lifestyle behaviours, clinical measurements, and biomarkers associated with blood sugar regulation.
b) **Data Pre-processing**: Perform meticulous data pre-processing to address anomalies like missing values, outliers, and inconsistencies in formatting. Standardize numerical attributes and encode categorical variables to ensure uniformity across the dataset.
c) **Feature Engineering and Selection**: Employ advanced strategies to extract informative attributes and derive novel variables. Techniques such as correlation [5] analysis, dimensionality reduction, and domain expertise integration guide feature selection.
d) **Model Development**: Set up three machine learning models: AB optimization, Random Forest, and Gradient Boosting. Train each model to capture both early-stage and overall diabetes risk factors.
e) **Model Evaluation**: Evaluate model performance with metrics including accuracy, precision, recall, the F1 score.
f) **Integration and Ensemble Learning**: Combine results from classifiers to create an ensemble model. Techniques such as model averaging and stacking improve forecast accuracy.
g) **Model Deployment and Interpretability**: Deploy the ensemble model within a user-friendly interface. Facilitate model interpretability through feature importance analysis, SHAP values, or LIME explanations.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 11 Issue 6 June 2024

h) **Continuous Monitoring and Updating**: Implement mechanisms for ongoing model performance monitoring and regular updates with new data. Solicit feedback for iterative refinement.

a) Datasets:

The Early-Stage Diabetes Risk Prediction Dataset is accessible on Kaggle and comes from Bangladesh's Sylhet Diabetes Hospital. It consists of medical professional-approved data gathered via direct patient questionnaires. This dataset facilitates the evaluation of early-stage diabetes risk. Similarly, key input variables for predicting diabetes risk are provided by the Diabetes Risk Prediction Dataset, which was also sourced via Kaggle. Pregnancies, blood pressure, glucose levels, skin thickness, insulin levels, BMI, the function of the diabetic pedigree, and age are some of these characteristics. By using both datasets, researchers can create predictive models that improve diabetes treatment and diagnosis.

