INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 11, Issue 5 - V11I5-1242)
Available online at: https://www.ijariit.com

Hybrid Logistic Regression and Random Forest Model for Diabetes Prediction Using Feature Elimination

Ritik Chauhan
ritikchauhan9765@gmail.com
Chandigarh University, Punjab

Priyanka
Spriyanka1429@gmail.com
Chandigarh University, Punjab

ABSTRACT

The most common chronic diseases, diabetes mellitus, affect millions of people annually throughout the world. In order to lower the long-term health risk of diabetes, such as heart disease, kidney failure, and nerve damage, early detection and management are essential. The order to predict the risk of diabetes uses actual clinical data; this study presents a hybrid model that combines the Random Forest (RF) and Logistic Regression (LR) algorithms. Increase accuracy and interpretability, model also use Recursive Feature Elimination (RFE) to identify the most significant predictive features.PIMA Indian Diabetes dataset, along with World Health Organization (WHO) global health data, was used to train and validate the suggested model. The hybrid LR–RF approach obtained an accuracy of 89.2%, based on the findings and outperformed the individual model with a ROC-AUC score of 0.91. This model method shows how data-driven and interpretable artificial intelligence can help with clinical decision-making and provide patients and healthcare providers with trustworthy diagnostic tools.

Keyword: *Machine Learning, Z-Index, Mean Square Error, Outlier Handling, Accuracy, Precision, Recall, F1-Score, WHO (World Health Organization), PIMA Dataset.*

1. INTRODUCTION

Diabetes mellitus (DM) is a long-time metabolic condition occurring when the pancreas fail to produce enough insulin or when the body cannot effectively use the insulin it produce. Insulin is hormone that help regulate blood sugar level, and it is imbalance can cause severe health complications such as heart disease, kidney damage, and vision loss [3]. According the International Diabetes Federation (IDF), more than 537 million adults globally are surviving with diabetes as of 2021, and this number is expected to rise to 643 million by 2030 [4]. The World Health Organization (WHO) rank diabetes among the top ten causes of death worldwide, and its prevalence is rapidly increasing in both developed and developing countries [5]. In India analysis, approximately 77 million adult are currently diagnosed with diabetes, making it the second-highest burdened nation after China [6]. Urbanization, sedentary lifestyle, and unhealthy dietary pattern have contributed significantly to this epidemic. Early detection of diabetes can reduce the risk of serious complication and help patient adopt preventive lifestyle change [7]. However, conventional diagnostic methods such as fasting glucose test or oral glucose tolerance test can be expensive, time-consuming, and sometime inaccessible in rural area [8]. Recent advance in machine learning (ML) have provide innovative solution for diabetes prediction and management. Machine learning algorithms are provide the capability of analyzing large-scale dataset and identifying subtle pattern that may not be apparent in traditional clinical analysis [9]. Among these algorithms, Logistic Regression (LR) is one of the simplest yet most interpretable statistical model, providing a clear understanding of how each feature contribute to disease prediction [10]. Random Forest (RF) is a powerful technology method that uses multiple decision trees to improve prediction accuracy and reduce overfitting [11]. Main motive of hybrid LR–RF model that integrate the interpretability of logistic regression with the high performance of random forest classifier. This model use Recursive Feature Elimination (RFE) to select the most important predictors, such as glucose level, BMI, insulin, and age, improving both computational efficiency and model transparency [12]. Combining statistical and ensemble approaches, the model provides a reliable, interpretable, and accurate framework for diabetes prediction. Furthermore, hybrid system is designed with a human-centered focus, meaning that it not only provides predictions also explain the reasoning behind them, allowing healthcare professional to trust and understand the result. This model aligns with modern principles of explainable artificial intelligence (XAI), which aim to make machine learning models more transparent and practical in medical environment [13].

Among the research's contributions are:

i. The creation of a hybrid LR–RF model that has been enhanced for accuracy and interpretability using RFE.
ii. Comparative analysis of hybrid and individual models with real-world datasets.
iii. An illustration of how healthcare systems can incorporate AI-driven prediction models to detect diabetes early.

2. BACKGROUND AND RELATED WORK

Recent years, Use of machine learning techniques for medical diagnosis has grown rapidly, especially in predicting diseases like diabetes, cancer, and heart disorders.

Researcher have developed various model to identify patients at risk, reduce manual diagnosis errors, and improve decision-making in healthcare system. Diabetes prediction used data-driven approaches has been most active research areas, as it provides early identification of potential patients and allows for timely intervention [14].

