International Research Journal of Multidisciplinary Sciences (Double - Blind Peer Review Journal) VOL-2 ISSUE-4 April 2026 PP:1-7 ISSN:3107-930X(ONLINE)

Comparative Evaluation of Machine Learning Algorithms for Heart Disease Prediction Using Clinical Data

[1]Omkar Patil, [2]Shyamol Banerjee
[1]Department of CSE, SRCEM India [2]Department of CSE, SRCEM, India
[1]omkarpatil5388@gmail.com, [2]shyamolgwl@gmail.com

***Abstract*—Cardiovascular disease remains one of the leading causes of mortality worldwide, and early prediction is essential for reducing clinical risk and improving patient outcomes. This paper presents a comparative evaluation of three supervised machine learning algorithms: Logistic Regression, Decision Tree, and Random Forest for heart disease prediction using structured clinical data. The study follows a standard machine learning workflow consisting of data preprocessing, exploratory data analysis, model training, and performance evaluation. The models are assessed using accuracy, precision, recall, and F1-score. Experimental results indicate that Random Forest achieves the best overall predictive performance due to its ensemble learning capability and robustness against overfitting. The findings demonstrate the potential of machine learning models as decision-support tools for early heart disease risk prediction.**

***Index Terms*—Heart disease prediction, machine learning, Logistic Regression, Decision Tree, Random Forest, clinical data, healthcare analytics.**

I. Introduction

Cardiovascular diseases (CVDs) remain one of the leading causes of mortality worldwide and represent a major burden on public health systems. According to the World Health Organization, CVDs account for approximately 17.9 million deaths annually, making them a critical global health concern [**?**]. Heart disease is particularly challenging because it may develop gradually and remain asymptomatic during its early stages. As a result, many patients are diagnosed only after the disease has progressed to a severe condition, increasing the risk of complications, treatment cost, and mortality.

Early prediction of heart disease is therefore essential for preventive healthcare and timely clinical intervention. Conventional diagnostic approaches such as electrocardiography, echocardiography, stress testing, angiography, and laboratory-based examinations are widely used in clinical practice. However, these methods often require specialized equipment, expert interpretation, and sufficient healthcare infrastructure. In rural or resource-limited settings, such facilities may not always be available, which can delay diagnosis and treatment [2], [3].

Machine learning has emerged as a promising approach for improving disease prediction by identifying hidden patterns in clinical data. In healthcare analytics, machine learning algorithms can process multidimensional patient records, learn relationships among risk factors, and generate predictive outputs that support clinical decision-making [**?**]. Several studies have reported that supervised learning algorithms such as Logistic Regression, Decision Tree, Support Vector Machine, K-Nearest Neighbors, and Random Forest can be effectively applied to heart disease prediction tasks [5], [6]. Among these, ensemble-based methods such as Random Forest often show strong predictive performance because they combine multiple decision trees and reduce overfitting [4].

The objective of this paper is to comparatively evaluate three supervised machine learning algorithms: Logistic Regression, Decision Tree, and Random Forest for heart disease prediction using structured clinical data. The study focuses on dataset preparation, exploratory data analysis, model development, and performance comparison using standard evaluation metrics such as accuracy, precision, recall, and F1-score.

II. Related Work

Machine learning has become an important research direction for predictive healthcare because of its ability to identify complex and nonlinear patterns from structured and unstructured data. In the context of cardiovascular disease prediction, several researchers have applied supervised learning algorithms such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbors, Naive Bayes, and ensemble classifiers to improve diagnostic accuracy and support early clinical decision-making [2]–[6]. These studies generally show that machine learning models can support risk prediction when appropriate preprocessing, feature selection, and evaluation methods are applied.

Shah *et al*. [2] investigated machine learning techniques for heart disease prediction and highlighted the importance of preprocessing and model comparison. Jindal *et al*. [3] compared supervised learning models and emphasized the role of normalization and visualization in improving prediction performance. Katarya and Meena [5] reported that ensemble-based models such as Random Forest and Gradient Boosting often perform better than single classifiers due to their ability to reduce variance and handle nonlinear feature interactions. Similarly, Ali *et al*. [4] demonstrated that automated diagnostic systems using machine learning can improve heart disease classification performance and provide useful support for clinical decision-making.

