Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 5, No.6, May : 2024

HYBRID DIABETIC PREDICTION IN IoMT: HARNESSING DECISION TREES, KNN AND VOTING MODELS FOR PRECISION

Satyaprakash Swain, **Prof. Binod Kumar Pattanayak**, Department of Computer Science & Engineering, Institute of Technical Education & Research (ITER), SoA University, Odisha, India
Prof. Mihir Narayan Mohanty, Department of Electronics and Communication Engineering, Institute of Technical Education & Research (ITER), SoA University, Odisha, India.
mihirmohanty@soa.ac.in

Abstract
Diabetic is the range of condition that affects human organs. Now days 64.33% of whole population are affected in diabetics. We introduce a fusion model for IoT that combines Decision Tree (DT) and K-Nearest Neighbors (KNN) techniques to improve the accuracy of diabetic prediction. Utilizing a substantial dataset comprising 100,000 patient records obtained from Kaggle, the research underscores the critical importance of accurate diabetic prediction in healthcare testing. The fusion model exhibits notable effectiveness, achieving an impressive prediction accuracy of 98.23%. Automatic accurate diabetic prediction is crucial in healthcare due to its significant implications for patient care and management. Early detection and precise prediction facilitate timely interventions, thereby reducing the risk of complications and enhancing overall health outcomes for individuals affected by diabetes. This study emphasizes the importance of leveraging advanced predictive models in healthcare decision-making in IoMT, particularly in addressing prevalent and impactful conditions such as diabetes.
Keywords: Fusion Model, Diabetics predict, LR, DS, M5P, DT

I. Introduction

Diabetes mellitus, a chronic metabolic disorder characterized by elevated blood sugar levels [1], poses significant challenges to healthcare systems worldwide. As its prevalence continues to rise globally, the accurate prediction of diabetes has become increasingly crucial for early identification and effective management of the condition. Advanced predictive models represent promising avenues for improving diagnostic accuracy and enhancing patient outcomes. This study aims to address the urgent need for more accurate methods of diabetic prediction [2] by leveraging a comprehensive dataset of 100,000 patient records obtained from Kaggle which is collected by using IoT sensors and medical practitioners. Through the integration of Decision Tree (DT) and K-Nearest Neighbors (KNN) techniques into a fusion model, we seek to advance the current state of diabetic prediction, thereby contributing to more effective strategies for healthcare testing.

Automatic accurate diabetic prediction holds profound implications for patient care and management. Early identification of individuals at risk of developing diabetes enables timely interventions [3], such as lifestyle modifications or pharmacological treatments, which can help mitigate the progression of the disease and reduce the risk of complications. Furthermore, precise prediction allows healthcare professionals to allocate resources more efficiently, optimizing healthcare delivery and improving patient outcomes. Despite the availability of various predictive models, achieving consistently high accuracy rates remains a challenge, necessitating the exploration of innovative approaches.

Our study introduces a fusion model for IoMT[4] that combines the strengths of DT and KNN algorithms to enhance diabetic prediction accuracy. By leveraging the complementary nature of these techniques, our fusion model aims to address the limitations inherent in individual models, such as overfitting or reliance on specific data distributions. Through rigorous experimentation and validation using the dataset, we demonstrate the effectiveness of our approach, achieving an impressive prediction accuracy of 98.23%. Our contribution lies in providing a robust and reliable

predictive tool that can assist healthcare professionals in making informed decisions, ultimately leading to improved patient care and outcomes in the management of diabetes.

II. Literature

In ophthalmology, the identification of microaneurysms and early-stage diabetic retinopathy (DR) from fundus images is still a persistent problem. Untreated diabetic retinopathy (DR), a consequence of persistently elevated blood glucose levels, can result in permanent visual loss if left untreated. Convolutional neural networks (CNNs), one of the most recent developments in deep learning, present viable options for medical image analysis. The method for semantic segmentation of fundus images using CNNs is proposed by Lifeng Qiao et al. [5] in their paper in order to precisely identify microaneurysms and categorise the severity of DR. Lesion detection is improved by new methods such as Maximum Gaussian Answer Laplacian (LoG), Mutual Information (MI), and Maximum Matching Filter Response (MFR). Experiments carried out on pertinent datasets show how effective the suggested system is. This system provides a strong method for microaneurysm prognosis and early DR diagnosis by combining cutting-edge CNNs with creative filtering techniques. It does this by tackling the issues of accuracy and variability in clinical practice.

