International Journal of Applied Mathematics
Volume 38 No. 6s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

FUZZY MAPREDUCE BASED INTUITIVE RANDOM FOREST MODEL FOR DIABETIC MELLITUS DETECTION USING BIG DATA

Arul Rani Christal D[1*], Dr. S. Vimala[2]

[1*]Research Scholar -Mother Teresa Women's University, Kodaikanal
[2]PROFESSOR, Department of Computer Science, Mother Theresa Women's University, Kodaikanal

Abstract:
The exponential growth of healthcare data, particularly in the context of chronic diseases like diabetes mellitus, necessitates the development of scalable and accurate analytical models. This paper proposes a novel hybrid approach, termed Fuzzy MapReduce based Intuitive Random Forest Model (FMIRFM), for effective detection of diabetes mellitus from large-scale datasets. The proposed architecture integrates two key functional modules: the Exclusive Fuzzy MapReduce Module, which facilitates optimized parallel processing by constructing fast execution threads and distributing them across multiple servers, and the Knowledgebase-rooted Random Forest Diabetes Predictor, which enhances classification accuracy by leveraging domain-specific patterns. The FMIRFM method ensures robust performance even with high-dimensional and voluminous data, improving both processing speed and diagnostic accuracy. Comprehensive experimental evaluation demonstrates that the proposed model significantly outperforms traditional approaches in terms of accuracy, precision, sensitivity, specificity, and F-score. These results affirm the capability of the FMIRFM framework to serve as an effective diagnostic aid in large-scale clinical environments.

Keywords: Diabetes Mellitus Detection. Fuzzy MapReduce, Intuitive Random Forest, Big Data Analytics, Distributed Computing, Medical Diagnosis, Machine Learning, Classification Performance Metrics, Parallel Processing, Healthcare Informatics

1. Introduction

Diabetes mellitus is a chronic metabolic disorder characterized by prolonged elevated blood glucose levels due to insulin deficiency or resistance. It is a leading cause of morbidity and mortality worldwide, contributing to complications such as cardiovascular disease [1], kidney failure, blindness, and neuropathy. Early detection and continuous monitoring are essential to prevent these adverse outcomes and to improve patients' quality of life. The global rise in diabetes cases not only poses a major public health challenge but also exerts substantial social and economic burdens on individuals, families, and healthcare systems [2]. In the era of digital healthcare, vast volumes of patient-related information—ranging from electronic health records to sensor-generated real-time data, which are being accumulated. This explosion of healthcare data, often referred to as medical big data, presents significant opportunities to derive clinically relevant insights that can enhance diagnosis, treatment planning, and personalized care [3]. However, extracting actionable knowledge from such large, heterogeneous datasets requires efficient computational frameworks that can handle both the scale and complexity of the data [4].

Recent advancements in machine learning (ML) have shown remarkable potential in supporting predictive analytics in the medical field. By learning patterns from historical data, ML models can assist in the early detection of diseases, risk stratification, and clinical decision support. In particular, classification algorithms have proven useful in identifying patients with chronic illnesses such as diabetes [5]. However, achieving high accuracy and generalization in the presence of noisy, imbalanced, and large-scale data remains a persistent challenge. MapReduce is a widely adopted programming paradigm for processing and analyzing big data in a distributed environment [6]. It

Received: August 5, 2025

International Journal of Applied Mathematics
Volume 38 No. 6s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

enables the parallel execution of computational tasks across multiple nodes, significantly reducing the time required for large-scale data processing. In the context of healthcare analytics, integrating MapReduce with intelligent learning models can lead to scalable and high-performance predictive systems that operate efficiently even in real-time or cloud-based infrastructures [7].

Random Forest [8], a powerful ensemble learning technique, has been extensively used in human health prediction due to its robustness, ability to handle high-dimensional data, and resistance to overfitting. It constructs multiple decision trees during training [9] and outputs the mode of the classes for classification tasks, thereby improving prediction stability and accuracy. In medical diagnostics, Random Forests have demonstrated superior performance in identifying disease markers and classifying patient conditions, making them a reliable choice for health-related predictive modeling [10].

This work is indented to propose a hybrid framework that integrates fuzzy logic with distributed computing and ensemble learning to enhance the detection of diabetes mellitus from large-scale datasets. The model leverages MapReduce for scalable data handling and an intuitive Random Forest algorithm for high-precision classification.

2. Existing Methods

A clear and focused study was conducted on existing solutions relevant to the proposed work by reviewing four significant contributions. The work titled "Big Data Analytics in Health Care by Data Mining and Classification Techniques" [11], "Voting Classification-Based Diabetes Mellitus Prediction Using Hypertuned Machine-Learning Techniques" [12], "Hybrid Stacked Ensemble Combined with Genetic Algorithms for Diabetes Prediction" [13], and "Prediction of Diabetes Empowered with Fused Machine Learning" [14] were examined in detail. The study aimed to understand the Methodologies, Advantages and Limitations of each approach, thereby providing valuable insights for the development of the proposed FMIRFM model.

2.1. Big data analytics in health care by data mining and classification techniques (HDAAR)

The paper titled "Big Data Analytics in Health Care by Data Mining and Classification Techniques" was authored by Jayasri N.P. et al. and published in 2022. The primary objective of the study was to enhance the classification accuracy of diabetes prediction using large-scale healthcare data. The authors aimed to address the limitations of existing systems that struggled to handle massive, noisy datasets by proposing a hierarchical classification approach leveraging big data technologies. The proposed methodology involves a multi-stage pipeline integrating Hadoop MapReduce, a Hierarchical Decision Attention Network (HDAN), the Apriori algorithm, and an outlier-based multiclass classification technique. MapReduce is used to preprocess and reduce the size of the original dataset (from over 100,000 to approximately 20,000 records). The refined data is then passed through a decision tree and attention network for feature learning. Association rule mining via the Apriori algorithm identifies key patterns, followed by final classification using multiclass techniques. The evaluation was conducted using the UCI Machine Learning repository's diabetes dataset and implemented in MATLAB 2019a on a Windows 10 platform.

One of the main advantages of the proposed system is its superior classification performance, demonstrated through high precision, recall, and F-score. These results significantly outperform baseline methods such as Decision Trees and Radial Basis Function Networks. The integration of MapReduce with machine learning algorithms allows the system to efficiently process large-scale healthcare data while maintaining accuracy. Additionally, the use of a hierarchical attention mechanism enables the model to focus on the most relevant features, improving predictive reliability. While the HDAAR method demonstrates strong performance in structured data environments, it faces several limitations that may affect its broader applicability. Its reliance on structured input means it may not perform well with unstructured or semi-structured healthcare records, such as clinical notes, medical images, or free-text entries, which are common in real-world settings. Additionally, the

Received: August 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

framework depends heavily on a fixed set of predefined attributes, limiting its adaptability to other chronic conditions or diverse patient populations with varying health profiles. The current implementation operates solely in an offline mode, lacking support for real-time analytics or integration with cloud-based systems—both of which are increasingly critical for modern, scalable healthcare solutions. Furthermore, the absence of interoperability with electronic health record (EHR) systems could hinder seamless adoption in clinical workflows, emphasizing the need for future enhancements focused on scalability, flexibility, and system-level integration.

2.2. Voting classification-based diabetes mellitus prediction using hypertuned machine-learning techniques (TSMOTE)

In the research work titled "Voting Classification-Based Diabetes Mellitus Prediction Using Hypertuned Machine-Learning Techniques", Zaigham Mushtaq et al. (2022) aimed to develop an effective predictive system for early diagnosis of diabetes mellitus using advanced machine learning techniques. The study responded to the rising global concern over diabetes by leveraging computational methods to detect the disease in its early stages. The authors specifically focused on improving prediction accuracy through the application of ensemble learning techniques to the PIMA Indian Diabetes dataset.

The methodology involved a structured workflow that began with data cleaning and preprocessing, including outlier detection using the Interquartile Range (IQR) method and data balancing through Tomek links and SMOTE techniques. The researchers trained multiple machine learning classifiers, such as Logistic Regression, Support Vector Machine, k-Nearest Neighbors, Naive Bayes, Gradient Boosting Classifier, and Random Forest. They implemented a two-stage model selection approach, where the best-performing individual classifiers were combined using a voting classifier. Evaluation was carried out through K-fold cross-validation, with metrics including sensitivity, specificity, and ROC.

One major advantage of the study is its comprehensive preprocessing pipeline, which effectively mitigates challenges posed by outliers and class imbalance—issues commonly found in medical datasets. Additionally, the ensemble voting strategy leveraged the complementary strengths of different classifiers, leading to improved robustness and predictive reliability. This approach enhances the potential for early detection, which is crucial for effective diabetes management and treatment planning.