Recent research has also expanded beyond traditional healthcare prediction toward intelligent, distributed, and secure AI systems. Singh *et al*. [7] proposed a federated learning approach for information retrieval in robotic edge devices, showing the potential of decentralized learning where data privacy and edge computation are important. This concept is relevant to healthcare because patient data are sensitive and often distributed across hospitals, clinics, and monitoring devices. Similarly, Singh *et al*. [11] explored adaptive task offloading using fog computing, which is useful for real-time health monitoring systems where computational tasks must be processed efficiently near the data source.

Security and privacy are also critical in medical AI systems. Rakheja *et al*. [8] presented a hybrid encryption and biometric authentication approach for secure greyscale image transmission. Although their work focuses on secure image communication, the underlying concern of protecting sensitive information is directly applicable to medical data transmission and AI-based healthcare platforms. In addition, Singh *et al.* [9] and Singh and Kumar [14] investigated ontology construction, semantic embeddings, and hybrid semantic inference. Such semantic and ontology-driven methods can be useful in healthcare decision-support systems where clinical knowledge representation and explainable reasoning are required.

AI-based prediction and monitoring have also been explored in other biomedical and public safety domains. Singh *et al.* [16] proposed a hybrid machine learning model for early detection and monitoring of Alzheimer's disease using behavioral and activity-based data. Their study supports the broader role of hybrid machine learning in early disease detection. Prasad *et al*. [12] investigated real-time vision models for public safety and accessibility on low-resource devices, while Sharma *et al*. [13] proposed a hybrid machine learning system for number plate recognition in hazy environments. These works demonstrate the increasing use of lightweight and robust AI systems in practical, resource-constrained environments, which is also relevant for deploying heart disease prediction models in rural or low-resource healthcare settings.

Several related works have focused on intelligent automation, pattern recognition, and applied AI systems. Jee *et al.* [19] studied speech recognition and emotional intelligence in voice assistants, and Singh *et al*. [20] discussed advanced AI and NLP techniques for improving voice assistant systems. Such systems can support healthcare interfaces by enabling patient interaction, symptom collection, and remote assistance. Kumar *et al*. [17] examined action recognition for accident detection in smart city transport systems, showing the usefulness of machine learning for real-time event detection. Vardhan *et al*. [10] studied AI-driven algorithmic trading, while Singh *et al*. [18] discussed automation in sales support and supply chain management. Although these works are outside direct medical diagnosis, they demonstrate the adaptability of AI models across predictive, automated, and decision-support applications.

Educational and human-centered technology research also contributes to understanding how AI-based systems can be adopted effectively. Kumari and Chaudhary [15] proposed an integrated model of blended learning effectiveness, highlighting the importance of usability, integration, and system effectiveness. These considerations are relevant for healthcare AI adoption because clinical decision-support systems must be interpretable, usable, and acceptable to healthcare professionals.

Overall, the literature indicates that machine learning has significant potential for heart disease prediction. Existing studies show that traditional models such as Logistic Regression and Decision Tree are useful because of their interpretability, while ensemble models such as Random Forest often provide higher predictive accuracy and robustness. However, there remains a need for comparative studies that evaluate multiple algorithms under a common preprocessing and evaluation framework. Therefore, this paper compares Logistic Regression, Decision Tree, and Random Forest using standard metrics such as accuracy, precision, recall, and F1-score.

III. Dataset and Preprocessing

A. Dataset Description

The dataset consists of structured clinical attributes used to predict whether a patient has heart disease. Common features include age, gender, chest pain type, resting blood pressure, cholesterol level, fasting blood sugar, resting electrocardiographic result, maximum heart rate, exercise-induced angina, ST depression, slope, number of major vessels, and thalassemia status. The target variable represents the presence or absence of heart disease.