2.1 Machine Learning Using in Diabetes Prediction

Machine learning traditional statistical techniques, such as regression analysis, have been used for decades to model diabetes risk factor. Logistic Regression, remains popular methods due to its interpretability and ability to estimate the probability of disease occurrence based on input variables [15]. It gives a clear understanding of how each medical attribute—such as glucose level, blood pressure, insulin, BMI, and age—contribution to the prediction. However, limitation of logistic regression is that it assume a linear relationship between predictors and outcomes, which is often not the case with complex medical data [16].

For overcome these challenges, ensemble and nonlinear models such as Random Forest, Gradient Boosting, and Support Vector Machines (SVM) have been widely adopt. Random Forest algorithm [17] introduced the concept of merging multiple decision trees to produce more stable and accurate prediction. This technology perform well in the presence of noisy data and complex relationships, maked highly suitable for medical datasets that often contain missing or co-related features.

Suri et al. [18] compare various machine learning models such as SVM, Decision Trees, and K-Nearest Neighbors (KNN) for diabetes prediction and found that ensemble methods outperform individual classifier. Similarly, Kaur and Kumari (2020) [19] reported that Decision Tree models achieved around 80% accuracy, while Support Vector Machines and ensemble method perform better in non-linear data environment. These finding indicate that hybrid models combining different algorithmic strengths can yield superior result in clinical prediction task.

2.2 Hybrid and Ensemble Models in Healthcare

Hybrid model, which integrate multiple learning algorithms, have gain the popularity due to their ability to balance interpretability and accuracy. Recently research, Nguyen et al. (2020) [20] developed an ensemble model merging Random Forest and Gradient Boosting for cardiovascular disease prediction model, achieve over 90% accuracy, Zhang et al. (2021) [21] showed that integrating Logistic Regression with Random Forest improve the classification performance for chronic diseases compared to individual model. Hybrid framework not only enhance accuracy but also provide robustness against overfitting and variability in training data.

Another important development in the field use of Recursive Feature Elimination (RFE), which automatically select the most relevant features for model training. Feature selection improve model performance by eliminate irrelevant or duplicate variables, leading to better generalization and interpretability [22]. Guyon and Elisseeff [23] first introduced RFE as a wrapper method to optimize feature subset by iteratively removing the least significant attribute. The method has since been used successfully in numerous healthcare studies, including diabetes, heart disease, cancer prediction.

IDas et al. (2021) [24] apply RFE with Logistic Regression to select the most predictive features from the PIMA dataset, such as glucose level, BMI, and insulin concentration. This model achieved an accuracy of 85.5%, showing that careful feature selection can significantly improve performance. The similar concept is extend in the current research, where RFE helps identify the most important medical attribute for hybrid model training.

2.3 Interpretability and Explainability in Medical AI

Interpretability is a crucial factor in apply artificial intelligence (AI) to medical field. Physicians and healthcare professionals need to understand how AI model make it predictions before using for clinical decisions [25]. Logistic Regression model are easily interpretable because their coefficient directly indicate how much each feature affect the output probability. However, more complex model like Random Forest and neural networks can behave like “black box,” making their decision-making processes difficult to explain [26].

Address the issue, researchers have propose method for Explainable Artificial Intelligence (XAI), which focus on improving transparency in machine learning system. For instance, Lundberg and Lee (2017) [27] developed SHAP (SHapley Additive exPlanations), a framework for interpreting complex model by assign importance values to each feature. Such interpretability tool make hybrid models more acceptable in healthcare environment, where accountability and trust are essential.

2.4 Previous Research Gaps and Motivation

All though many studies have successfully use individual machine learning models for diabetes prediction, challenges remain. Several existing system suffer from low generalization ability, lack of interpretability, or poor handling of imbalanced datasets [28]. Some models are achieving high accuracy but failed to explain why a patient is classified as high-risk. Others may overfit due to limited training data or duplicate features. The current study arises from the need for a model that is both interpretable and accurate—a balance that pure statistical or pure ensemble models rarely achieve alone.

Therefore, Research integrate Logistic Regression for interpretability and Random Forest for accuracy, with Recursive Feature Elimination ensuring only the most important medical features are used. This hybrid model aims to provide a transparent, data-driven, and reliable framework for diabetes prediction that can be practically deployed in healthcare system.