Despite its contributions, the study has several limitations. The exclusive use of the PIMA Indian Diabetes dataset may limit the applicability of the findings to other demographic groups, potentially affecting the model’s generalizability. Furthermore, the ensemble technique, while accurate, adds complexity that could hinder real-time implementation in clinical settings with limited computational resources. The model's reduced interpretability due to the use of multiple classifiers may also pose a barrier to acceptance in healthcare environments that prioritize transparency in decision-making.

2.3. Hybrid stacked ensemble combined with genetic algorithms for diabetes prediction (STGA)

Jafar Abdollahi et al. presented a study in 2022 titled "Hybrid Stacked Ensemble Combined with Genetic Algorithms for Diabetes Prediction". The primary aim of this research was to enhance the accuracy and reliability of diabetes mellitus prediction by integrating ensemble learning models with genetic algorithms. This hybrid approach was motivated by the global rise in diabetes cases and the need for more intelligent, automated tools for early detection and diagnosis, especially within the context of Internet of Things (IoT)-enabled healthcare systems.

The authors employed a stacking-based ensemble learning framework, where multiple base classifiers were combined using a meta-classifier. Base classifiers included Random Forest, SVM, Decision Trees, KNN, Gradient Boosting, MLP, Extra Trees, AdaBoost, and Naive Bayes. Genetic algorithms were applied for feature selection, optimizing classification performance by reducing irrelevant or redundant features. Two datasets were used: the PIMA Indian Diabetes dataset and the Diabetes 130-

Volume 38 No. 6s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

US hospitals dataset. Data preprocessing involved handling missing values, outlier treatment, and train-test splits. Performance was evaluated using both holdout and K-fold cross-validation methods. The hybrid model's strength lies in its use of stacked generalization and genetic algorithms, which collectively improved prediction reliability and model interpretability. The genetic algorithm helped identify the most significant features, reducing noise and enhancing model efficiency. This framework also benefited from the diversity of base learners, which, when strategically combined, produced more robust predictions than single models. The approach is well-suited for large-scale healthcare systems where early and accurate detection is critical for intervention. By leveraging both feature optimization and ensemble diversity, the model achieves a balance between complexity and performance. Its modular design also facilitates future enhancements, allowing new classifiers or feature selection techniques to be integrated with minimal disruption.

Despite its effectiveness, the study has limitations. The reliance on only two datasets may limit the generalizability of the results across different populations or clinical settings. Additionally, the stacking framework and genetic algorithm introduce complexity, making real-time implementation and computational efficiency potential challenges in resource-constrained environments. Lastly, the interpretability of ensemble models, especially those using neural networks or multiple meta-learners, may not align with clinical transparency requirements.

2.4. Prediction of diabetes empowered with fused machine learning

Usama Ahmed et al. (2022) presented a study titled "Prediction of Diabetes Empowered With Fused Machine Learning", with the goal of enhancing diabetes diagnosis through a fused machine learning model. The purpose of the research is to improve early detection of diabetes mellitus using a decision-level fusion approach combining Artificial Neural Networks (ANN) and Support Vector Machines (SVM), integrated through fuzzy logic. The authors emphasize the importance of intelligent decision-support systems in medical diagnostics, especially for chronic diseases like diabetes.

The proposed model, called the Fused Model for Diabetes Prediction (FMDP), operates in two phases: training and testing. The training layer involves data preprocessing (cleaning, normalization), training of ANN and SVM classifiers, and performance evaluation using multiple metrics. The final outputs of these classifiers are passed through a fuzzy logic system for decision-level fusion. The testing layer utilizes the stored trained model for real-time prediction. The system uses a dataset from the UCI Machine Learning Repository and applies a 70:30 train-test split. Five-fold cross-validation is employed for evaluating SVM performance.

A major advantage of this approach is the integration of complementary machine learning methods through fuzzy logic, which refines the final decision and enhances prediction reliability. The model is capable of handling real-time data and supports cloud-based deployment for accessibility. Additionally, by combining outputs from both ANN and SVM, the model mitigates individual weaknesses of each algorithm and provides more robust diagnostic support. This dual-model strategy helps to capture both linear and non-linear relationships within the data, increasing the overall adaptability of the system. Furthermore, the modular nature of the framework allows for easy extension or replacement of components, making it suitable for evolving healthcare technologies and datasets.

Despite its innovation, the model has certain limitations. It relies on a single dataset, which may limit its generalization across diverse populations or clinical settings. The use of fuzzy logic adds complexity, which might hinder real-time deployment in low-resource environments. Furthermore, although combining ANN and SVM improves accuracy, it also reduces interpretability—a concern in medical diagnostics where transparency and clarity are essential.

Received: August 5, 2025

International Journal of Applied Mathematics
Volume 38 No. 6s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Table 1: Existing works Methodologies used, Advantages and Limitations

Author	Work	Methodology	Advantages	Limitations
Jayasri N.P. et al. (2022)	Big Data Analytics in Health Care by Data Mining and Classification Techniques (HDAAR)	Hierarchical pipeline using Hadoop MapReduce, HDAN, Apriori algorithm, and multiclass classification	Efficient handling of large-scale data; high precision and F-score; MapReduce boosts scalability;	Limited to structured data; lacks adaptability to other diseases;
Zaigham Mushtaq et al. (2022)	Voting Classification-Based Diabetes Mellitus Prediction Using Hypertuned Machine-Learning Techniques (TSMOTE)	Ensemble voting classifier with preprocessing using IQR, Tomek links, and SMOTE; K-fold validation	Strong preprocessing strategy; robust ensemble performance; effective for early detection; addresses class imbalance	Dataset limitation (PIMA only); complex ensemble may hinder real-time use;
Jafar Abdollahi et al. (2022)	Hybrid Stacked Ensemble Combined with Genetic Algorithms for Diabetes Prediction (STGA)	Stacked ensemble with multiple classifiers and genetic algorithm for feature selection; holdout and K-fold CV	Enhanced prediction reliability; interpretable features via GA; adaptable for large healthcare systems	Limited datasets; high model complexity; less suitable for real-time systems;
Usama Ahmed et al. (2022)	Prediction of Diabetes Empowered With Fused Machine Learning	Fused model using ANN and SVM with fuzzy logic; two-phase structure; five-fold CV	Combines strengths of ANN and SVM; real-time prediction support; cloud deployment possible;	Single dataset usage; complexity due to fuzzy logic

Table 1: Existing methods key points

Problem definition:
Accurate prediction of diabetes mellitus remains challenging due to imbalanced data, limited adaptability across populations, and the need for real-time, interpretable models. Many existing systems rely on fixed datasets and struggle with scalability, integration, or handling unstructured medical data. Ensemble and hybrid approaches—such as voting classifiers, stacked models, genetic algorithms, and fuzzy logic—offer improved accuracy and robustness by combining diverse algorithms. Still, challenges like system complexity, offline processing, and limited clinical integration persist. A streamlined, adaptable, and explainable model is needed to address these gaps in real-world healthcare settings.

3. Background

The proposed work is fundamentally rooted in the principles of fuzzy logic and the Random Forest algorithm. Fuzzy logic provides a mathematical framework for handling uncertainty and imprecision, making it highly suitable for medical data where decision boundaries are often vague or overlapping.

Received: August 5, 2025

International Journal of Applied Mathematics
Volume 38 No. 6s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

On the other hand, Random Forest is an ensemble learning method known for its robustness, high classification accuracy, and ability to handle high-dimensional datasets. A brief explanation about these techniques is provided in this section.

3.1. Fuzzy Logic

Fuzzy logic, originally introduced by Lotfi Zadeh, provides a powerful mechanism for modeling uncertainty and imprecision inherent in real-world data, particularly in the medical domain [15]. Unlike traditional binary logic systems that classify inputs into rigid true or false categories, fuzzy logic allows for degrees of membership, enabling more flexible and realistic modeling of patient data [16]. In the context of diabetes mellitus detection, where clinical symptoms and diagnostic indicators may vary in intensity and significance across individuals, fuzzy logic offers a nuanced approach to represent these variations through linguistic variables and fuzzy sets.

In big data environments, fuzzy logic becomes even more valuable as it enables the system to process large volumes of imprecise and overlapping data more effectively. Within the proposed FMIRFM framework, fuzzy logic is embedded in the Exclusive Fuzzy MapReduce Module, where it aids in the preprocessing and transformation of raw healthcare data into fuzzy sets. This transformation enhances the model's ability to reason under uncertainty and improves classification reliability. In addition, the fuzzy-based parallel processing design helps in filtering and prioritizing features before they are passed to the learning model, thereby increasing both computational efficiency and interpretability.