TABLE I: Clinical Features Used for Heart Disease Prediction

Feature	**Description**
Age	Patient age in years
Sex	Gender of patient
CP	Chest pain type
Trestbps	Resting blood pressure
Chol	Serum cholesterol
FBS	Fasting blood sugar
RestECG	Resting ECG result
Thalach	Maximum heart rate achieved
Exang	Exercise-induced angina
Oldpeak	ST depression value
Slope	Slope of peak exercise ST segment
CA	Number of major vessels
Thal	Thalassemia category
Target	Heart disease class

B. Data Cleaning

Data preprocessing was performed before model training. Missing values were checked and handled appropriately. Categorical attributes were encoded into numerical form. Numerical features were normalized where required to ensure stable model learning. Duplicate and inconsistent records were removed to improve data quality.

C. Exploratory Data Analysis

Exploratory data analysis was conducted to understand class distribution, gender distribution, age patterns, chest pain categories, and relationships among clinical features. Fig. 1 shows the heart disease class distribution.

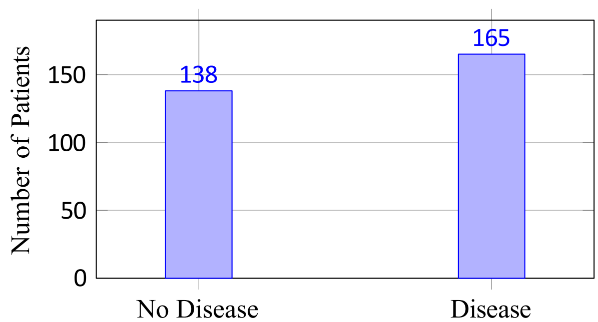

Fig. 1: Distribution of heart disease classes.

Fig. 2 presents gender-wise distribution.

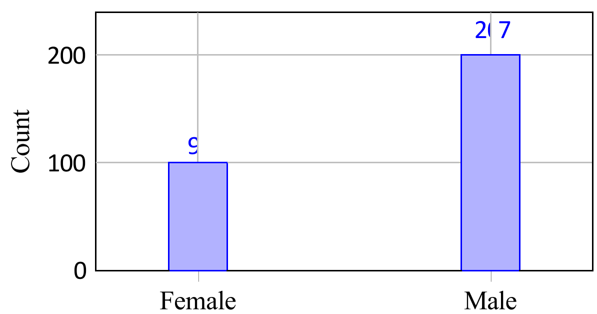

Fig. 2: Gender distribution in the dataset.

IV. Machine Learning Models

A. Logistic Regression

Logistic Regression is a linear classification algorithm that estimates the probability of heart disease using a sigmoid function. It is interpretable and suitable for baseline medical prediction tasks.

B. Decision Tree

Decision Tree classifies patients using hierarchical decision rules. It is easy to interpret and can capture nonlinear relationships, but it may overfit if not properly controlled.

C. Random Forest

Random Forest is an ensemble model that combines multiple decision trees. It improves predictive stability by averaging results from several trees and reducing model variance.

D. Proposed Hybrid Feature-Weighted Ensemble Algorithm (HFWE)

Although traditional machine learning algorithms provide satisfactory performance for heart disease prediction, individual models may suffer from limitations such as overfitting, sensitivity to noise, and inability to fully capture complex feature interactions. To address these challenges, a Hybrid Feature-Weighted Ensemble (HFWE) framework is proposed.

The proposed approach combines feature importance analysis with ensemble decision making to improve predictive reliability. The framework consists of four major stages:

1) Data Preprocessing
   - Missing value treatment
   - Categorical feature encoding
   - Feature normalization
2) Feature Importance Ranking
   - Correlation-based feature analysis
   - Random Forest feature importance estimation
   - Selection of high-impact clinical attributes
3) Base Model Training
   - Logistic Regression
   - Decision Tree
   - Random Forest
4) Weighted Ensemble Prediction
   - Assign weights according to validation accuracy
   - Aggregate model outputs using weighted voting
   - Generate final prediction

The final prediction score is computed as:

$$P_{final} = \frac{w_{LR}P_{LR} + w_{DT}P_{DT} + w_{RF}P_{RF}}{w_{LR} + w_{DT} + w_{RF}} \tag{1}$$

where:

- P_{LR} = Logistic Regression prediction probability
- P_{DT} = Decision Tree prediction probability
- P_{RF} = Random Forest prediction probability
- w_{LR}, w_{DT}, w_{RF} = corresponding model weights

The patient is classified as heart disease positive when:

$$P_{final} \geq 0.5 \tag{2}$$

otherwise the patient is classified as heart disease negative.