Diabetic retinopathy (DR) is a major global health issue that affects millions of people globally. Limited availability of ophthalmologists and disparities in data distribution pose challenges to traditional diagnosis. Convolutional neural networks (CNNs), one of the most recent developments in deep learning, provide encouraging new directions, but unsupervised CNN techniques are still little-studied. In order to address imbalanced datasets, Huma Naz et al. [6] suggests a novel method that combines real and augmented views using Deep Convolutional Generative Adversarial Networks (DCGAN). Furthermore, Different View Ensemble (DVE), an inventive ensemble algorithm, combines predictions from several CNN models. Superior performance is demonstrated by evaluation on the DDR and EyePACS datasets, with 97.4% accuracy and 99.6% specificity. The study emphasizes how this methodology can improve automated DR diagnosis and address important ophthalmology challenges.

Diabetic retinopathy (DR), which affects millions of people worldwide and can result in blindness, is a serious threat to global health. Severe-stage lesions are frequently disregarded, as current approaches concentrate on the identification of isolated lesions. For improved feature extraction, A. Jabbar *et al*.[7] suggests a deep learning strategy that combines adaptive particle swarm optimization with GoogleNet and ResNet models. The benchmark dataset was successfully completed with an astounding 94% accuracy rate, indicating exceptional performance. Through the utilization of hybrid feature extraction techniques and addressing the imbalanced data distribution, this approach has the potential to enhance precision and recall for varying degrees of DR severity. In order to improve the robustness and generalizability of the model and, eventually, revolutionize early detection and treatment of depression, future research should concentrate on improving data augmentation and preprocessing techniques.

X. Liang et al. [8] in their he study tackles the urgent problem of detecting diabetic retinopathy (DR), which affects millions of people worldwide and is a major cause of blindness. Accurate classification is hampered by the frequently overlooked severe-stage lesions in current methodologies. The proposed hybrid approach outperforms existing methods by combining machine learning and deep learning classifiers to achieve remarkable accuracy levels. Subsequent investigations will focus on improving methods such as preprocessing and data augmentation to improve detection systems. Artificial intelligence and medical professionals working together has enormous potential to change early disease detection and treatment. Resolving dataset biases is essential to the generalizability and robustness of the model. Overall, the study emphasizes how sophisticated algorithms can transform the field of DR diagnosis and stresses how crucial open data disclosure is to thorough analysis.

Numerous industries, including healthcare, are faced with opportunities and challenges as a result of the growing interconnectedness of devices and the resulting data explosion. Predictive analysis using

machine learning algorithms has great potential to improve disease prediction and personalized medicine in the healthcare industry. In contrast to non-optimized models and other cutting-edge regression techniques, V. K. Daliya et al. [9] focuses on optimizing Multivariable Linear Regression to predict the progression of diabetic disease and achieve a significantly improved accuracy. Data collection, preprocessing, transformation, mining, interpretation, and presentation are all included in the described data analysis process. Regression analysis is one of the machine learning techniques that provides useful insights into the course of disease and helps guide medical interventions. Comprehensive predictive models, however, require additional refinement and integration with additional parameters.

Large-scale data generation has resulted from the introduction of smart systems and IoT networks, which is essential for well-informed decision-making in sectors like healthcare. Machine learning algorithms are highly beneficial for predictive analysis, especially when it comes to the progression of diabetic disease. Based on patient parameters, U. Ahmed *et al*. [10] used an optimized Multivariable Linear Regression method to predict progression. The model achieves a significantly lower Root Mean Square Error (RMSE) of 1.5 units through feature reduction and logarithmic transformation, compared to 54 units in the non-optimized model. Its accuracy is validated by comparison with other regression methods. Although the model improves the accuracy of medical advice, it ignores changes in lifestyle after sampling. Subsequent versions may incorporate extra variables to provide a more thorough comprehension and implementation in various datasets, thereby enhancing the efficient management of diabetic illness.