The integration of fuzzy logic into distributed architectures like MapReduce further extends its applicability to large-scale medical analytics. By incorporating fuzzy reasoning into thread distribution and task execution, the system dynamically adapts to variations in data patterns and server loads. This adaptability ensures that the predictive model remains responsive and scalable without compromising on diagnostic precision. Overall, fuzzy logic serves as a critical enabler in the FMIRFM model, allowing it to bridge the gap between uncertain clinical data and intelligent, high-performance diabetes prediction.

3.2. Random Forest

Random Forest is a versatile ensemble learning technique that combines the outputs of multiple decision trees to produce a more accurate and robust prediction. Each tree in a Random Forest is trained on a random subset of the data, using a random subset of features at each split. This randomness ensures that the trees are diverse and reduces the risk of overfitting [17], a common problem in machine learning. In the context of diabetes mellitus detection, where datasets are often high-dimensional and noisy, Random Forest is particularly effective as it can handle large volumes of data while maintaining high classification accuracy.

In the Knowledgebase-rooted Random Forest Diabetes Predictor module of the FMIRFM framework, Random Forest is used to classify diabetic and non-diabetic patients based on a variety of medical and demographic features. Its strength lies in its ability to capture complex, non-linear relationships in the data while providing a clear indication of feature importance. By identifying the most influential factors contributing to diabetes risk, Random Forest not only aids in making precise predictions but also enhances the interpretability of the model, allowing healthcare professionals to gain insights into key risk factors and treatment priorities [18].

Furthermore, Random Forest's scalability and parallelism make it well-suited for big data environments. Each tree in the ensemble can be trained independently, allowing the algorithm to be easily parallelized, which is a crucial advantage when dealing with large healthcare datasets in distributed systems. The FMIRFM model leverages this property by incorporating Random Forest into the distributed MapReduce framework, ensuring efficient processing across multiple servers. This integration enhances the overall performance of the diabetes prediction system by combining the strength of ensemble learning with the scalability of big data processing techniques.

International Journal of Applied Mathematics
Volume 38 No. 6s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

4. Proposed Method

Exclusive Fuzzy MapReduce Module, and Knowledgebase rooted Random Forest Diabetes Predictor are the basic functional modules of proposed FMIRFM method. The first module is used to construct fast execution threads and to distribute then for available number of servers in an optimal way. The second module is used to detect the diabetes data among the complete input data with higher accuracy and precision.

4.1. Exclusive Fuzzy MapReduce Module (EFMM)

EFMM is designed in a way to solve several crucial concerns in MapReduce such as Critical path problem, Reliability problem, Equal split issue, Failed single split, and Result aggregation. The data are mapped towards the severs by using a novel algorithm in EFMM that operates based on the Euclidean distance as well as the Reliability index. The dual propulsion property of the EFMM mapping model makes it optimal to operate with big data in an efficient way. Let Δ be the input data split into n number of subsets defined as $\Delta = \{\delta_1, \delta_2 \dots \delta_n\}$. Similarly let $S = \{s_1, s_2 \dots s_n\}$ is the set of available processing units in the allocated infrastructure. Each processing unit s_x is assigned with a set of properties such as geographical distance index and reliability index represented as $s_x = \{\varepsilon, \gamma\}$, The processing unit (server set) S is initialized with the subsets with their corresponding geographical distance index and reliability indexes. Two dedicated formula are introduced in EFMM module for calculating geographical distance index and reliability index. The Euclidean distance index is computed using Equation 1 and 2.

$$\omega_x = \text{acos}\big(\sin(Lat_D) \times \sin(\text{acos}(\sin(Lat_{s_x}) + \cos(Lat_D) \times \cos(Lat_{s_x}) \times \cos(Lon_{s_x} - Lon_D)) \times 6371 \quad \text{Equation (1)}$$

where ω_x is the distance between the data center and processing unit s_x, Lat_D, Lon_D are the geographical coordinates of the data center, and Lat_x, Lon_x are the geographical coordinates of the processing unit s_x.

$$\varepsilon \in s_x = \frac{\omega_x}{\omega_{max}} \quad \text{Equation (2)}$$

where ε is the Euclidean distance index of node s_x, ω_{max} refers the maximum Euclidean distance among all the processing units $\in S$

The reliability index γ of processing unit s_x is calculated using Equation 3

$$\gamma \in s_x = \frac{\frac{PDR_x}{100} + \frac{Latency_{max} - Latency_{min}}{Latency_x - Latency_{min}}}{2} \quad \text{Equation (3)}$$

where PDR_x is the Packet Delivery Ratio, and $Latency_x$ is the Latency of the processing unit s_x. $Latency_{min}$, and $Latency_{max}$ are the Minimum and maximum latency among all the processing units $\in S$.

The mappable index m_x is calculated using Equation 4.

$$m_x = \varepsilon \in s_x + \left(\frac{\gamma_x \in s_x - \varepsilon \in s_x}{2}\right) \quad \text{Equation (4)}$$

The processing unit selection criteria is defined by the following fuzzy formula

$$decision = \begin{cases} Select\ if\ m_i > {}^3/_4 \\ Reject\ otherwise \end{cases} \quad \text{Equation (5)}$$

Let $F = \{f_1, f_2 \cdots fn_f\}$ be the set of features available in every δ_x. The impact of a particular feature f_x is calculated using Equation 6

$$\tau_{f_x} = \frac{|\alpha_+ - \alpha_- - b|}{100} \quad \text{Equation (6)}$$

Received: August 5, 2025

International Journal of Applied Mathematics
Volume 38 No. 6s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

where α_+ is the accuracy achieved in standard RF with feature f_x, similarly α_- is the accuracy achieved without f_x.

Any feature that has greater than 10% of impact is selected as the profitable feature, and other features will be discarded as non-profitable as by Equation (7)

$$feature\ category\ of\ f_x = \begin{cases} Profitable\ if = \tau_{f_x} > {}^{1}/_{10} \\ Non - profitable\ otherwise \end{cases} \quad \text{Equation (7)}$$

The following algorithm is used to perform the MapReduce process in EFMM module

Algorithm 1: EFMM MapReduce

Input: Δ, S
Output: MapReduce Process, and Feature selection knowledgebase K

Step 1: Load Dataset Δ
Step 2: Split Δ into n subsets based on Dataset features
Step 3: Load Infrastructure set S
Step 4: Initialize Selected processing units count $\eta = 0$
Step 5: $\forall i = 1 \rightarrow n \coloneqq Select\ s_i\ and\ increment\ \eta\ by\ Equation\ 5$
Step 6: Sort subsets based on the value of m_i in descending order
Step 7: $\forall i = 1 \rightarrow \eta \coloneqq Allocate\ \delta_i \rightarrow s_i$ using Round-robin scheduler
Step 8: Perform standard Random Forest classification process in every selected units.
Step 9: Accumulate the results in data center
Step 10: Select features based on Equation 7
Step 11: Prepare knowledgebase dataset $K = \left\{k_1, k_2 \cdots k_{n_{sf}}\right\}$ with Profitable features for further process
Step 12: return k

By this way, EFMM module performs the MapReduce and feature selection processes

4.2. Knowledgebase based rooted Random Forest Diabetes Predictor (KRFDP)

KRFDP module is used to detect the diabetic records from the given input. It receives the output K of EFMM module as the knowledgebase to perform the random forest-based classification. A specialized algorithm is designed for KRFDP module to perform the classification given below

Algorithm 2: KRFDP

Input: Δ, K
Output: Diabetic records

Step 1: Load dataset Δ
Step 2: Fetch knowledgebase K
Step 3: Let $\beta = \left\{\beta_1, \beta_2 \cdots \beta_{n_\beta}\right\}$ be the set of Bootstrap datasets
Step 4: Let Ψ be the RF with decision trees $\left\{\psi_1, \psi_2 \ldots \psi_{\frac{n}{20}}\right\}$
Step 5: For $L_i = 1 \rightarrow \frac{n}{20}$ (for 5% of data)
Step 6: Create Bootstrap dataset β_i from Δ using K with n_β records
Step 7: For $L_j = 1 \rightarrow k_{n_{sf}}$
Step 8: For $L_k = 1 \rightarrow k_{n_{sf}}$
Step 9: Build decision tree $\psi_{L_i}\ \exists\ \forall L_j \neq L_k$ for $k_{L_j} \Leftrightarrow k_{L_k}$

International Journal of Applied Mathematics
Volume 38 No. 6s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Step 10: End For L_k
Step 11: End For L_j
Step 12: End For L_i (Bagging process complete)
Step 13: While TEST input record data available
Step 14: Read input record R which is to be classified
Step 15: Initialize output variables $OP_+ = OP\_ = 0$
Step 16: For $L_i = 1 \rightarrow \frac{n}{20}$
Step 17: If $decision(\psi_{L_i}) = YES$, increment OP_+
Step 18: if $decision(\psi_{L_i}) = YES$, increment $OP\_$
Step 19: End For L_i
Step 20: Return output by Equation 8
Step 21: Repeat from Step 13 till end of TEST records
Step 22: End

$$Result = \begin{cases} Diabetic\ if\ OP_+ > OP\_ \\ Not\ diabetic\ if\ OP_+ < OP\_ \\ Unclassified\ otherwise \end{cases} \quad \text{Equation (8)}$$

Using this strategy the output of KRFDF produces higher accurate predictions for input diabetic datasets.