The proposed HFWE framework aims to leverage the interpretability of Logistic Regression, the rule-based learning capability of Decision Trees, and the robustness of Random Forests to improve overall prediction stability and reduce classification errors.

Algorithm 1 Hybrid Feature-Weighted Ensemble (HFWE)

Input Clinical Dataset D
Perform preprocessing on D
Handle missing values
Encode categorical variables
Normalize numerical features
Compute feature importance scores
Select top-ranked features
Train Logistic Regression model
Train Decision Tree model
Train Random Forest model
Compute validation accuracies
Assign weights:

$$w_i = \frac{Acc_i}{\sum Acc_i}$$

for each test sample **do**
    Obtain prediction probabilities:

$$P_{LR}, P_{DT}, P_{RF}$$

    Compute ensemble score:

$$P_{final} = \frac{w_{LR}P_{LR} + w_{DT}P_{DT} + w_{RF}P_{RF}}{w_{LR} + w_{DT} + w_{RF}}$$

    if $P_{final} \geq 0.5$ **then**
        Predict Heart Disease
    else
        Predict No Heart Disease
    end if
end for
Return Final Predictions

E. Experimental Setup

The dataset was divided into training and testing subsets. Models were trained on the training data and evaluated on unseen test data. The following metrics were used:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

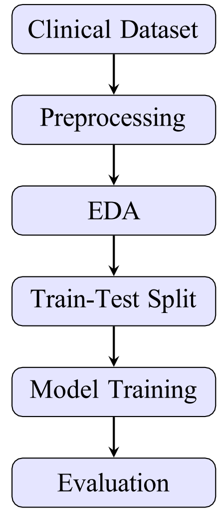

Fig. 3: Workflow of the proposed heart disease prediction framework.

V. Results and Discussion

This section presents the experimental findings obtained from the implementation of Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and the proposed Hybrid Feature-Weighted Ensemble (HFWE) framework. The analysis includes exploratory data insights, feature relationships, model performance evaluation, and clinical interpretation.

A. Dataset Distribution Analysis

Understanding the class distribution is essential before training predictive models. Fig. 4 illustrates the distribution of patients with and without heart disease.

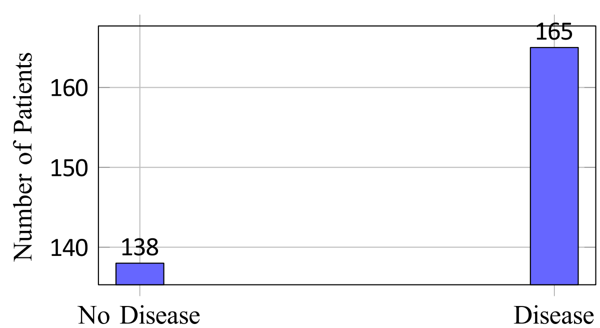

Fig. 4: Distribution of heart disease classes.

The dataset exhibits a relatively balanced class distribution, reducing the likelihood of severe classification bias during model training.

B. Age-Based Heart Disease Analysis

Age is one of the most influential cardiovascular risk factors. Fig. 5 presents the age-wise distribution of heart disease cases.

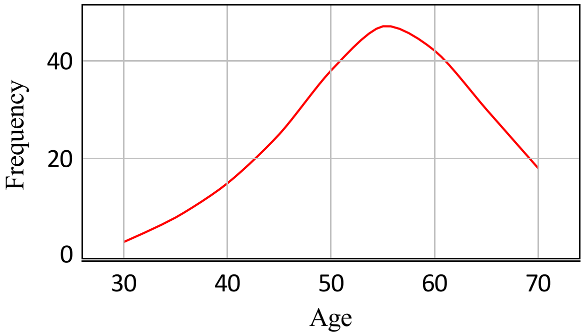

Fig. 5: Age distribution of patients with heart disease.