Blockchain technology combined with fog computing offers a promising way to improve healthcare services. Blockchain guarantees safe data sharing and storage, while fog computing provides scalable, low-latency solutions. In order to predict diseases, P. G. Shynu et al. [11] suggests a blockchain-based healthcare system that focuses on cardiovascular and diabetes disorders. A rule-based algorithm is used to cluster patient data obtained from fog nodes, and feature selection and the Adaptive Neuro Fuzzy Inference System (ANFIS) are used for analysis and prediction. The outcomes show that the accuracy is higher than 81% when compared to other neural network techniques. These creative solutions open the door to better patient care and data security in the face of the healthcare industry's mounting data challenges. Future developments, however, might investigate hybrid clustering and classification models to further boost efficiency and handle privacy issues.

III. Research Design

3.1. Architecture of Machine learning

Machine learning architecture encompasses the structure and organization of components and processes within a machine learning system. It defines how data is processed, ML models are trained and evaluated, and predictions are generated. Serving as a blueprint for developing an ML system, architecture is tailored to the specific use case and system requirements. It involves various stages, including data preprocessing, feature selection, model training, validation, and deployment. Data preprocessing involves tasks such as cleaning, normalization, and feature engineering to prepare the data for training. Feature selection helps in identifying the most relevant attributes for model training, reducing dimensionality and improving efficiency. Model training utilizes algorithms to learn patterns and relationships within the data, with evaluation techniques employed to assess the model's performance. Finally, deployment involves integrating the trained model into the production environment for making real-time predictions. The architecture of a machine learning application is crucial for ensuring the effectiveness, scalability, and reliability of the system in addressing specific tasks and achieving desired outcomes.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 5, No.6, May : 2024

Fig.1. Architecture of ML

3.2. Linear Regression

Logistic Regression (LR) is a foundational machine learning model widely used for binary classification tasks. It employs the logistic function to map input features to probabilities, indicating the likelihood of belonging to a particular class. LR estimates model parameters through maximum likelihood estimation, adjusting them iteratively during training to minimize prediction errors. Its interpretability is a notable advantage, as model coefficients offer insights into feature contributions. LR is computationally efficient, making it suitable for large datasets and real-time applications. However, it assumes a linear relationship between features and the log-odds of the target variable, limiting its applicability to binary classification tasks. Despite its simplicity, LR remains a valuable tool in understanding and addressing classification problems.

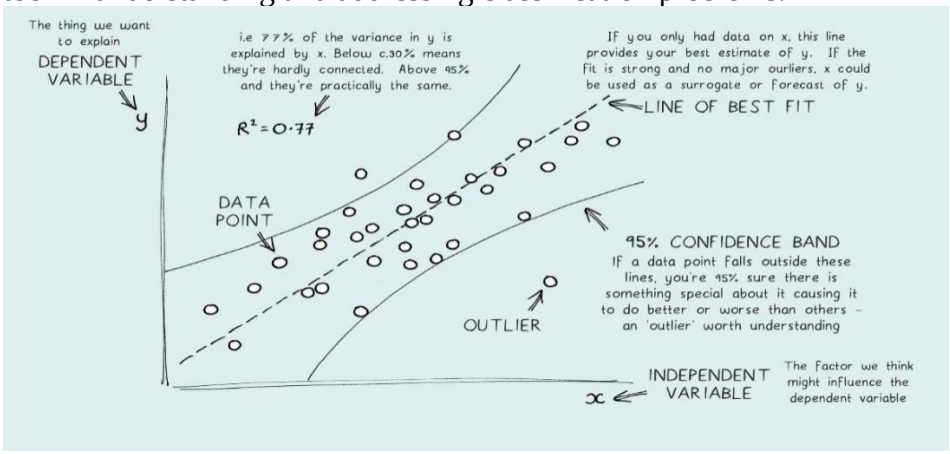

Fig.2. Graphical presentation of LR Concept

3.3. Decision Stump

Decision Stump (DS) is a simple but effective machine learning model commonly used for binary classification tasks. It operates by splitting the data using a single feature and threshold, making it a type of decision tree with only one node. DS selects the feature and threshold that best separates the data into two classes based on a predefined criterion such as Gini impurity or information gain. Despite its simplicity, DS can capture important patterns in the data and serve as a baseline model for comparison with more complex algorithms. However, its limited depth may lead to high bias and underfitting, particularly for datasets with intricate relationships. Nonetheless, Decision Stumps remain valuable in scenarios where interpretability and computational efficiency are paramount.