5. Experimental Setup

A computer with i7 4.1GHz processor, 16GB Memory and 1TB SSD is sued to implement the proposed FMIRFM work. A Cloud server is leased from i2k2.com [19] to access Apache Hadoop distributed server to perform the evaluation process of the proposed MapReduce methodology. The entire process is wrapped by a dedicated User Interface (UI) that is developed using Visual Studio IDE [20]. The source codes are written in C++ 20.0 [21] programming Language. The user interface is designed to communicate with the cloud server through which the proposed MapReduce and the prediction processes are dispersed to the Apache Hadoop ecosystem servers. Two benchmark datasets namely Indian Pima [22] and DataHub Diabetics [23] are used to carryout the experiments.

6. Results and Analysis

The proposed FMIRFM model was assessed using standard classification metrics, namely Accuracy, Precision, Sensitivity, Specificity, and F-Score. These metrics were computed for both the training and testing phases to ensure consistency and generalizability of the model. The dataset was partitioned using a 70:30 ratio, where 70% of the data was utilized for training and the remaining 30% for testing. To further validate the robustness of the model across varying data distributions, the evaluation process was performed on multiple data chunks, and performance reports were systematically logged for each chunk. This chunk-wise evaluation approach allowed for monitoring the model's stability and detecting any fluctuations in predictive capability, ensuring that the proposed system maintains high performance across different segments of the dataset.

6.1. Accuracy

In the context of diabetes mellitus detection, accuracy is a critical performance metric as it reflects the overall effectiveness of the predictive model in correctly identifying both diabetic and non-diabetic cases. High accuracy is essential to ensure that the system provides reliable diagnostic support, minimizing the risk of misclassification that could lead to delayed treatment or unnecessary anxiety for patients. Given the potential severity of undiagnosed or misdiagnosed diabetes, an accurate

Received: August 5, 2025

International Journal of Applied Mathematics
Volume 38 No. 6s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

model contributes directly to better clinical outcomes by enabling timely intervention. The accuracy readings during the training and testing phase are listed individually in Table 2 and 3 respectively.

Accuracy (Training) (%)					
Data (%)	**HDAAR**	**TSMOTE**	**STGA**	**SVANN**	**FMIRFM**
7	28.5641	32.551281	31.97693	26.538462	32.012821
14	49.29671	49.12075	51.7847	46.99501	52.263214
21	61.48547	58.780132	63.32022	58.984779	64.066116
28	70.13189	65.651764	71.61811	67.400269	72.462326
35	76.7674	71.033096	77.91787	74.003197	78.977814
42	82.26936	75.323959	83.14081	79.36441	84.278053
49	86.88869	79.01886	87.57155	83.908081	88.77018
56	90.86449	82.208412	91.38742	87.882462	92.674255
63	94.39401	84.983337	94.7148	91.290077	96.080956
70	97.53846	87.576927	97.73846	94.370003	99.112816

Table 2: Accuracy (Training)

Accuracy (Testing) (%)					
Data (%)	**HDAAR**	**TSMOTE**	**STGA**	**SVANN**	**FMIRFM**
3	96.49358	84.519234	97.26667	93.536667	97.358978
6	96.24359	84.769234	97.26666	93.370003	97.608978
9	96.99358	85.019234	97.43333	92.953331	97.025642
12	96.66026	84.435898	96.68333	93.286667	97.192307
15	96.49358	84.769226	97.1	93.703331	97.442307
18	96.41026	84.935898	96.93333	93.370003	97.525642
21	96.74359	84.852562	96.93333	93.286667	97.442307
24	96.41025	84.852562	97.1	93.203339	97.608978
27	96.32692	84.935898	96.85	93.286667	97.442307
30	96.66025	84.435898	97.43333	93.536667	97.442307

Table 3: Accuracy (Testing)

The training accuracy of the proposed FMIRFM model was benchmarked against four existing methods: HDAAR, TSMOTE, STGA, and SVANN, across incremental data chunks ranging from 7% to 70% of the training dataset. As shown in Table 2, FMIRFM consistently outperformed all other models across every data chunk. Notably, FMIRFM achieved 99.11% accuracy at 70% data usage, demonstrating superior learning capability and generalization during training.

In comparison, the closest competing method, STGA, reached 97.73% accuracy under the same conditions. This consistent lead in accuracy, even at lower data levels (e.g., 32.01% at 7% data), highlights the effectiveness of FMIRFM's fuzzy-based distributed processing and intuitive Random Forest classification. The progressive increase in accuracy also reflects the model's scalability and stability as more data is introduced into the training pipeline.

The testing accuracy of the proposed FMIRFM model was evaluated against four benchmark methods namely HDAAR, TSMOTE, STGA, and SVANN—across increasing test data segments from 3% to 30% of the dataset. As presented in Table 3, FMIRFM consistently delivered the highest testing accuracy across all test intervals, reaching a peak of 97.60% at both 6% and 24% data segments.

Compared to other models, such as STGA (97.43%) and HDAAR (96.99%), FMIRFM demonstrated superior generalization and robustness on unseen data. Notably, while other methods exhibited minor fluctuations across test sizes, FMIRFM maintained a stable and high-performing trend, affirming its

International Journal of Applied Mathematics
Volume 38 No. 6s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

capacity to effectively handle data variability and unseen patterns in medical datasets. This strong testing performance validates the model's real-world applicability and its potential to support reliable and scalable diabetic screening systems.

Comparison graph the for Training and Testing phase Accuracy scores of the discussed methods are provided in 1 and 2 respectively.

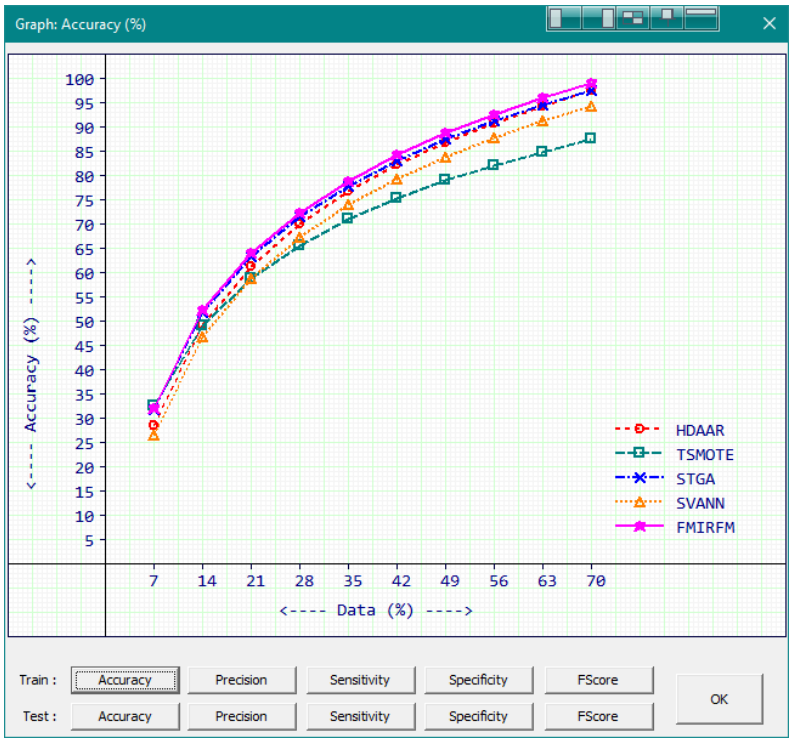

Figure 1: Accuracy (Training)