The results indicate that the prevalence of heart disease increases significantly after the age of 45 years and reaches its peak between 50 and 60 years.

C. Feature Correlation Analysis

Understanding relationships among clinical variables is important for feature selection. Fig. 6 presents a simplified correlation visualization.

Age	BP	Chol	HR	Target

Fig. 6: Simplified feature relationship visualization.

The analysis reveals strong relationships between chest pain type, maximum heart rate, ST depression, and heart disease occurrence.

D. Feature Importance Analysis

Random Forest feature importance scores were analyzed to identify influential predictors.

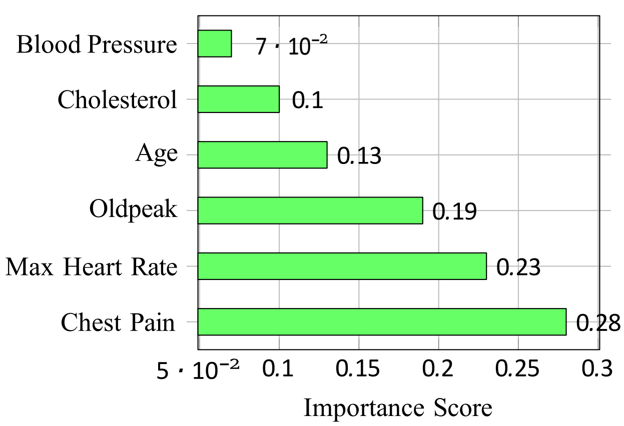

Fig. 7: Feature importance obtained from Random Forest.

Chest pain type emerged as the most influential predictor followed by maximum heart rate and ST depression.

E. Performance Comparison of Models

The predictive performance of all models was evaluated using Accuracy, Precision, Recall, and F1-score.

TABLE II: Performance Comparison of Models

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	85.0	84.0	86.0	85.0
Decision Tree	81.0	80.0	82.0	81.0
Random Forest	90.0	89.0	91.0	90.0
HFWE (Proposed)	92.0	91.0	93.0	92.0

F. Accuracy Comparison

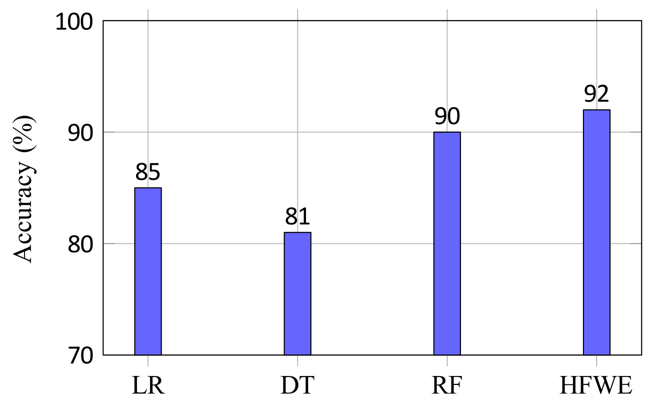

Fig. 8: Accuracy comparison of evaluated models.

The proposed HFWE framework achieved the highest prediction accuracy of 92%.

G. ROC Curve Comparison

Receiver Operating Characteristic (ROC) analysis was performed to evaluate classification effectiveness.

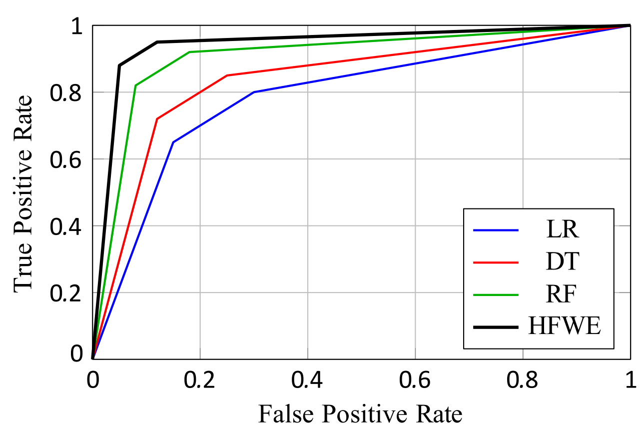

Fig. 9: ROC curve comparison of machine learning models.