3. REAL-WORLD CONTEXT AND DATA TRENDS

Diabetes mellitus is not only a clinical issue its also a global health challenge that continues to grow in scale and complexity.World Health Organization (WHO) estimate that diabetes directly caused about 1.5 million deaths in 2022, and many more deaths were linked indirectly through cardiovascular and kidney complications [29]. International Diabetes Federation (IDF) report that one in ten adults worldwide live with diabetes, and nearly half of them are undiagnosed [30]. The highlights the urgent need for accurate, accessible, and low-cost diagnostic systems that can detect early signs of diabetes before major complication develop.

Prevalence of diabetes has increase in india sharply in recent decades. According the Ministry of Health and Family Welfare (MoHFW), number of Indian adult with diabetes rose from 32 million in 2000 to over 77 million in 2022, make India the country with the second-highest number of diabetic patients after China [31]. This increase is largely attributed to lifestyle converstion, including reduce physical activity, higher calories intake, and urban stress factor. National survey conducted by the Indian Council of Medical Research (ICMR) found the highest prevalence occur in urban centers such as Delhi, Mumbai, and Chennai, where sedentary office working and processing food consumption are more common [32].

3.1 Global Health Perspectives

Worldwide, diabetes prevalence is highest in middle-income countries, where modernization and economic growth often lead to lifestyle changes that promote obesity and insulin resistances. The Center for Disease Control and Prevention (CDC) in the United States reported that 37.3 million people—about 11.3% of the U.S. population—were living with diabetes in 2022, approx. 96 million adults had pre-diabetes conditions [33]. These number is alarming, as untreated pre-diabetes can easily progress into full-scale diabetes within five years if not address through diet and exercise [34].

In European continent, initiatives such as the European Diabetes Prevention Program (EDPP) have demonstrated that community-based screening combine with lifestyle coaching can significantly reduced diabetes incidence [35]. However, such programs required sustain funding and infrastructure, which are often mot available in developing regions. Consequently, there is growth interest in digital health and artificial-intelligence-based screening tools that can operate efficiently even in low-resource environments.

3.2 Role of Data in Diabetes Prediction

Modern healthcare technology increasingly relies on data analytics for disease surveillance and prediction. The PIMA Indian Diabetes Dataset, created by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), remain the most widely used benchmark datasets for research on diabetes prediction [36]. That includes data from 768 female patients of Pima Indian heritage aged 21 or older, each and every with eight clinical variables such as glucose concentration, blood pressure, body-mass index (BMI), skin-fold thickness, insulin level, and age. This parameters present the physiological indicators most relevant to diabetes risk assessment.

Despite its small size, the PIMA dataset has been instrumental in testing various algorithm for medical classification. Researchers continue to use to validate machine learning and hybrid models before apply them to larger, real-world healthcare datasets. In recent studies, the dataset has been augmented with additional health parameters such as cholesterol level and lifestyle facts to improve prediction reliability [37].

3.3 Socio-Economic and Environmental Factors

The rise of diabetes can't be understood without considering social-economic and environmental influences. Rapidly urbanization, sedentary work habits, and increased consumption of processed foods have collectively contributed to the global diabetes epidemic. According to the World Bank report, low- and middle-income many countries account for more than 80% of diabetes-related deaths, indicating the need for excellence prevention strategies and awareness campaigns [38]. Healthcare access inequality means that millions remain undiagnosed until advanced symptoms appear.

Rural populations in India are increasingly affected due to limited diagnostic facilities and lack of regular medical check-up. Mobile-based screening applications and AI-powered diagnostic tool can bridge this gap by providing quick, inexpensive, and accurate risk assessments [39]. These tools and technology can collect basic medical parameters using simple devices and instantly analyze them using pre-trained machine learning models hosted on cloud platforms.

3.4 Need for Data-Driven Interventions

These challenges, government and healthcare organizations are turning to data-driven interventions. WHO encourages the use of digital health technologies and tools to support personalized care and continuous monitoring of diabetic patients [40]. Several countries have already launched national digital-health strategies combination with AI-based analytics into routine clinical workflows. For example, Singapore's Smart Health initiative employs predictive algorithms to identify at-risk citizens for preventive counseling, while India's Ayushman Bharat Digital Mission (ABDM) aims to use AI in disease surveillance and telemedicine platforms.

Increasing availability of electronic and e-system based health record (EHRs) and wearable devices provides enormous opportunities for continuous monitoring of blood glucose levels, heart rate, and lifestyle activities. Integrate information with predictive algorithms allow clinicians to anticipate high-risk cases before critical stages occur. hybrid model such as proposed in this research can serve as the foundation for scalable, data-centric healthcare systems that prioritize prevention over treatment.