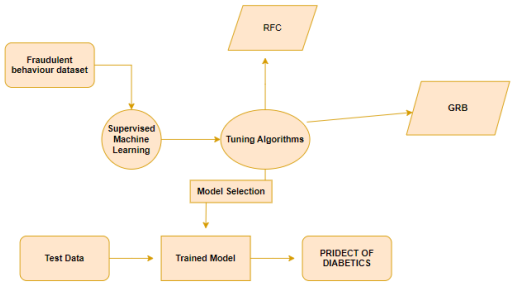

Figure 2. Entity Relationship Diagram

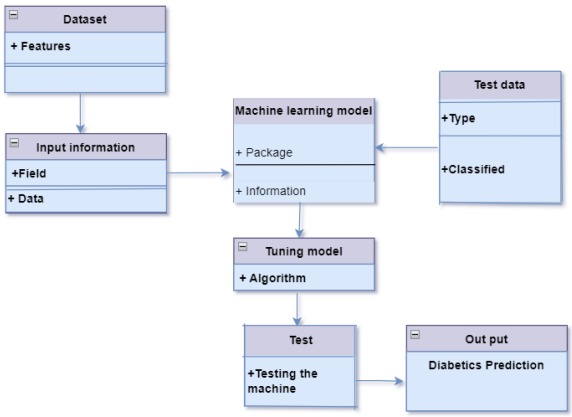

Figure 3. UML Diagram

V. MODULE DESCRIPTION:

A. Data Pre-Processing:

Validation methods in ML ensure accurate model predictions by assessing model performance on unseen data. In real-world scenarios, where data samples may not fully represent the population, validation becomes crucial. Techniques [2] like cross-validation and train-test splitting provide unbiased evaluation, helping to fine-tune model hyperparameters. Understanding data properties aids in choosing appropriate algorithms, while data cleaning tasks, such as handling missing values using Python's Pandas library, streamline the process. Identifying different types of missing data informs imputation strategies, contributing to robust model development.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 520 entries, 0 to 519
Data columns (total 17 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Age                 520 non-null    int64
 1   Gender              520 non-null    int32
 2   Polyuria            520 non-null    int32
 3   Polydipsia          520 non-null    int32
 4   sudden weight loss  520 non-null    int32
 5   weakness            520 non-null    int32
 6   Polyphagia          520 non-null    int32
 7   Genital thrush      520 non-null    int32
 8   visual blurring     520 non-null    int32
 9   Itching             520 non-null    int32
 10  Irritability        520 non-null    int32
 11  delayed healing     520 non-null    int32
 12  partial paresis     520 non-null    int32
 13  muscle stiffness    520 non-null    int32
 14  Alopecia            520 non-null    int32
 15  Obesity             520 non-null    int32
 16  class               520 non-null    int32
dtypes: int32(16), int64(1)
memory usage: 36.7 KB
```

Figure 4. Input Features.

B. Data Visualization:

Data visualization is integral to our comprehensive diabetes prediction approach, offering insights into data attributes and patterns. Initially, exploratory data analysis identifies trends, outliers, and relationships, guiding feature selection. Visualization aids in model interpretation, clarifying significant feature contributions to risk prediction. It also facilitates performance evaluation, comparing model metrics like accuracy and ROC [7] curves. For user interaction, interactive dashboards enable individuals to input data, visualize personalized risk assessments, and understand contributing factors. Overall, visualization enhances understanding and engagement, crucial for effective diabetes risk assessment and management.

Figure 5. Instances Of Each Attribute in Data Set.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 11 Issue 6 June 2024

C. Algorithm Implementation:

a) Performance Metrics to Calculate:

When doing binary classification jobs, two performance indicators are used: True Positive Rate (TPR) and False Positive Rate (FPR). The True Positive Ratio (TPR), alternatively referred to as Sensitivity or Recall, quantifies the percentage of accurately anticipated positives, or true positives, among all positive occurrences. The ratio of True Positives (TP) to the total of True Positives and False Negatives (FN) is used to compute it. Conversely, false positive rate (FPR) is the ratio of false positives (erroneously projected positives) to all real negative cases. The ratio of False Positives (FP) to the total of False Positives and True Negatives (TN) is used to compute it.

The ratio of properly predicted observations (TP and TN) to the total number of observations is used to determine accuracy, which is a measure of the overall correctness of the model's predictions.

By calculating the percentage of true positive predictions among all positive predictions, precision [3] measures how accurate the model is at making positive assertions.[2]

The model's recall (also known as sensitivity) quantifies how well it can distinguish true positive examples from all real positive examples.

Precision and recall are harmonic means, and the F1 Score strikes a balance between both. When working with imbalanced datasets, where the proportion of positive and negative examples varies significantly, it is very helpful.

b) Random Forest:

Multiple decision trees trained on random subsets of the data and features are combined in Random Forest, an ensemble learning technique. A majority vote determines the final product, with predictions made by each tree. Let Y be the target variable and X represent input features mathematically. Random Forest comprises decision trees T_i, each making a prediction Ŷ_i. The final prediction Ŷ is obtained through majority voting. Random Forest is robust to noise, scalable, and provides feature importance estimation. Its ability to mitigate overfitting while maintaining high accuracy makes it widely used in classification and regression tasks across domains.

c) Gradient Boosting:

It is an ensemble learning technique that creates decision trees one after the other to fix mistakes produced by the ones before it. By repeatedly fitting fresh trees to the residuals of earlier forecasts, iteratively minimizing a loss function until convergence or a predetermined limit is reached. It is efficient at capturing subtle feature-target links in complex datasets. Its flexibility, allowing different loss functions and regularization techniques, makes it popular for regression and classification tasks in diverse domains.

d) Ada Boost:

AdaBoost (Adaptive Bosting) is a collaborative learning method that builds a strong classifier by repeatedly combining weak classifiers. It focuses on misclassified instances by adjusting their weights in each iteration. Each weak learner is trained on a modified dataset, emphasizing previously misclassified samples [10]. The weighted forecasts of all weak learners are combined to get the final prediction. This model is renowned for its versatility and robustness in binary classification tasks, often outperforming individual classifiers. Its effectiveness, simplicity, and resistance to overfitting make it a popular choice in several fields, such as bioinformatics, natural language processing, and computer vision.

VI. CONCLUSION AND FUTURE SCOPE

In conclusion, the Comprehensive Diabetes Prediction system marks a significant milestone in healthcare analytics, amalgamating early-stage and overall risk assessment. Leveraging RF and AdaBoost for early-stage prediction, achieving accuracies of 86.4% and 77% respectively, and Gradient Boosting (GB) for diabetes prediction with 100% accuracy underscores its efficacy. Future endeavours should prioritize model refinement, feature exploration, and interdisciplinary collaboration to ensure regulatory compliance and real-world applicability. Continuous performance monitoring, coupled with iterative updates, is imperative to sustain predictive accuracy. Embracing emerging technologies like federated learning and blockchain holds promise for further enhancing system utility and scalability.

REFERENCES

[1] S Sivaranjani, S Ananya, J Aravinth, R Karthika, "Diabetes Prediction using Machine Learning Algorithms with Feature Selection and Dimensionality Reduction", 2021, 10.1109/ICACCS51430.2021.9441935

[2] Sneha N, Gangil T. Analysis of diabetes mellitus for early prediction using optimal features selection. J Big Data. 2019; 6:1. 10.1186/s40537-019-0175-6

[3] Alsulami S, et al. Effect of dietary fat intake and genetic risk on glucose and insulin-related traits in Brazilian young adults. J Diabetes Metab Disord. 2021;1337–47. 10.1007/s40200-021-00863-7. [PMC free article] [PubMed]

[4] Mohammadi H, Eshtiaghi R, Gorgani S, Khoramizade M. Assessment of Insulin, GLUT2 and inflammatory cytokines genes expression in pancreatic β-Cells in zebrafish (Danio rario) with overfeeding diabetes induction w/o glucose. J. Diabetes Metab. Disord. 2021;20(2):1567–1572. doi: 10.1007/s40200-021-00903-2. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

[5] International Diabetes Federation. Eighth edition. 2017;2017.

[6] Kaur P, Sharma M. Analysis of Data Mining and Soft Computing Techniques in Prospecting Diabetes Disorder in Human Beings: a Review. Int. J. Pharm. Sci. Res. 2018;9(7):2700–2719. doi: 10.13040/IJPSR.0975-8232.9(7). 2700-19. [CrossRef] [Google Scholar]

[7] R. Sengamuthu, R. Abirami, and D. Karthik, "Various Data Mining Techniques Analysis to Predict," 2018.

[8] A. Anand and D. Shakti, "Prediction of diabetes based on personal lifestyle indicators," Proc. 2015 1st Int. Conf. Next Gener. Comput. Technol. NGCT 2015, no. September, pp. 673–676, 2016, doi: 10.1109/NGCT.2015.7375206.

[9] Jha RP, Shri N, Patel P, Dhamnetiya D, Bhattacharyya K, Singh M. Correction to: Trends in the diabetes incidence and mortality in India from 1990 to 2019: a joinpoint and age-period-cohort analysis. J. Diabetes Metab. Disord. 2021;20(2):1741. doi: 10.1007/s40200-021-00865-5. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

[10] Diabetes Federation International and IDF, IDF Diabetes Atlas 2019, 9th Editio. 2019.

[11] Naz H, Ahuja S. Deep learning approach for diabetes prediction using PIMA Indian dataset. J. Diabetes Metab. Disord. 2020;19(1):391–403. doi: 10.1007/s40200-020-00520-5. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

[12] Nissa N, Jamwal S, Mohammad S. Early Detection of Cardiovascular Disease using Machine learning Techniques an Experimental Study. Int. J. Recent Technol. Eng. 2020;9(3):635–641. doi: 10.35940/ijrte.c46570.99320. [CrossRef] [Google Scholar]

[13] S. M. Ganie, M. B. Malik, and T. Arif, "Machine Learning Techniques for Diagnosis of Type 2 Diabetes Using Lifestyle Data," in International Conference on Innovative Computing and Communications, 2022, pp. 487–497.

[14] Ramesh D, Katheria YS. Ensemble method based predictive model for analyzing disease datasets: a predictive analysis approach. Health Technol. (Berl). 2019;9(4):533–545. doi: 10.1007/s12553-019-00299-3. [CrossRef] [Google Scholar]

[15] Ganie SM, Malik MB, Arif T. Various Platforms and Machine Learning Techniques for Big Data Analytics. A Technological Survey. 2018;3(6):679–687. [Google Scholar]

[16] Choubey DK, Paul S. Classification techniques for diagnosis of diabetes: A review. Int. J. Biomed. Eng. Technol. 2016;21(1):15–39. doi: 10.1504/IJBET.2016.076730. [CrossRef] [Google Scholar]

[17] Georga EI, Protopappas VC, Bellos CV, Fotiadis DI. Wearable systems and mobile applications for diabetes disease management. Health Technol. (Berl). 2014;4(2):101–112. doi: 10.1007/s12553-014-0082-y. [CrossRef] [Google Scholar]

[18] Mohebbi A, Aradottir TB, Johansen AR, Bengtsson H, Fraccaro M, Morup M. A deep learning approach to adherence detection for type 2 diabetics. Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. 2017; EMBS:2896–2899. doi: 10.1109/EMBC.2017.8037462. [PubMed] [CrossRef] [Google Scholar]

[19] R. Barhate and P. Kulkarni, "Analysis of Classifiers for Prediction of Type II Diabetes Mellitus," Proc. - 2018 4th Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2018, pp. 1–6, 2018, doi: 10.1109/ICCUBEA.2018.8697856.

[20] M. Kowsher, M. Y. Turaba, T. Sajed, and M. M. Mahabubur Rahman, "Prognosis and treatment prediction of type-2 diabetes using deep neural network and machine learning classifiers," 2019 22nd Int. Conf. Comput. Inf. Technol. ICCIT 2019, no. December, pp. 18–20, 2019, doi: 10.1109/ICCIT48885.2019.9038574.