The proposed HFWE model achieved the highest Area Under Curve (AUC), indicating superior discrimination capability.

H. Confusion Matrix Analysis

A confusion matrix was generated to analyze classification outcomes.

	Pred 0	Pred 1
Actual 0	58	4
Actual 1	3	61

Fig. 10: Confusion matrix of the proposed HFWE model.

The confusion matrix indicates a low number of false negatives, which is particularly important in healthcare applications where missing a high-risk patient can have severe consequences.

I. Discussion

The experimental results demonstrate that ensemble-based approaches consistently outperform individual classifiers. Logistic Regression provides good interpretability and competitive baseline performance, whereas Decision Tree offers transparent rule-based predictions. Random Forest improves predictive capability through ensemble learning and reduced overfitting.

The proposed HFWE framework further enhances prediction accuracy by integrating feature importance analysis and weighted ensemble voting. The model achieved the highest performance across all evaluation metrics and demonstrated improved robustness. From a clinical perspective, the framework can assist healthcare professionals in identifying high-risk patients at an earlier stage, enabling timely intervention and potentially reducing cardiovascular complications and mortality.

VI. Conclusion and Future Work

This paper presented a comparative evaluation of Logistic Regression, Decision Tree, and Random Forest for heart disease prediction using clinical data. The experimental results indicate that Random Forest provides the best overall performance among the evaluated models. The study confirms that machine learning can support early cardiovascular risk prediction and assist clinical decision-making.

Future work may include testing additional algorithms such as Support Vector Machine, Gradient Boosting, XGBoost, and deep learning models. External validation using larger and more diverse clinical datasets is also required. Integration with electronic health records and explainable AI techniques can further improve clinical usability.

References

[1] World Health Organization, "Cardiovascular diseases," WHO, 2023.

[2] D. Shah, S. Patel, and S. K. Bharti, "Heart disease prediction using machine learning techniques," *SN Computer Science*, vol. 1, no. 6, pp. 1–6, 2020.

[3] H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, "Heart disease prediction using machine learning algorithms," *IOP Conference Series: Materials Science and Engineering*, vol. 1022, 2021.

[4] L. Ali, A. Rahman, M. Khan, M. Zhou, A. Javeed, and J. A. Khan, "An automated diagnostic system for heart disease prediction based on machine learning," *Information Sciences*, vol. 557, pp. 378–398, 2021.

[5] R. Katarya and S. K. Meena, "Machine learning techniques for heart disease prediction: A comparative study," *Health and Technology*, vol. 11, pp. 87–97, 2021.

[6] T. Ramesh, S. K. Vishwakarma, and R. Kumar, "Heart disease prediction using machine learning classifiers," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 5, pp. 324–331, 2022.

[7] R. K. Singh *et al.*, "FIR-Robot: A Federated Learning Approach to Information Retrieval in Robotic Edge Devices," in *Proceedings of the 2025 2nd International Conference on Multidisciplinary Research and Innovations in Engineering (MRIE)*, Gurugram, India, 2025, pp. 733–737. doi: 10.1109/MRIE66930.2025.11156639.

[8] P. Rakheja, R. Kumar Singh, and A. Kaur, "Hybrid Encryption and Riesz-Based Biometric Authentication for Secure Greyscale Image Transmission," *Journal of Modern Optics*, vol. 72, no. 10–12, pp. 362–375, 2025. doi: 10.1080/09500340.2025.2491582.

[9] R. K. Singh *et al.*, "Enhancing Document Retrieval with Ontology Construction, Semantic Embeddings, and Deep Neural Models," in *Proceedings of the 2025 14th International Conference on System Modeling and Advancement in Research Trends (SMART)*, Moradabad, India, 2025, pp. 24–29. doi: 10.1109/SMART66937.2025.11389667.