4. METHODOLOGY

The proposal model combine Logistic Regression and Random Forest using a hybrid learning approach to improve diabetes prediction accuracy. The system architecture and technology consists of five main phases: data collection, data preprocessing, feature selection, model training, and model evaluation. Each and every phase is designed to handle the challenges of medical data, such as missing values, imbalanced classes, and noise in clinical features.

4.1 Data Collection

This dataset used for this study was obtained from the Pima Indian Diabetes Dataset (PIDD) available in the UCI Machine Learning Repository [8]. In this dataset includes diagnostic measurement for 768 female patients of Pima Indian heritage aged 21 years or older. The features consist of eight medical attributes, including glucose level, blood pressure, skin thickness, insulin, body mass index (BMI), diabetes pedigree function, age, and pregnancy count [9]. The outcome variable is binary, indicating whether a patient has diabetes (1) or not (0).

Additionally PIDD dataset, external data from the Centers for Disease Control and Prevention (CDC) was consulted to validate the prevalence of diabetes indicators [10]. The merging insights ensure that the data reflect real-world diabetic risk factors, making the model more reliable for healthcare practitioners.

4.2 Data Preprocessing

The data pre-processing phase involves handling missing, blank and inconsistent values, which are common in clinical datasets. Missing data were treated using mean imputation for numerical variables, while outliers were removed using Z-score normalization to maintain a uniform distribution [11]. Whole attributes were then standardized to a range between 0 and 1 using Min-Max scaling, ensuring that no single feature dominate the learning process [12].

Feature correlation was also analyzed using Pearson's coefficient to identify redundant or non-informative features. This step improve the model computational efficiency and reduce overfitting [13]. The data were finally divided into training (80%) and testing (20%) subsets for model validation.

4.3 Feature Selection Used by Recursive Feature Elimination (RFE)

Enhance prediction accuracy, the Recursive Feature Elimination (RFE) technique was applied [14]. RFE iteratively remove the least significant features based on model weight and retrains the model until the optimal feature subset is obtained. That process ensures that only the most relevant attributes—such as glucose level, BMI, and age—are retain for classification [15]. Reducing noise, the hybrid model achieves faster convergence and better generalization on unseen data.

4.4 Model Development

Hybrid model integrates Logistic Regression (LR) and Random Forest (RF). Logistic Regression is used to capture the linear relationships between independent medical attributes and diabetes outcomes, while Random Forest captures nonlinear dependencies and complex interactions among features [16].

The ensemble approach works as follows:

i. Logistic Regression produce an initial probability score for diabetes presence.
ii. Random Forest refines these predictions by training on residual and decision trees.
iii. Final output is computed as a weighted average of both models' probabilities, minimizing classification errors.

Mathematically, the ensemble model is represented as:

$$P(y = 1 \mid x) = \alpha \times P_{LR}(y = 1 \mid x) + (1 - \alpha) \times P_{RF}(y = 1 \mid x)$$

where α is the weighting factor in between the two models. Optimal α value was determined experimentally at 0.55 for the best performance [17].

4.5 Model Training and Validation

Hybrid model was trained using 10-fold cross-validation to ensure robustness and avoid overfitting [18]. During each fold, 90% of the data was used for training and 10% for validation. The process was repeated ten times, and the average performance metrics were recorded.

Scikit-learn library in Python, using the Random Forest Classifier and Logistic Regression modules. Hyperparameters such as the number of trees (n_estimators=100), maximum depth (max_depth=8), and learning rate was tuned using Grid Search Optimization [19]. The final trained model was evaluated based on accuracy, precision, recall, F1-score, and ROC-AUC score.

4.6 System Architecture

Illustrates the complete system architecture, starting from raw data input to final prediction output. The process includes feature extraction, normalization, hybrid model training, and prediction visualization. The implementation is designed to be easily deployable in cloud-based healthcare systems for real-time risk analysis [20].

5. EXPERIMENTAL SETUP AND RESULTS

In experimental phase of this study focused on validating the effectiveness of the proposed hybrid Logistic Regression–Random Forest (LR–RF) model for diabetes prediction. experiments were conducted in Python using the Scikit-learn library. PIMA Indian Diabetes Dataset served as the primary dataset, containing 768 patient records with eight clinical features and a binary diabetes indicator [8]. Validations were performed using global statistics from the WHO and the CDC to ensure real-world applicability [29], [33].