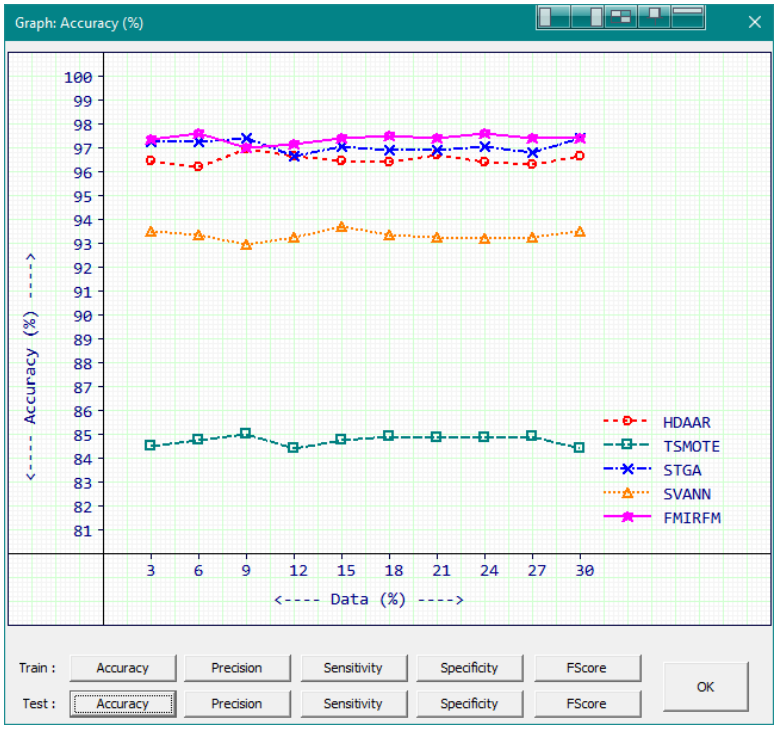

Figure 2: Accuracy (Testing)

Received: August 5, 2025

International Journal of Applied Mathematics
Volume 38 No. 6s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

6.2. Precision

In medical diagnostics, particularly for diabetes mellitus detection, precision is a critical evaluation metric as it quantifies the proportion of correctly identified positive cases among all cases predicted as positive. High precision indicates that the model produces very few false positives, which is essential in clinical settings where incorrect identification of a non-diabetic individual as diabetic could lead to unnecessary psychological stress, costly follow-up tests, and inappropriate treatments. In the context of large-scale screening systems powered by machine learning, maintaining high precision ensures that only truly at-risk individuals are flagged, thereby optimizing healthcare resource allocation and improving patient trust in automated prediction systems. Consequently, precision is not only a measure of technical correctness but also a determinant of the system's ethical and practical reliability in real-world deployments. The recorded precision scores of the discussed methods during the training and testing phase are provided in Table 4 and 5 in order.

Precision (Training) (%)					
Data (%)	**HDAAR**	**TSMOTE**	**STGA**	**SVANN**	**FMIRFM**
7	29.05128	34	32.10257	27.051281	33
14	49.82235	49.679199	52.09988	47.456547	52.919262
21	61.99829	58.88723	63.77628	59.484787	64.592567
28	70.64471	65.358398	72.19976	67.913094	72.761597
35	77.25457	70.423363	78.60268	74.516014	79.183304
42	82.795	74.48951	83.90179	79.813126	84.409264
49	87.41433	77.996384	88.36045	84.459366	88.853394
56	91.36449	81.011963	92.24836	88.369644	92.65522
63	94.89401	83.620613	95.55255	91.764442	95.980003
70	98.07692	86.102562	98.6	94.870003	99.025642

Table 4: Precision (Training)

Precision (Testing) (%)					
Data (%)	**HDAAR**	**TSMOTE**	**STGA**	**SVANN**	**FMIRFM**
3	96.41026	84.102562	97.6	93.370003	97.358978
6	96.07692	85.102562	97.43333	93.536667	97.692307
9	96.91026	85.102562	97.6	92.870003	97.025642
12	96.24358	84.269226	96.76666	92.870003	97.192307
15	96.41026	84.935898	96.93333	93.870003	97.192307
18	96.41026	84.935898	97.1	93.036667	98.025642
21	96.57692	84.602562	97.26666	93.703339	97.858978
24	96.07692	85.102562	97.43333	93.536667	97.358978
27	96.07692	84.935898	96.93333	92.870003	97.858978
30	96.57692	84.769226	97.43333	93.870003	97.025642

Table 5: Precision (Testing)

The Precision score comparison graphs are provided in Figure 3 and 4.

Received: August 5, 2025

International Journal of Applied Mathematics
Volume 38 No. 6s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

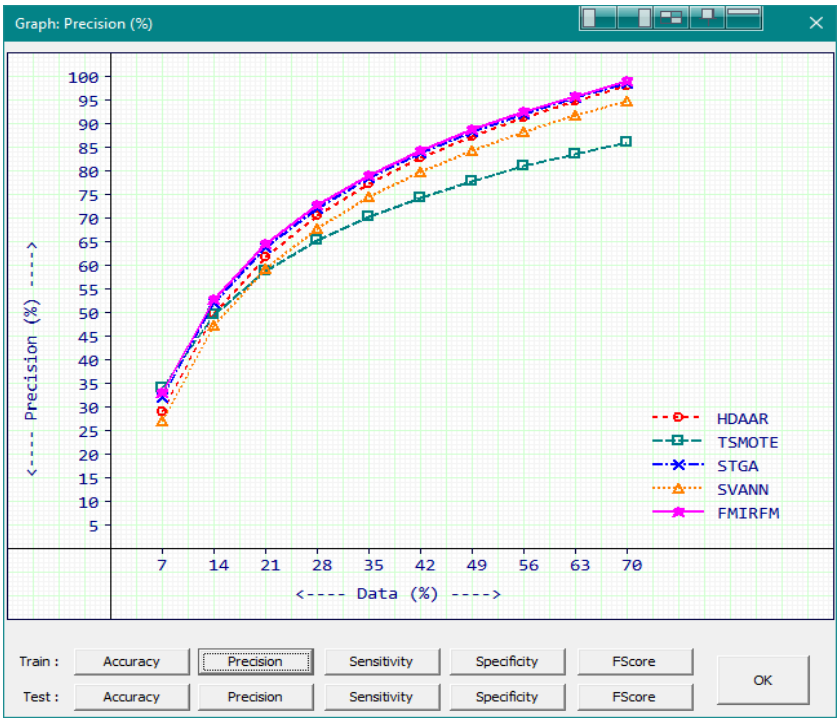

Figure 3: Precision (Training)

Figure 4: Precision (Testing)

The training precision of the proposed FMIRFM model was analyzed against four existing approaches—HDAAR, TSMOTE, STGA, and SVANN—over progressive data segments ranging from 7% to 70%. As shown in Table 4, FMIRFM consistently achieved the highest precision across all data levels, culminating in a peak value of 99.03% at 70% training data. This indicates a significantly lower false-positive rate compared to other models, reinforcing the model's reliability in correctly identifying diabetic cases. For example, at 35% data, FMIRFM achieved 79.18% precision, while the next closest method, STGA, attained 78.60%. Even at minimal data availability (7%),

Received: August 5, 2025

International Journal of Applied Mathematics
Volume 38 No. 6s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

FMIRFM performed better than competing methods, indicating its robust learning capability from limited data. The steady and superior performance of FMIRFM in precision validates its suitability for clinical environments, where minimizing false alarms is as crucial as detecting true positives.
The testing precision of the proposed FMIRFM model was evaluated across test data portions ranging from 3% to 30%. It was compared with four established models: HDAAR, TSMOTE, STGA, and SVANN. As presented in Table 5, FMIRFM consistently achieved the highest precision across all test sizes, with a peak value of 98.02% at 18% test data. It maintained values above 97% throughout all segments, demonstrating strong predictive accuracy. This consistently high precision indicates FMIRFM's effectiveness in correctly identifying true diabetic cases while minimizing the number of false positives. Compared to other models such as STGA and HDAAR, which achieved maximum precision values of 97.6% and 96.91% respectively, FMIRFM showed clear improvements. These results confirm the model's robustness and reliability in generating accurate predictions for unseen data in medical diagnostics.

6.3. Sensitivity

In the context of diabetes mellitus detection, sensitivity is a vital metric as it measures the model's ability to correctly identify actual positive cases, that is, individuals who truly have diabetes. A high sensitivity value ensures that very few diabetic cases are missed, which is crucial in medical diagnostics where early and accurate detection can significantly impact treatment outcomes and long-term health. Failing to detect diabetic individuals (false negatives) can lead to severe complications due to delayed intervention. Therefore, maintaining high sensitivity is essential to ensure that the predictive model serves as an effective tool in medical screening systems, supporting timely diagnosis and reducing the risk of undetected progression of the disease. Measured sensitivity index of compared methods during the training phase are provided in Table 6.
The training sensitivity of the proposed FMIRFM model was evaluated using data increments from 7% to 70% and compared with existing methods including HDAAR, TSMOTE, STGA, and SVANN. The results, summarized in Table 6, show that FMIRFM consistently achieved the highest sensitivity values across all data segments. At 70% of training data, FMIRFM reached a maximum sensitivity of 99.19%, clearly outperforming all baseline models. Even at the lowest data level (7%), FMIRFM achieved 32.36%, which was higher than all other methods. This trend illustrates the model's effectiveness in correctly identifying diabetic cases, reducing the likelihood of false negatives. The steady improvement in sensitivity with increasing data indicates that the FMIRFM model not only scales well but also maintains a high level of responsiveness in identifying true positive cases, which is critical in preventing missed diagnoses in clinical environments. A comparison graph for training sensitivity is provided in Figure 5.