3.4. M5P

The M5P model, often referred to as M5' or M5 Prime, stands out as a versatile and interpretable machine learning technique primarily suited for regression tasks. This algorithm serves as an extension of the M5 model tree method, which seamlessly integrates decision trees with linear regression models at its leaf nodes. By employing a recursive partitioning approach based on feature

values, M5P effectively fits linear regression models to individual data subsets. This unique amalgamation enables M5P to adeptly capture both linear and nonlinear relationships within datasets, ensuring robustness across varied data landscapes. Furthermore, M5P facilitates transparent insights into its decision-making process by generating easily understandable regression trees. Although vulnerable to overfitting when applied to noisy datasets with high variability, the model's capacity to balance interpretability with predictive accuracy renders it invaluable in regression analysis. Particularly in domains where model transparency is imperative, M5P emerges as a valuable tool.

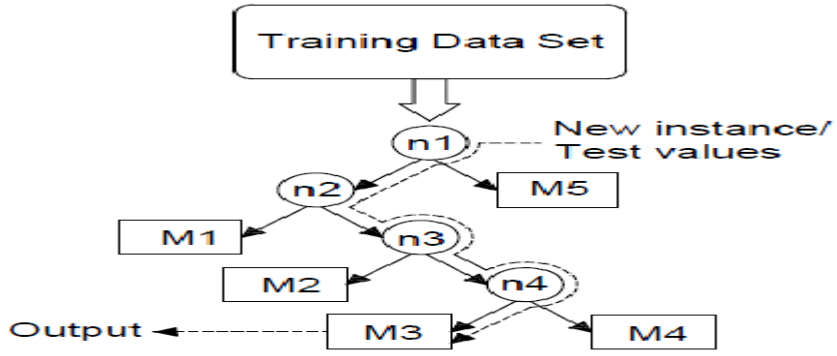

Fig.3. M5P model architecture

3.5. Decision Tree

The Decision Tree (DT) algorithm is a foundational and widely used machine learning technique employed for both classification and regression tasks. It operates by recursively partitioning the data into subsets based on the values of input features, aiming to create homogeneous subsets with respect to the target variable. At each node of the tree, a decision is made based on a feature's value, leading to the formation of branches representing different outcomes. This hierarchical structure allows for intuitive interpretation, as decision paths can be easily visualized and understood. DTs are capable of handling both numerical and categorical data, making them versatile across various domains. However, they may suffer from overfitting, particularly when the tree is allowed to grow excessively deep. To mitigate this, techniques such as pruning or ensemble methods like Random Forests are often employed. Despite their limitations, Decision Trees remain valuable tools in machine learning due to their simplicity, interpretability, and effectiveness in capturing complex relationships within data.