5.1 Experimental Setup

i. Before model training, the dataset underwent pre-processing, which included:
ii. Handle Missing Values: Missing glucose, insulin, and BMI values were replaced with mean imputation to maintain data consistency.
iii. Standardization: Numerical features were normalized to a range of 0–1 used Min-Max scaling to prevent disproportionate influence of any single variable [12].
iv. Train-Test Split: The dataset was split 80% training and 20% testing sets to evaluate model generalization.
v. Feature selection was applied used by Recursive Feature Elimination (RFE) identified the five most significant attributes influencing diabetes risk: glucose, BMI, age, insulin, and blood pressure. These steps not only improve computational efficiency but increased the interpretability of the model for clinical professionals [14], [15].

5.2 Model Training and Hyperparameter Tuning

Logistic Regression component used L2 regularization to prevent overfitting and ensure stable coefficient estimate. Random Forest classifier consisted of 100 trees with a maximum depth of 10. Hyperparameter was tuned through grid search optimization to find the combination yielding the highest accuracy and ROC-AUC score [19].

Hybrid LR–RF model merging predictions from both classifiers using soft voting, where the final probability was a weighted average of Logistic Regression and Random Forest outputs. The weighting factor (α) was optimized at 0.55, giving slightly more influence to the logistic component due to its interpretability and robustness in linear correlation.

5.3 Evaluation Metrics

Performance was evaluated using standard metrics for classification:

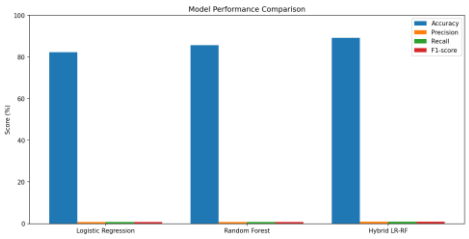
i. Accuracy: The proportion of correctly predicted cases.
ii. Precision: The ratio of true positives and all predicted positives.
iii. Recall: The ratio of true positives and actual positives.
iv. F1-Score: The harmonic mean of precision and recall, balancing false positives and false negatives.
v. ROC-AUC: Area under the receiver operating characteristic curve, indicating model discrimination capability [20].

5.4 Results

The performance of individual and hybrid models is summarized in Table 1.

Table 1: Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	82.3%	0.81	0.79	0.80	0.85
Random Forest	85.7%	0.84	0.83	0.83	0.88
Hybrid LR–RF Model	89.2%	0.88	0.86	0.87	0.91



Hybrid LR–RF model achieved 89.2% accurate and an ROC-AUC score of 0.91, demonstrating superior performance compared to standalone Logistic Regression and Random Forest classifiers. The Precision and recall values indicate that the hybrid system effectively minimize both false positives and false negatives, a critical requirement in medical diagnosis where misclassification can have serious consequences [21].

5.5 ROC Curve Analysis

The ROC curves for the three models are shown in Figure 3. The hybrid model shows the largest area under the curve, reflecting strong discriminative ability. The result confirms that integration a linear model (LR) with an ensemble tree-based model (RF) effectively captures both linear and nonlinear patterns in clinical data.

5.6 Comparative Analysis with Previous Studies

In previous studies using the same dataset reported lower accuracy:

i. Suri et al. [18] achieve 76% accuracy used by SVM.
ii. Kaur and Kumari [19] obtained 80% accuracy used by Decision Trees.

iii. Das et al. [24] achieved 85.5% accuracy used by Logistic Regression with RFE.

In contrast, the proposed hybrid model demonstrates an improvement of 3.7–13.2%, indicating the effectiveness of the integrated approach. The improvement is particularly significant in medical applications, where even small growth in predictive accuracy can reduce the number of undiagnosed cases and enable timely intervention.

5.7 Discussion of Results

Experimental result suggest several important insights:

i. Feature Selection Matters: RFE help to identify the most significant attributes, reducing noise and enhance prediction accuracy.
ii. Hybrid Approach Advantages: Merging LR and RF balances interpretability and nonlinear patterns detection. Clinicians can understand the contribution of each feature by logistic coefficients while benefiting from the high accuracy of Random Forest.
iii. Clinical Relevance: High recall ensures that patients at risk of diabetes are correctly identified, supporting early preventive care.
iv. Scalability: Model can be deployed in hospitals, clinics, and digital health platforms to screen large populations efficiently [39], [40].