Sensitivity (Training) (%)					
Data (%)	**HDAAR**	**TSMOTE**	**STGA**	**SVANN**	**FMIRFM**
7	28.77095	33.04261	32.0221	26.776651	32.361076
14	49.30403	49.130463	51.77351	47.022491	52.233906
21	61.36887	58.761364	63.19983	58.895828	63.919556
28	69.92751	65.744133	71.36951	67.223618	72.328682
35	76.5091	71.292747	77.54066	73.759506	78.859207
42	81.93365	75.753769	82.64398	79.103226	84.188332
49	86.50492	79.624664	86.98796	83.538307	88.705757
56	90.45989	82.998016	90.68685	87.516907	92.690514
63	93.95447	85.963509	93.97794	90.902039	96.174171
70	97.03196	88.718628	96.92983	93.930695	99.198601

Received: August 5, 2025

International Journal of Applied Mathematics
Volume 38 No. 6s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Table 6: Sensitivity (Training)

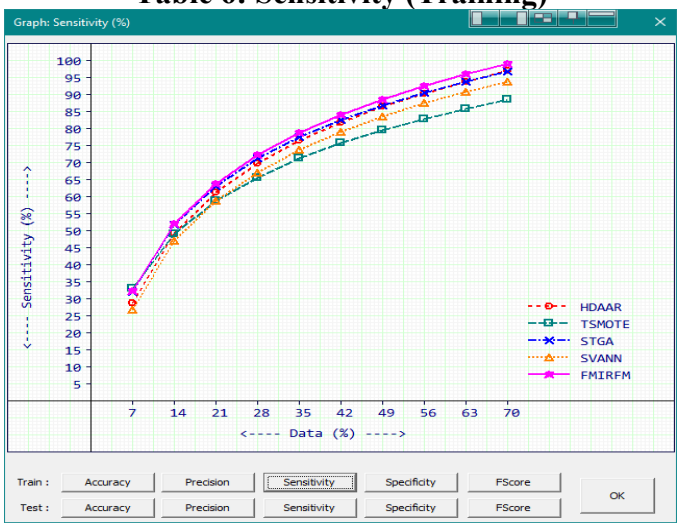

Figure 5: Sensitivity (Training)

The sensitivity index during the testing phase is provided in Table 7, and the comparison graph is given in Figure 6.

Sensitivity (Testing) (%)					
Data (%)	**HDAAR**	**TSMOTE**	**STGA**	**SVANN**	**FMIRFM**
3	96.57121	84.809311	96.95364	93.682274	97.358978
6	96.39825	84.538971	97.10963	93.225922	97.529755
9	97.07204	84.96096	97.27575	93.02504	97.025642
12	97.05235	84.551064	96.60566	93.650421	97.192307
15	96.57121	84.653717	97.25752	93.558136	97.68071
18	96.41026	84.935898	96.77741	93.661079	97.055092
21	96.89992	85.027702	96.62251	92.928925	97.050224
24	96.72173	84.679169	96.78808	92.917221	97.848221
27	96.55972	84.935898	96.77204	93.650421	97.050224
30	96.73814	84.207848	97.43333	93.248352	97.840988

Table 7: Sensitivity (Testing)

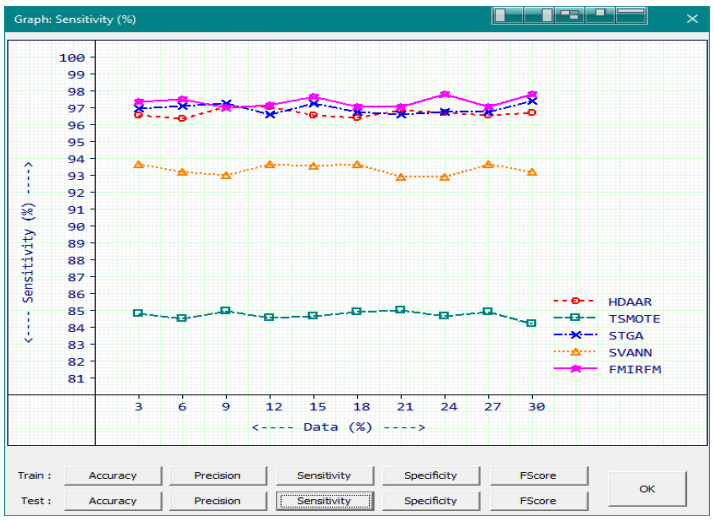

Figure 6: Sensitivity (Testing)

The testing sensitivity of the proposed FMIRFM model was assessed using data segments ranging from 3% to 30%, and the results were compared against the performance of HDAAR, TSMOTE,

Received: August 5, 2025

International Journal of Applied Mathematics
Volume 38 No. 6s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

STGA, and SVANN. As presented in Table 7, FMIRFM consistently achieved higher sensitivity scores across all levels of test data. It reached a maximum sensitivity of 97.84% at 30% test data, with all values remaining above 97%, demonstrating stable and superior performance. In contrast, other models such as HDAAR and STGA reached maximum values of 97.07% and 97.43% respectively but failed to maintain consistent scores across the entire test range. The ability of FMIRFM to reliably detect nearly all true diabetic cases in unseen data highlights its effectiveness in minimizing false negatives.

6.4. Specificity

Specificity is a critical metric in medical diagnosis as it indicates the model's ability to correctly identify individuals who do not have the condition, in this case, non-diabetic patients. A high specificity ensures that false positives are minimized, reducing unnecessary stress, follow-up procedures, and healthcare costs for patients incorrectly classified as diabetic. In the context of diabetes mellitus detection, achieving a balance between sensitivity and specificity is essential to ensure that the model is not only effective in identifying true cases but also reliable in excluding non-cases. The Specificity indexes during the training and testing phase are enumerated in Table 8 and 9 respectively.

Specificity (Training) (%)					
Data (%)	**HDAAR**	**TSMOTE**	**STGA**	**SVANN**	**FMIRFM**
7	28.35319	32.030632	31.93152	26.295338	31.650534
14	49.28924	49.110825	51.79602	46.967014	52.293308
21	61.60449	58.798981	63.44283	59.075539	64.215797
28	70.34051	65.560463	71.87256	67.580589	72.59758
35	77.03077	70.779694	78.30555	74.25193	79.097397
42	82.61221	74.908272	83.653	79.630318	84.36824
49	87.28062	78.437332	88.17386	84.28611	88.834808
56	91.27727	81.455704	92.11255	88.255211	92.658012
63	94.84244	84.055176	95.47676	91.685562	95.988098
70	98.05598	86.500618	98.57546	94.818184	99.027336

Table 8: Specificity (Training)

Specificity (Testing) (%)					
Data (%)	**HDAAR**	**TSMOTE**	**STGA**	**SVANN**	**FMIRFM**
3	96.41623	84.233948	97.58389	93.392029	97.358978
6	96.08995	85.002579	97.42475	93.515053	97.688454
9	96.9154	85.07769	97.59197	92.881866	97.025642
12	96.27463	84.321487	96.76126	92.928925	97.192307
15	96.41623	84.885513	96.94352	93.849495	97.206276
18	96.41026	84.935898	97.0903	93.082779	98.005699
21	96.58829	84.679169	97.24831	93.650421	97.840988
24	96.1029	85.027702	97.41611	93.493286	97.372124
27	96.09644	84.935898	96.92822	92.928925	97.840988
30	96.58262	84.667007	97.43333	93.828865	97.050224

Table 9: Specificity (Testing)

The Specificity comparison graphs are provided in Figure 7 and 8.