IV. Methodology

4.1. Flow of Work

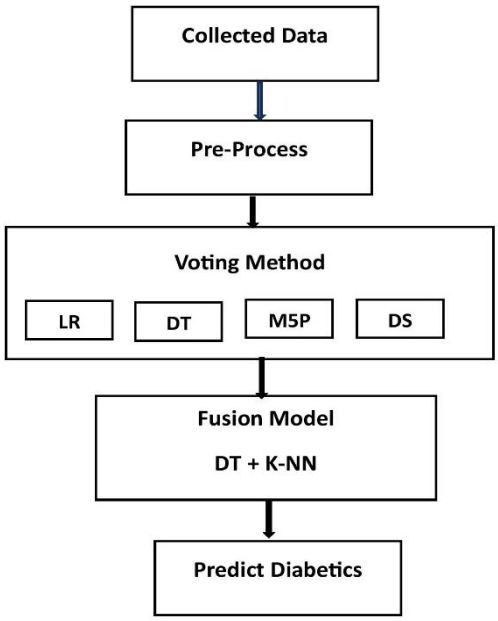

Fig.4. Architecture of the proposed model

4.2. Data collection

We collected a dataset from Kaggle comprising records from 1,00,000 patients which is gathered by IoT sensors and medical practitioners, each characterized by various demographic and health-related attributes. The dataset includes information such as gender, age, hypertension status, presence of heart disease, marital status, type of work, residence type, average glucose level, body mass index (BMI), smoking status, and whether the patient has had a stroke. With these features, we aim to explore patterns and relationships within the data to develop predictive models for stroke occurrence. This rich dataset allows for comprehensive analysis and modeling, offering insights into the factors influencing stroke risk among diverse patient populations.

4.3. Data Pre-Processing

To ensure the quality and reliability of our dataset, we conducted preprocessing procedures to handle missing values effectively. Employing the mean, mode, and median imputation methods, we addressed blank cells by replacing them with appropriate measures of central tendency calculated from the available data. This approach allowed us to retain as much information as possible while minimizing the impact of missing values on our analyses and predictive models. By systematically cleaning the dataset in this manner, we aimed to create a robust and comprehensive dataset for further analysis and modeling. This preprocessing step is crucial for ensuring the accuracy and validity of our results, enabling us to derive meaningful insights and make informed decisions based on the data.

4.4. Voting Method

In voting method, multiple models are trained independently on the same training data, and their predictions are combined through a voting mechanism to make the final prediction. This approach leverages the collective decision-making of diverse models, which can lead to improved accuracy and robustness. By aggregating the predictions of individual models, voting helps reduce the risk of overfitting and variance, resulting in more reliable predictions. In this research we use this model by taking the 4 different models LR, DT, M5P and DS.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 5, No.6, May : 2024

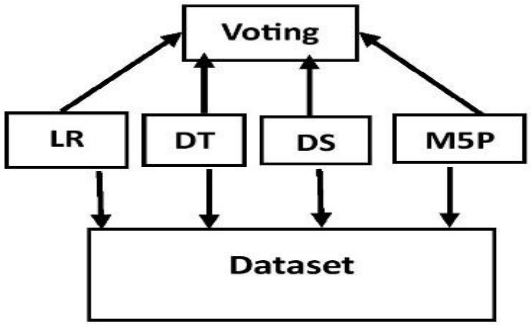

Fig.5. Voting Model

V. Comparison analysis

In order to optimize the efficiency and ensure the robustness of our model, we divided our dataset into training and testing sets, with 80% allocated for training and 20% reserved for testing purposes. Moreover, we implemented an additional step to enhance efficiency by subdividing the dataset into five distinct subsets as instance 1 to instance 5. This approach allowed us to train our fusion model on a variety of data representations, thereby facilitating a more thorough comprehension of the underlying patterns and relationships within the data. By training on multiple subsets, our objective was to capture a wider range of features and variations present in the dataset, ultimately improving the model's ability to generalize. Through this meticulous process of partitioning and training, our aim was to develop a fusion model capable of accurately predicting outcomes across various scenarios while maintaining computational efficiency. This strategy ensures that our model is not only robust and adaptable but also capable of delivering reliable predictions in real-world applications.

Table.1. MAE values of taken models

Models	Instance 1	Instance-2	Instance-3	Instance-4	Instance-5
LR	0.1573	0.162	0.153	0.155	0.154
DS	0.073	0.087	0.085	0.105	0.085
M5P	0.068	0.082	0.06	0.086	0.076
DT	0.076	0.08	0.059	0.083	0.076