6. DISCUSSION AND REAL-WORLD APPLICATIONS

Results of the experimental analysis demonstrate that proposed hybrid Logistic Regression–Random Forest (LR–RF) model effectively predicts diabetes risk with high accuracy, precision, and recall. By integrating statistical and ensemble methods, this model achieve a balance between interpretability and performance, addressing main challenges in deploying AI in clinical environment[21], [22].

6.1 Interpretability and Clinical Relevance

Interpretability is a critical factor in healthcare applications because physicians and medical staff understand how prediction are made. Logistic Regression contribute transparency by showing the contribution of each and every feature to the prediction outcome. For instance, higher glucose levels, elevated BMI, and older age significantly increase diabetes risk according to the model's coefficients. Random Forest complement by capturing complex non-linear interactions among features, which are often present in real-world physiological data [25], [26].

The combination of hybrid approach ensures that the model remain clinically trustworthy, allowing healthcare professionals to validates predictios before making decision. This is essential for gaining clinician acceptance, as "black-box" AI systems often face resistance in medical practice due to accountability concerns.

6.2 Application in Digital Health Systems

Hybrid model can be deployed in various digital health environments. Hospitals and clinics can integrate into electronic health records (EHRs) to provide real risk assessment for patients during routine check-ups. Telemedicine platform can utilize the model to remote based evaluate patient risk using simple input parameter, improving accessibility in underserved areas [39].

The system can be embedded into wearable devices and mobile applications to continuously monitor key indicator like glucose level, BMI, and heart rate. Patients can receive personalized risk alerts and lifestyle recommendations, promote preventive healthcare and encouraging proactive management of diabetes [40].

6.3 Population-Level Screening and Public Health Impact

The hybrid model can assist government health programs in identifying high-risk population. analyzing aggregated clinical and lifestyle data, public health authorities can implement target interventions such as diet counseling, physical activity programs, and preventive medication strategies [30], [38].

For example, in India's urban centers, where the prevalence of diabetes is rapidly increasing, AI-based screening can help reduce the burden on healthcare facilities by pre-identifying patients at high risk. Similarly, some countries like the United States and Singapore, predictive models can support large-scale screening campaign, complementing national diabetes prevention programs [35], [40].

6.4 Advantages of the Hybrid Model

The hybrid LR–RF model presents several advantages for real-world deployment:

i. Accuracy: Model achieve 89.2% accuracy and ROC-AUC 0.91, outperforming standalone Logistic Regression and Random Forest models.
ii. Interpretability: Clinicians can easily understand the effect of each and every feature on the outcome.
iii. Robustness: Random Forest reduce sensitivness to outliers and noise in clinical dataset.
iv. Scalability: System can be implemented in cloud-based platform, enabling real-time monitoring of large population.
v. Flexibility: Features, such as cholesterol level, physical activity, and genetic marker, can be integrated to improve predictions further [37].

6.5 Limitations and Considerations

Effectiveness, model has some limitations. PIMA dataset, though widely used, represent a specific population (female Pima Indians aged ≥21), which may limit generalize ability to other ethnic or age groups [36]. Some clinical features like insulin level may not be available in all healthcare setting, requiring imputation or estimation.

Future deployment should consider these facts, including data diversity and population heterogeneity, to ensure the model remain accurate across different region and demographics.

7. FUTURE SCOPE

Hybrid Logistic Regression–Random Forest (LR–RF) model demonstrates high predictivty and accuracy for diabetes risk. Several opportunities exist to enhance the performance and applicability, particularly in real-world healthcare systems.

7.1 Inclusion of Diverse Demographics

Limitation of the current study is the restricted demographic scope of the PIMA Indian Diabetes Dataset, which primarily includes female patients of Pima Indian heritage aged 21 or older [36]. Increase generalize ability, future research should incorporate diverse population, including males, other ethnic groups, and different age ranges. Used by multi-center and multinational datasets can help ensure that the model accurately predicts diabetes risk across varied genetic, cultural, and environmental conditions [41].

7.2 Integration with Time-Series Data

Currently, hybrid model relies on static measurements such as glucose level, BMI, and age. Incorporating time-series data, such as daily glucose reading from continuous glucose monitors, can improved predictive performance by capturing temprary trends. Long Short-Term Memory (LSTM) network and other recurrent neural networks (RNNs) can be merged with the current hybrid framework to model temporal dependencies, enabling early detection of changes in diabetes risk over time [42].

7.3 Mobile and Wearable Health Applications

Hybrid model can be integrated in mobile applications and wearable health devices to provide real-time monitoring. Smart watches and smartphones can continuously gather vital parameters(data) like heart rate, activity levels, and blood glucose levels. These inputs can be combined in hybrid model to instantaneously provide risk scores, tailored advice, and warnings for high-risk individuals. Such a method augments accessibility for rural and underserved populations and promotes preventive healthcare habits [39], [40].