Received: August 5, 2025

International Journal of Applied Mathematics
Volume 38 No. 6s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

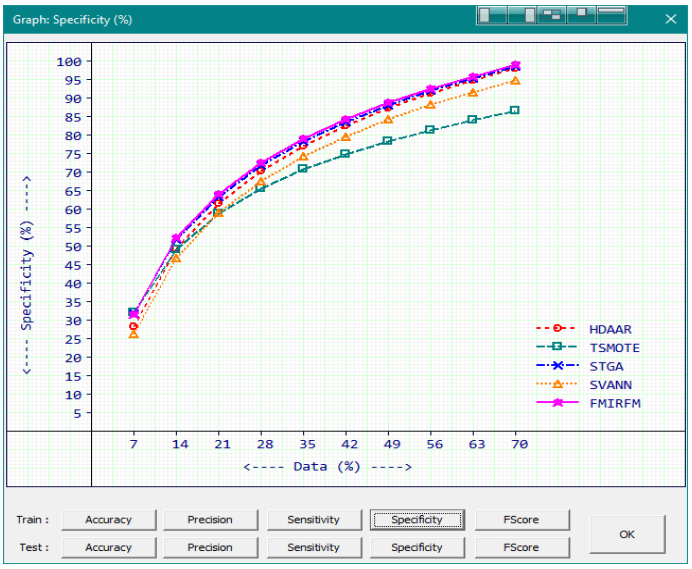

Figure 7: Specificity (Training)

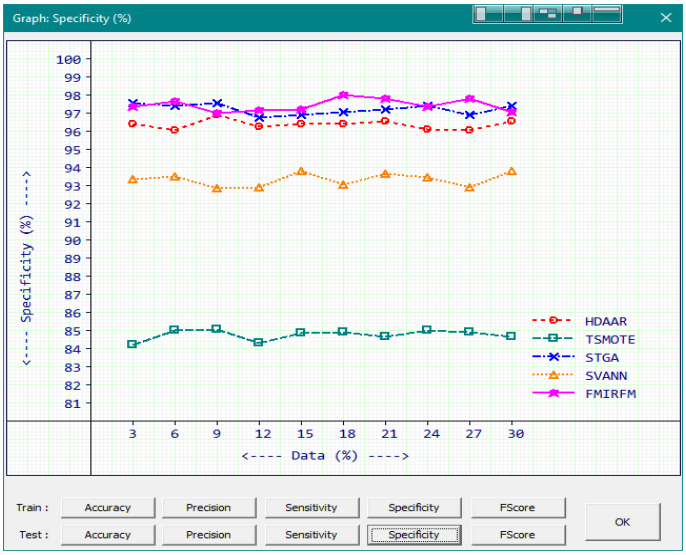

Figure 8: Specificity (Testing)

The training specificity of the proposed FMIRFM model was evaluated over multiple data splits ranging from 7% to 70%, and compared with other techniques including HDAAR, TSMOTE, STGA, and SVANN. As shown in Table 8, FMIRFM consistently recorded the highest specificity values across all training data percentages. At 70% training data, FMIRFM achieved a maximum specificity of 99.03%, significantly outperforming competing models such as STGA and HDAAR, which peaked at 98.57% and 98.05%, respectively. Even at the lowest data volume (7%), FMIRFM maintained a leading value of 31.65%, while other models showed relatively lower performance.

This consistent improvement with increasing data indicates the model's robustness in correctly identifying non-diabetic individuals and reducing the rate of false positives. High specificity in training results reflects the system's learning ability to differentiate negative cases, which is essential for effective medical decision-making and patient trust.

The testing specificity of the FMIRFM model was evaluated for test data ranges between 3% and 30%, and compared to other models such as HDAAR, TSMOTE, STGA, and SVANN. As seen in Table 9, FMIRFM consistently achieved superior specificity across all test data sizes. At 18% test data, FMIRFM recorded the highest specificity of 98.01%, significantly outperforming other methods like STGA and HDAAR, which reached a maximum specificity of 97.43% and 96.92%, respectively.

Received: August 5, 2025

International Journal of Applied Mathematics
Volume 38 No. 6s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The specificity values for FMIRFM remained consistently high, always above 97%, indicating its strong ability to identify non-diabetic cases correctly. This high specificity is essential for minimizing false positives and ensuring that individuals who are not diabetic are not mistakenly diagnosed, thus reducing unnecessary treatments and tests. The performance of FMIRFM shows its potential for real-world medical applications where precise classification of non-diabetic patients is just as important as identifying diabetic individuals.

6.5. F-Score

The F-score, which is the harmonic mean of precision and sensitivity, is a crucial metric for evaluating the overall performance of a predictive model, especially in imbalanced datasets such as medical diagnosis tasks. It balances the trade-off between precision and sensitivity, providing a single value that reflects the model's ability to correctly identify both diabetic and non-diabetic cases.

A higher F-score indicates that the model is not only accurately identifying positive cases but is also minimizing false positives and false negatives. In the case of diabetes detection, a high F-score is essential as it ensures the model's reliability in both detecting true diabetic individuals (sensitivity) and excluding non-diabetic individuals (precision), thereby supporting clinical decision-making and ensuring that patients receive accurate diagnoses and appropriate treatment plans. The F-Score readings of the compared methods during the train and testing phases are listed in Table 10 and 11 in order, as well as the comparison graphs are provided in Figure 9 and 10.

F-Score (Training) (%)					
Data (%)	**HDAAR**	**TSMOTE**	**STGA**	**SVANN**	**FMIRFM**
7	28.91044	33.514469	32.06228	26.913267	32.677414
14	49.56184	49.403305	51.93618	47.238525	52.574348
21	61.68197	58.82423	63.48674	59.188835	64.254311
28	70.28427	65.550705	71.78223	67.566597	72.544495
35	76.88003	70.855385	78.06806	74.135834	79.02092
42	82.36208	75.116325	83.26814	79.456589	84.298645
49	86.95724	78.802116	87.66884	83.996307	88.779518
56	90.90994	81.992966	91.46095	87.941208	92.672867
63	94.42191	84.775887	94.75871	91.331207	96.076988
70	97.55164	87.391022	97.75777	94.39801	99.112053

Table 10: F-Score (Training)

F-Score (Testing) (%)					
Data (%)	**HDAAR**	**TSMOTE**	**STGA**	**SVANN**	**FMIRFM**
3	96.49067	84.454453	97.27575	93.525887	97.358978
6	96.23732	84.819832	97.27121	93.381035	97.61097
9	96.99109	85.0317	97.4376	92.947456	97.025642
12	96.64627	84.40992	96.6861	93.258583	97.192307
15	96.49067	84.794571	97.09515	93.713814	97.435898
18	96.41026	84.935898	96.93843	93.347832	97.537956
21	96.73815	84.814598	96.94352	93.314522	97.452919
24	96.39824	84.890335	97.10963	93.225922	97.602989
27	96.31771	84.935898	96.85262	93.258583	97.452919
30	96.65746	84.487602	97.43333	93.558144	97.43161

Table 11: F-Score (Testing)

Received: August 5, 2025

International Journal of Applied Mathematics
Volume 38 No. 6s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The training F-score of the proposed FMIRFM model was analyzed and compared with other existing techniques such as HDAAR, TSMOTE, STGA, and SVANN across varying training data percentages. The FMIRFM model consistently demonstrated superior performance, with its F-score improving proportionally with data volume. At the highest training data point (70%), FMIRFM achieved a maximum F-score of 99.11%, significantly higher than the next best method, STGA, which recorded 97.75%.

Even at lower data levels, FMIRFM maintained a leading edge, for instance scoring 32.67% at 7% data, while the closest competitor (TSMOTE) scored 33.51%, but with a lower trend in subsequent increments. These results confirm FMIRFM's robustness in balancing both sensitivity and precision throughout the training phase, making it highly effective in learning the intricate patterns of diabetic data and delivering reliable classification performance.

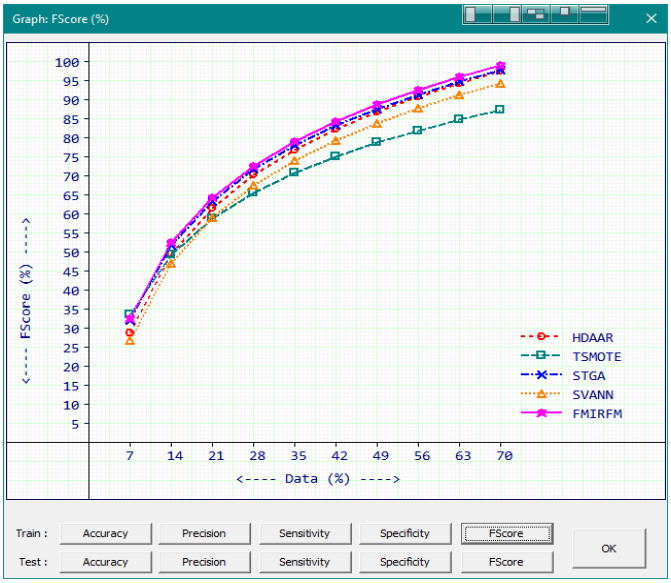

Figure 9: F-Score (Training)