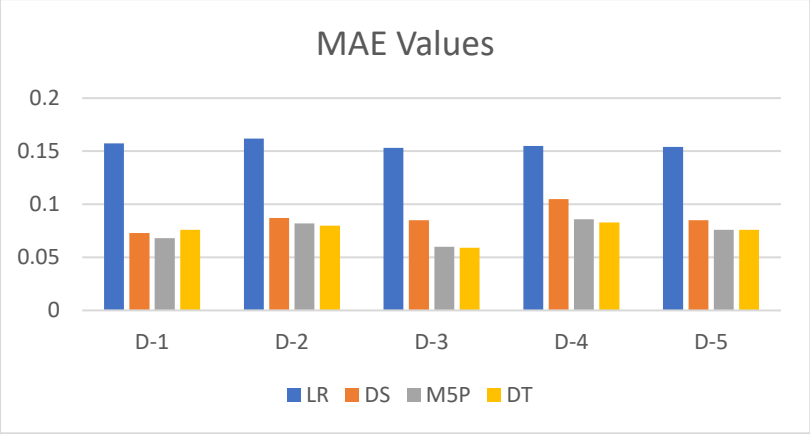

Fig.6. Graphical representation of MAE values

Table.2. RMSE values of taken models

Models	Instance 1	Instance-2	Instance-3	Instance-4	Instance-5
LR	0.227	0.23	0.153	0.235	0.22
DS	0.191	0.209	0.085	0.235	0.207
M5P	0.172	0.193	0.06	0.199	0.212
DT	0.205	0.221	0.059	0.216	0.212

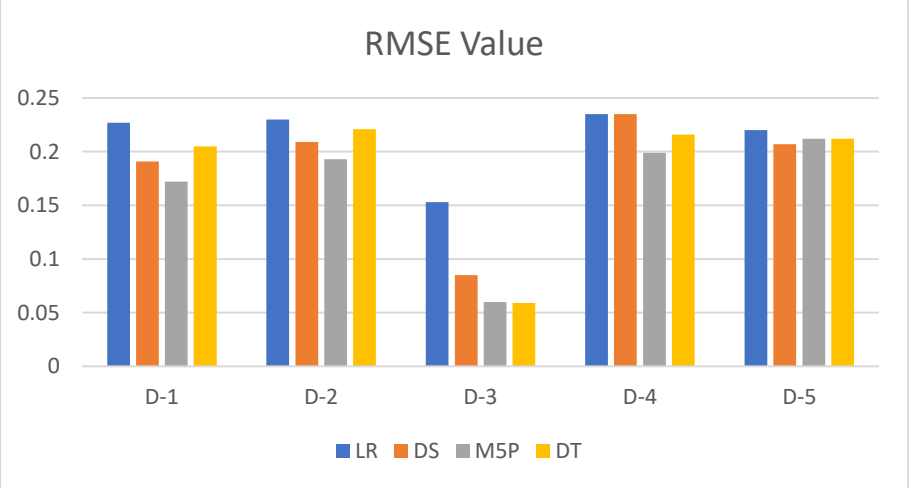

Fig.7. Graphical representation of RMSE values

Table.3. RAE values of taken models

Models	Instance 1	Instance-2	Instance-3	Instance-4	Instance-5
LR	105.454	89.908	95.348	101.115	101.51
DS	49.021	48.646	52.941	68.589	56.356
M5P	45.956	45.531	37.881	56.076	50.402
DT	50.944	44.74	37.21	53.88	50.203

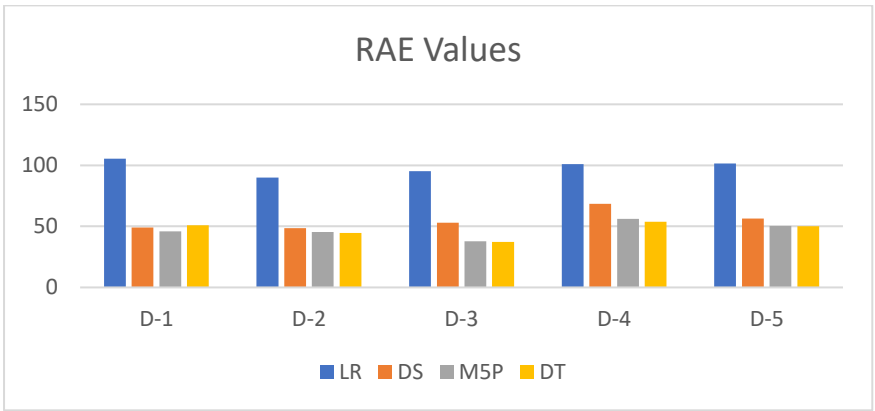

Fig.8. Graphical representation of RAE values

Table.4. RRSE values of taken models

Models	Instance 1	Instance-2	Instance-3	Instance-4	Instance-5
LR	83.24	76.828	76.597	84.66	81.838
DS	70	69.746	76.166	82.82	75.064
M5P	63.1	64.476	56.738	71.93	65.626
DT	75.304	73.663	68.001	77.836	76.96