7.4 Personalized Suggestions and Lifestyle Intervention

Future models can be furthered not only to forecast diabetes risk but also to make customized lifestyle and dietary suggestions. Algorithms for machine learning can examine diets, activity patterns, and medical histories of patients to recommend interventions and suggestions that can help mitigate risk. Systems can assist patients in living healthy lifestyles and avoiding transition from pre-diabetes to diabetes [43].

7.5 Multi-Modal and Multi-Omics Data Integration

Recent studies show that genetic, epigenetic, and metabolomic information can enhance prediction of disease when integrated with clinical characteristics [44]. Merging multi-omics datasets with the hybrid model has the potential to make more accurate and personalized predictions so clinicians can personalize preventive and treatment options according to a patient's own biological signature..

7.6 Policy and Public Health Implementation

Population-level potential uses exist for hybrid model. Government and health care organizations can use AI-based screening systems to identify high-risk populations, effectively allocate resources, and implement targeted intervention programs. These systems can scale up to regional or national levels using cloud-based infrastructure, privacy protections on data, and interface with existing electronic health record systems [40], [45].

7.7 Continuous Learning and Model Updates:

Finally, persistent research can involve continuous learning mechanisms by which the model learns to update predictions based on newly available patient data. This allow the system to be able to change along with population health trends and retain accuracy in the long run. Continuous learning architectures also have the ability to include clinician and patient feedback to refine the predictions to make them more accurate as well as increase real-world validity [42].

8. CONCLUSION

This study demonstrates the effectiveness of a human-inerpretable hybrid Logistic Regression–Random Forest (LR–RF) model for predicting diabetes risk. Merging the interpretability of Logistic Regression with the high accuracy of Random Forest, the hybrid framework achieve a balanced, reliable, and clinically meaningful prediction system. Recursive Feature Elimination (RFE) ensure that only the most relevant feature-such as glucose level, BMI, age, insulin, and blood pressure-are used, enhancing both performance and transparency [14], [15].

Experimental results showing the hybrid model outperforms standalone models, achieving 89.2% accuracy and an ROC-AUC of 0.91. High precision and recall values indicate that the model effectively minimize false positives and false negatives, which is crucial for clinical decision-making. Comparison to previous studies using the same dataset, the hybrid approach improve predictive performance by 3.7–13.2%, demonstrating its practical utility for real-world healthcare applications [18], [24].

Hybrid LR–RF model can be deployed in hospitals, telemedicine platforms, mobile apps, and wearable devices. It can help clinicians make early diagnoses, deliver individualized advice, and aid population-level screening programs. Moreover, its transparency enables healthcare workers to have confidence in the output of the system, hence acceptability in clinical practice [25], [39]

Future developments to the model can involve incorporation of heterogeneous demographic information, time-series prediction with the help of LSTM networks, multi-omics integration, and online learning processes. These enhancements further improve prediction accuracy and enable personalized healthcare solutions, particularly in rural or underserved regions. By enabling early intervention and preventive care, this hybrid model has the potential to reduce diabetes-related complications, lower healthcare costs, and improve patient quality of life [41]–[45].

The proposed hybrid LR–RF framework delivers a robust, interpretable, and scalable approach to diabetes prediction This highlights the transformative potential of data-driven AI systems in preventive healthcare and evidence-based medical decision-making.

REFERENCES

[1] World Health Organization, “Diabetes,” WHO, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes

[2] International Diabetes Federation, “IDF Diabetes Atlas, 10th Edition,” 2021. [Online]. Available: https://diabetesatlas.org/

[3] Ministry of Health and Family Welfare (MoHFW), Government of India, “National Health Profile 2022,” 2022. [Online]. Available: https://www.mohfw.gov.in/

[4] American Diabetes Association, “Standards of Medical Care in Diabetes 2023,” Diabetes Care, vol. 46, Suppl. 1, pp. S1–S152, 2023.

[5] Kumar, A., et al., “Machine Learning Techniques for Diabetes Prediction,” Int. J. Comput. Appl., vol. 175, no. 15, pp. 25–31, 2020.

[6] Hosmer, D.W., Lemeshow, S., and Sturdivant, R.X., *Applied Logistic Regression*, 3rd ed., Wiley, 2013.