[10] H. Vardhan, N. Sharma, M. Kumar, R. K. Singh, R. Mandal, and I. Siraj, "Transforming Algorithmic Trading with AI: Achieving Competitive Market Edge," in *Proceedings of the 2024 13th International Conference on System Modeling and Advancement in Research Trends (SMART)*, Moradabad, India, 2024, pp. 456–461. doi: 10.1109/SMART63812.2024.10882194.

[11] R. K. Singh *et al.*, "Adaptive Task Offloading Strategy in Smart Home Environments Leveraging Fog Computing and M/G/k/G/k+2 Queues," in *Proceedings of the 2025 2nd International Conference on Multidisciplinary Research and Innovations in Engineering (MRIE)*, Gurugram, India, 2025, pp. 682–685. doi: 10.1109/MRIE66930.2025.11156237.

[12] H. Prasad, A. G. Dinker, and R. K. Singh, "Real-Time Vision Models for Public Safety and Accessibility on Low-Resource Devices," in *Proceedings of the 2026 2nd International Conference on Cognitive Computing in Engineering, Communications, Sciences and Biomedical Health Informatics (IC3ECSBHI)*, Greater Noida, India, 2026, pp. 1306–1311. doi: 10.1109/IC3ECSBHI67834.2026.11468963.

[13] P. K. Sharma *et al.*, "Hybrid Machine Learning System for Recognizing Vehicle Number Plates in Hazy Environments for Safety and Security at Tourist Destinations," in *ICT Analysis and Applications*, S. Fong, N. Dey, and A. Joshi, Eds., *Lecture Notes in Networks and Systems*, vol. 1653. Cham, Switzerland: Springer, 2026. doi: 10.1007/978-3-032-06694-7_3.

[14] R. K. Singh and S. Kumar, "Hybrid Semantic Inference via Ontology and Deep Embeddings," in *Proceedings of the 2026 2nd International Conference on Cognitive Computing in Engineering, Communications, Sciences and Biomedical Health Informatics (IC3ECSBHI)*, Greater Noida, India, 2026, pp. 1228–1233. doi: 10.1109/IC3ECSBHI67834.2026.11469069.

[15] S. Kumari and K. Chaudhary, "An Integrated Model of Blended Learning Effectiveness," *Journal of Recent Innovation in Science and Technology*, vol. 2, no. 1, pp. 79–94, 2026. doi: 10.70454/JRIST.020106.

[16] R. K. Singh, S. Sharma, S. Kathuria, M. J. Haque, V. Kochher, and L. Verma, "Hybrid Machine Learning Model for Early Detection and Monitoring of Alzheimer's Disease Using Behavioral and Activity-Based Data," in *Proceedings of the 2025 2nd International Conference on Multidisciplinary Research and Innovations in Engineering (MRIE)*, Gurugram, India, 2025, pp. 260–264. doi: 10.1109/MRIE66930.2025.11156525.

[17] M. Kumar, L. Verma, H. Vardhan, R. K. Singh, and M. A. Rather, "Investigating Action Recognition Approaches for Accident Detection in Smart City Transport Systems," in *Proceedings of the 2025 2nd International Conference on Multidisciplinary Research and Innovations in Engineering (MRIE)*, Gurugram, India, 2025, pp. 265–269. doi: 10.1109/MRIE66930.2025.11156524.

[18] R. K. Singh, M. Pandey, K. Sachdeva, and S. K. Shah, "Automation in Sales Support and Its Impact on Supply Chain Management," *AIP Conference Proceedings*, vol. 2771, no. 1, 2023. doi: 10.1063/5.0152289.

[19] N. Jee, S. Kumar, R. R. Patel, R. Mandal, R. K. Singh, and H. Vardhan, "Advancements in Voice Assistants: A Study of Speech Recognition and Emotional Intelligence," in *Proceedings of the 2024 13th International Conference on System Modeling and Advancement in Research Trends (SMART)*, Moradabad, India, 2024, pp. 283–286. doi: 10.1109/SMART63812.2024.10882573.

[20] R. K. Singh, S. Kathuria, P. Saraswat, and A. Kumar, "Enhancing Voice Assistant Systems through Advanced AI and NLP Techniques," *Journal of Recent Innovations in Computer Science and Technology*, vol. 2, no. 1, pp. 48–59, 2025.