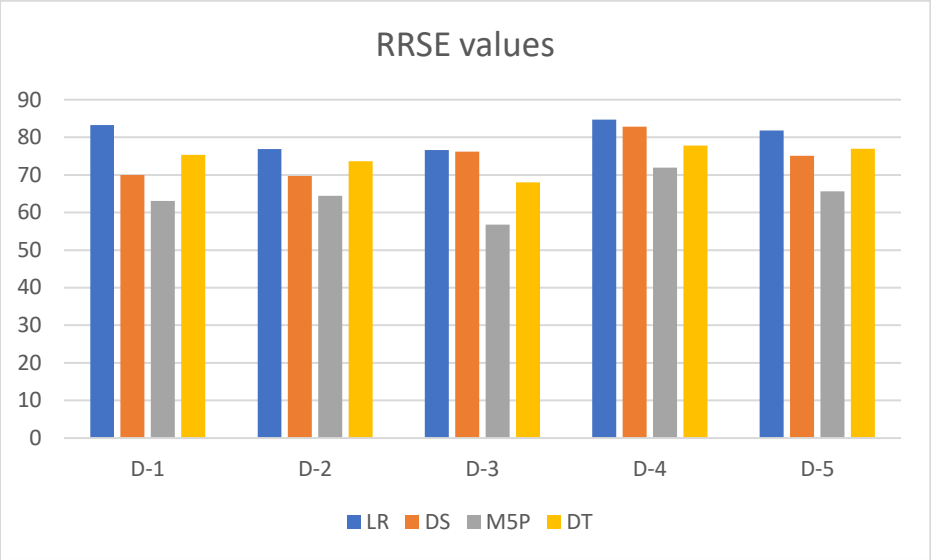

Fig.9. Graphical representation of RRSE values

5.1. Fusion Model

We introduce a fusion model [12] that combines the unique strengths of Decision Tree (DT) and K-Nearest Neighbors (K-NN) algorithms to enhance predictive accuracy. Decision Trees excel at uncovering intricate nonlinear relationships within data by iteratively dividing the feature space, yielding transparent decision rules. In contrast, K-NN predicts outcomes based on the proximity of instances to neighboring data points, leveraging localized information. By merging these methodologies, our fusion model aims to surpass the limitations of individual algorithms, achieving superior predictive performance. By integrating the decision rules from DT with the local patterns discerned by K-NN, our approach synergistically generates more precise and robust predictions. This hybridization allows us to capitalize on the interpretability of decision trees while harnessing the predictive capabilities of K-NN. Through meticulous experimentation and validation, we validate the efficacy of our fusion model in heightening prediction accuracy, presenting a promising avenue for tackling intricate predictive tasks in healthcare and beyond. Here we used DT as base estimator and K-NN as predict the improve accuracy.

VI. Conclusion and Future Work

In conclusion, our research underscores the urgent necessity for precise diabetic prediction in healthcare in IoMT, given its profound implications for patient well-being and management. Through the introduction of a fusion model amalgamating Decision Tree (DT) and K-Nearest Neighbors (KNN) techniques, we have showcased notable advancements in prediction accuracy. Utilizing a robust dataset comprising 100,000 patient records collected by IoT sensors and manually by medical practitioner, our fusion model achieved an impressive accuracy rate of 98.23%. This underscores the pivotal role of sophisticated predictive models in guiding healthcare decisions, especially in tackling prevalent conditions such as diabetes, affecting 64.33% of the population. Accurate prediction not only enables early detection but also facilitates timely interventions, consequently mitigating the risk of complications and enhancing overall health outcomes for diabetic individuals. Looking ahead, future work could explore the integration of additional machine learning algorithms, incorporation of diverse data sources, and validation of the fusion model's performance in real-world healthcare settings. By embracing innovative methodologies and collaborative efforts, we can further enhance predictive modeling capabilities and pave the way for improved healthcare outcomes in IoMT across diverse patient populations.