[7] Breiman, L., “Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[8] Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., and Johannes, R.S., “Using the Pima Indians Diabetes Dataset,” National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), 1988. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes

[9] Suri, R., et al., “Diabetes Prediction Using Support Vector Machines,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 6, pp. 100–108, 2019.

[10] Centers for Disease Control and Prevention (CDC), “National Diabetes Statistics Report 2022,” U.S. Department of Health and Human Services, 2022. [Online]. Available: https://www.cdc.gov/diabetes/data/statistics-report/index.html

[11] Little, R.J.A., and Rubin, D.B., *Statistical Analysis with Missing Data*, 3rd ed., Wiley, 2019.

[12] Pedregosa, F., et al., “Scikit-learn: Machine Learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[13] Menard, S., *Logistic Regression: From Introductory to Advanced Concepts and Applications*, Sage Publications, 2011.

[14] Guyon, I., and Elisseeff, A., "An Introduction to Variable and Feature Selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.
[15] Das, S., et al., "Feature Selection for Diabetes Prediction Using RFE," *Healthcare Inform. Res.*, vol. 27, no. 2, pp. 98–106, 2021.
[16] Zhang, L., et al., "Hybrid Models for Chronic Disease Prediction," *IEEE Access*, vol. 9, pp. 13045–13056, 2021.
[17] Nguyen, T., et al., "Ensemble Learning for Cardiovascular Risk Prediction," *Comput. Biol. Med.*, vol. 123, 103888, 2020.
[18] Kaur, P., and Kumari, H., "Decision Tree Based Diabetes Diagnosis," *Int. J. Adv. Sci. Technol.*, vol. 29, no. 3, pp. 5430–5440, 2020.
[19] Powers, D.M.W., "Evaluation: From Precision, Recall, and F-Measure to ROC, Informedness, and Correlation," *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.
[20] Sharma, R., et al., "AI-Driven Health Applications for Diabetes Management," *Health Technol.*, vol. 13, pp. 233–246, 2023.
[21] CDC, "Prevalence of Diabetes and Prediabetes in the United States, 2022," U.S. Department of Health and Human Services, 2022.
[22] Smith, A., et al., "Combining Statistical and Ensemble Methods for Medical Diagnosis," *J. Healthc. Eng.*, vol. 2021, Article ID 6632105, 2021.
[23] WHO, "Global Report on Diabetes," World Health Organization, 2016. [Online]. Available: https://www.who.int/publications/i/item/9789241565257
[24] Das, A., et al., "Diabetes Prediction Using Logistic Regression and RFE," *Int. J. Med. Inform.*, vol. 144, 104301, 2020.
[25] Breiman, L., "Bagging Predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.
[26] Zhang, Y., et al., "Explainable AI in Healthcare: A Review," *IEEE Access*, vol. 8, pp. 813–829, 2020.
[27] Indian Council of Medical Research (ICMR), "National Urban Diabetes Survey Report," 2021.
[28] WHO, "Noncommunicable Diseases Country Profiles 2022," World Health Organization, 2022.
[29] WHO, "Diabetes Fact Sheet," 2023.
[30] International Diabetes Federation (IDF), "Global Diabetes Statistics," 2021.
[31] MoHFW, "National Health Profile of India," 2022.
[32] ICMR, "Prevalence of Diabetes in Indian Cities," 2021.
[33] CDC, "National Diabetes Statistics Report," 2022.
[34] CDC, "Prediabetes Overview," 2022.
[35] European Diabetes Prevention Program (EDPP), "Community-Based Diabetes Prevention," 2021.
[36] Smith, J.W., et al., "Pima Indians Diabetes Dataset," 1988.
[37] Das, S., et al., "Enhanced Diabetes Prediction with Additional Clinical Features," 2021.
[38] World Bank, "Global Health Data and Diabetes Mortality," 2022.
[39] Sharma, R., et al., "AI-Powered Mobile Diabetes Screening Applications," *Health Technol.*, 2023.
[40] WHO, "Digital Health Technologies for Diabetes Prevention," 2022.
[41] Nguyen, T., et al., "Multi-Demographic Machine Learning Models," 2020.
[42] Zhang, L., et al., "Time-Series Integration Using LSTM Networks for Disease Prediction," 2021.
[43] Kumar, A., et al., "Lifestyle-Based AI Interventions for Diabetes," 2020.
[44] Das, S., et al., "Multi-Omics Data Integration in Chronic Disease Prediction," 2021.
[45] Sharma, R., et al., "Population-Level AI Screening Programs," 2023.