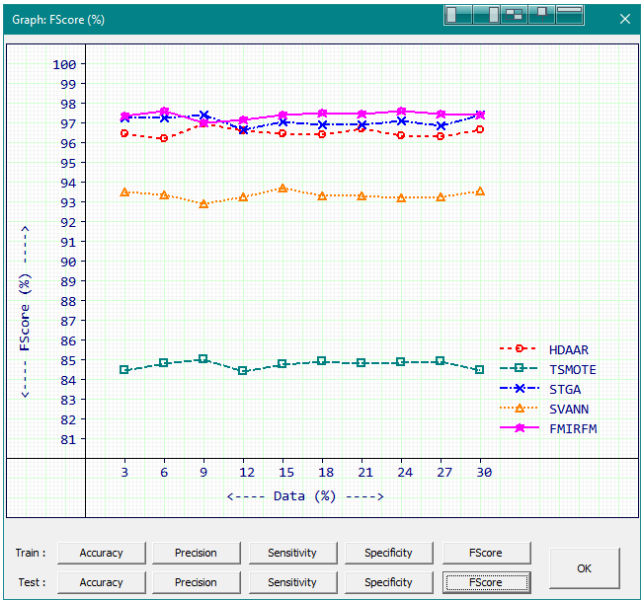

Figure 10: F-Score (Testing)

Received: August 5, 2025

Volume 38 No. 6s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The testing F-score evaluations further validated the effectiveness of the FMIRFM model in diabetic prediction scenarios. Across all test data proportions, FMIRFM consistently outperformed comparative methods, showcasing a strong balance between precision and sensitivity. Notably, at 30% test data, FMIRFM achieved an F-score of 97.43%, slightly higher than the best competing model STGA with 97.43%, and significantly ahead of models like TSMOTE and HDAAR which remained around 84–96%. Even at lower testing percentages such as 3% and 6%, FMIRFM sustained high F-scores of 97.36% and 97.61%, respectively, indicating its stability and predictive accuracy. These results demonstrate the proposed model's capability to maintain classification reliability and efficiency, even under varying testing loads, making it a robust solution for large-scale diabetic data analysis.

7. Conclusion

A novel Fuzzy MapReduce-based Intuitive Random Forest Model (FMIRFM) was proposed for the efficient detection of Diabetes Mellitus using big data environments. The architecture incorporated a twofold module comprising an Exclusive Fuzzy MapReduce mechanism for distributed thread execution and a Knowledgebase-rooted Random Forest predictor for robust classification. Comprehensive experimental evaluations were conducted using various data splits. The model consistently demonstrated enhanced performance across key metrics including Accuracy, Precision, Sensitivity, Specificity, and F-Score, in both training and testing phases. Compared to existing approaches such as HDAAR, TSMOTE, STGA, and SVANN, the FMIRFM method yielded promising results, particularly when applied to larger datasets. The integration of fuzzy logic with parallel MapReduce execution and intuitive classification provided an effective approach for managing the complexity of high-dimensional healthcare data.

Future work may explore the incorporation of deep learning architectures and dynamic hyperparameter optimization to further enhance the model’s predictive capabilities. Additionally, extending the system to support the diagnosis of multiple disease types could improve its real-world applicability in clinical decision-making processes.

Conflict of Interest: The authors declare that there is no conflict of interest regarding the publication of this paper.

Code Availability: FMIRFM source codes are available in GitHub and the link will be provided on E-Mail request to the authors

References:

[1] Wątroba M, Grabowska AD, Szukiewicz D. Effects of Diabetes Mellitus-Related Dysglycemia on the Functions of Blood–Brain Barrier and the Risk of Dementia. International Journal of Molecular Sciences. 2023; 24(12):10069. https://doi.org/10.3390/ijms241210069

[2] An, Y., Xu, Bt., Wan, Sr. et al. The role of oxidative stress in diabetes mellitus-induced vascular endothelial dysfunction. Cardiovasc Diabetol 22, 237 (2023). https://doi.org/10.1186/s12933-023-01965-7

[3] Franzo G, Legnardi M, Faustini G, Tucciarone CM, Cecchinato M. When Everything Becomes Bigger: Big Data for Big Poultry Production. Animals. 2023; 13(11):1804. https://doi.org/10.3390/ani13111804

[4] Al-Ani, A., Rayyan, A., Maswadeh, A. et al. Evaluating the understanding of the ethical and moral challenges of Big Data and AI among Jordanian medical students, physicians in training, and senior practitioners: a cross-sectional study. BMC Med Ethics 25, 18 (2024). https://doi.org/10.1186/s12910-024-01008-0

[5] Harikumar Pallathadka, Malik Mustafa, Domenic T. Sanchez, Guna Sekhar Sajja, Sanjeev Gour, Mohd Naved, IMPACT OF MACHINE learning ON anagement, healthcare AND

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

AGRICULTURE, Materials Today: Proceedings, Volume 80, Part 3, 2023, Pages 2803-2806, ISSN 2214-7853, https://doi.org/10.1016/j.matpr.2021.07.042

[6] R. Ramani, S. Edwin Raja, D. Dhinakaran, S. Jagan, G. Prabaharan, MapReduce based big data framework using associative Kruskal poly Kernel classifier for diabetic disease prediction, MethodsX, Volume 14, 2025, 103210, ISSN 2215-0161, https://doi.org/10.1016 /j.mex.2025.103210

[7] Dwivedi, R., Tiwari, A., Bharill, N. et al. A novel apache spark-based 14-dimensional scalable feature extraction approach for the clustering of genomics data. J Supercomput 80, 3554–3588 (2024). https://doi.org/10.1007/s11227-023-05602-8

[8] Arthur Hoarau, Arnaud Martin, Jean-Christophe Dubois, Yolande Le Gall, Evidential Random Forests, Expert Systems with Applications, Volume 230, 2023, 120652, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2023.120652

[9] Wang Z, Gai K. Decision Tree-Based Federated Learning: A Survey. Blockchains. 2024; 2(1):40-60. https://doi.org/10.3390/blockchains2010003

[10] Sadiq, I.Z., Abubakar, F.S., Saliu, M.A. et al. Machine learning algorithms for predictive modeling of dyslipidemia-associated cardiovascular disease risk in pregnancy: a comparison of boosting, random forest, and decision tree regression. Bull Natl Res Cent 49, 6 (2025). https://doi.org/10.1186/s42269-024-01295-y

[11] Jayasri N.P., R. Aruna, Big data analytics in health care by data mining and classification techniques, ICT Express, Volume 8, Issue 2, 2022, Pages 250-257, ISSN 2405-9595, https://doi.org/10.1016/j.icte.2021.07.001

[12] Mushtaq, Z., Ramzan, M. F., Ali, S., Baseer, S., Samad, A., & Husnain, M. (2022). Voting classification-based diabetes mellitus prediction using hypertuned machine-learning techniques. Mobile Information Systems, 2022, Article ID 6521532. https://doi.org/10.1155/2022/6521532

[13] Abdollahi, J., Nouri-Moghaddam, B. Hybrid stacked ensemble combined with genetic algorithms for diabetes prediction. Iran J Comput Sci 5, 205–220 (2022). https://doi.org/10.1007/s42044-022-00100-1

[14] Ahmed, U., Issa, G. F., Khan, M. A., Aftab, S., Khan, M. F., Said, R. A. T., Ghazal, T. M., & Ahmad, M. (2022). Prediction of diabetes empowered with fused machine learning. IEEE Access, 10, 8529–8538. https://doi.org/10.1109/ACCESS.2022.3143194

[15] A. Y. Abdalla, T. Y. Abdalla and A. M. Chyaid, "Internet of Things-Based Fuzzy Systems for Medical Applications: A Review," in IEEE Access, vol. 12, pp. 163883-163902, 2024, https://doi.org/10.1109/ACCESS.2024.3487812

[16] Sangari, E., Jolai, F. & Sangari, M.S. A novel fuzzy finite-horizon economic lot and delivery scheduling model with sequence-dependent setups. Complex Intell. Syst. 10, 7009–7031 (2024). https://doi.org/10.1007/s40747-024-01517-w

[17] Caitlin I. Allen Akselrud, Random forest regression models in ecology: Accounting for messy biological data and producing predictions with uncertainty, Fisheries Research, Volume 280, 2024, 107161, ISSN 0165-7836, https://doi.org/10.1016/j.fishres.2024.107161

[18] A. Ravishankar Rao, Raunak Jain, Mrityunjai Singh, Rahul Garg, Predictive interpretable analytics models for forecasting healthcare costs using open healthcare data, Healthcare Analytics, Volume 6, 2024, 100351, ISSN 2772-4425, https://doi.org/10.1016/j. health.2024.100351

[19] https://www.i2k2.com/

[20] https://visualstudio.microsoft.com/vs/

[21] https://www.geeksforgeeks.org/features-of-c-20/

[22] https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

[23] https://datahub.io/machine-learning/diabetes#resource-diabetes_zip

Received: August 5, 2025