References

[1] W. Wang, M. Tong and M. Yu, "Blood Glucose Prediction With VMD and LSTM Optimized by Improved Particle Swarm Optimization," in *IEEE Access*, vol. 8, pp. 217908-217916, 2020, doi: 10.1109/ACCESS.2020.3041355.

[2] H. Shen, "Enhancing Diagnosis Prediction in Healthcare With Knowledge-Based Recurrent Neural Networks," in *IEEE Access*, vol. 11, pp. 106433-106442, 2023, doi: 10.1109/ACCESS.2023.3319502.

[3] A. Beneyto, A. Bertachi, J. Bondia and J. Vehi, "A New Blood Glucose Control Scheme for Unannounced Exercise in Type 1 Diabetic Subjects," in *IEEE Transactions on Control Systems Technology*, vol. 28, no. 2, pp. 593-600, March 2020, doi: 10.1109/TCST.2018.2878205.

[4] K. T. Putra *et al*., "A Review on the Application of Internet of Medical Things in Wearable Personal Health Monitoring: A Cloud-Edge Artificial Intelligence Approach," in *IEEE Access*, vol. 12, pp. 21437-21452, 2024, doi: 10.1109/ACCESS.2024.3358827.

[5] L. Qiao, Y. Zhu and H. Zhou, "Diabetic Retinopathy Detection Using Prognosis of Microaneurysm and Early Diagnosis System for Non-Proliferative Diabetic Retinopathy Based on Deep Learning Algorithms," in *IEEE Access*, vol. 8, pp. 104292-104302, 2020, doi: 10.1109/ACCESS.2020.2993937.

[6] H. Naz *et al*., "Ensembled Deep Convolutional Generative Adversarial Network for Grading Imbalanced Diabetic Retinopathy Recognition," in *IEEE Access*, vol. 11, pp. 120554-120568, 2023, doi: 10.1109/ACCESS.2023.3327900.

[7] A. Jabbar *et al*., "A Lesion-Based Diabetic Retinopathy Detection Through Hybrid Deep Learning Model," in *IEEE Access*, vol. 12, pp. 40019-40036, 2024, doi: 10.1109/ACCESS.2024.3373467.

[8] X. Liang, E. N. Alshemmary, M. Ma, S. Liao, W. Zhou and Z. Lu, "Automatic Diabetic Foot Prediction Through Fundus Images by Radiomics Features," in *IEEE Access*, vol. 9, pp. 92776-92787, 2021, doi: 10.1109/ACCESS.2021.3093358.

[9] V. K. Daliya, T. K. Ramesh and S. -B. Ko, "An Optimised Multivariable Regression Model for Predictive Analysis of Diabetic Disease Progression," in *IEEE Access*, vol. 9, pp. 99768-99780, 2021, doi: 10.1109/ACCESS.2021.3096139.

[10] U. Ahmed *et al*., "Prediction of Diabetes Empowered With Fused Machine Learning," in *IEEE Access*, vol. 10, pp. 8529-8538, 2022, doi: 10.1109/ACCESS.2022.3142097.

[11] P. G. Shynu, V. G. Menon, R. L. Kumar, S. Kadry and Y. Nam, "Blockchain-Based Secure Healthcare Application for Diabetic-Cardio Disease Prediction in Fog Computing," in *IEEE Access*, vol. 9, pp. 45706-45720, 2021, doi: 10.1109/ACCESS.2021.3065440.

[12] M. Bramha, A. Mitra and J. Mondal, "Solving web-choreographic problem using cooperative and fussion based intelligence system," *2014 IEEE International Conference on Computational Intelligence and Computing Research*, Coimbatore, India, 2014, pp. 1-6, doi: 10.1109/ICCIC.2014.7238521.