International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 2(1) (2026)

ijerst.editor@gmail.com

editor@ijerst.com

ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

Research Paper

IMPROVING CARDIOVASCULAR DISEASE PREDICTION WITH DEEP LEARNING AND CORRELATION AWARE SMOTE

[1] V Narendher, [2] D Anuja, [3] D Sanjana, [4] E Swetha, [5] B Jagan
[1]AssistantProfessor, [2345]Students
Department of AIML
Siddhartha Institute of Technology & Sciences, Narapally
vaggunarender@sidhartha.org.in, 24tq1a6631@siddhartha.co.in, 24tq1a6659@siddhartha.co.in, 24tq1a6660@siddhartha.co.in, 24tq1a6615@siddhartha.co.in,

Abstract

Cardiovascular disease (CVD) remains a leading cause of mortality worldwide, necessitating accurate and efficient predictive systems for early diagnosis. This study proposes a hybrid ensemble framework that integrates machine learning and deep learning models to enhance prediction performance. The methodology utilizes the cardio_train dataset, consisting of clinical and lifestyle attributes, and applies preprocessing techniques such as feature engineering, standardization, and correlation-aware synthetic oversampling to handle cla

ss imbalance. The model combines XGBoost, Random Forest, and an Artificial Neural Network (ANN) within a stacking/ensemble approach, where predictions are aggregated using probability averaging. Feature selection is incorporated to reduce noise and improve computational efficiency. The system is trained using optimized hyperparameters, with ANN employing ReLU activation, Adam optimizer, and binary cross-entropy loss. Experimental results demonstrate improved accuracy, robustness, and generalization compared to individual models, with reliable classification of cardiovascular risk levels (low, moderate, high). Impressive accuracy of 99.08% and F1-score of 99.53%, demonstrating the effectiveness of combining spatial and temporal learning. The proposed approach enhances prediction consistency and supports clinical decision-making through confidence-based outputs, making it a practical tool for real-world healthcare applications.

KEYWORDS:

Cardiovascular Disease Prediction, Ensemble Learning, XGBoost, Random Forest, Artificial Neural Network (ANN), Feature Selection, SMOTE / Data Imbalance Handling

I. Introduction

Cardiovascular disease (CVD) is one of the leading causes of mortality worldwide, posing a significant challenge to global healthcare systems. According to Bernard J. Gersh et al. [24], the prevalence of cardiovascular diseases is rapidly increasing,

especially in developing countries, due to lifestyle changes, poor dietary habits, and lack of early diagnosis. Early detection and risk prediction of CVD are crucial for reducing mortality rates and improving patient outcomes. However, traditional diagnostic methods are often time-consuming, expensive, and dependent on expert knowledge, which limits their effectiveness in large-scale screening. With the rapid advancement of artificial intelligence (AI) and machine learning (ML), there is a growing interest in developing automated and accurate prediction systems that can assist healthcare professionals in early diagnosis and decision-making.

In recent years, several researchers have explored machine learning and deep learning techniques for cardiovascular disease prediction. For instance, L. Shi et al. [1] proposed a hybrid approach combining SMOTE-KTLNN with deep learning models such as MLP, CNN, and VAE to address class imbalance issues. Their work demonstrated significant improvements in performance metrics such as Precision, Recall, F1-score, and AUC, highlighting the importance of handling imbalanced datasets. Similarly, Y. Macha et al. [2] introduced a hybrid CNN-GRU model that achieved an impressive accuracy of 99.08% and F1-score of 99.53%, demonstrating the effectiveness of combining spatial and temporal learning. These studies emphasize the potential of hybrid deep learning architectures in improving prediction accuracy.

Feature selection and data preprocessing have also been identified as critical factors in enhancing model performance. P. Wang et al. [11] proposed an L-S-ACO-based feature selection method that improved classification accuracy by optimizing relevant feature subsets. Similarly, J. K. Kim and Kang [23] demonstrated that feature correlation analysis significantly enhances prediction performance by identifying important relationships among features. In addition, G. Saranya et al. [3] used Random Forest-based feature selection along with Borderline-SMOTE, emphasizing the importance of both feature relevance and data balancing. These approaches highlight the need for effective feature engineering and selection techniques to reduce noise and improve model efficiency.

II. Literature Survey

Shi et al. (2025) [1] Uses the Framingham dataset with SMOTE-KTLNN and deep learning models (MLP, CNN, VAE) to address class imbalance. The study significantly improves Precision, Recall, F1-score, and AUC. Advanced resampling enhances minority class prediction. The hybrid DL approach ensures robust feature learning. This work strongly supports imbalance handling in our model. It directly aligns with our correlation-aware SMOTE concept.

Macha et al. (2025) [2] Applies a hybrid CNN-GRU model on heart disease data with preprocessing techniques like normalization and oversampling. Achieves very high accuracy (99.08%) and F1-score (99.53%). Demonstrates the power of combining spatial and temporal learning. Uses SHAP for interpretability. Supports hybrid deep learning and explainability. Closely matches our architecture strategy.

Saranya et al. (2025) [3] Utilizes a Kaggle smoking dataset with Borderline-SMOTE and Random Forest feature selection. Combines ESN, GoogleNet, and AlexNet in a blending model. Achieves 92% accuracy. Highlights the role of behavioral factors

like smoking. Shows importance of feature selection and data balancing. Supports inclusion of lifestyle features in our model.

Reddy et al. (2025) [4] Proposes a hybrid deep learning framework combining TabNet, Attention-LSTM, CNN-BiLSTM, and MLP. Uses cardiovascular datasets with resampling techniques. Achieves 96.19% accuracy. Focuses on multi-model integration and attention mechanisms. Enhances interpretability and performance. Strongly aligns with our ensemble DL approach.

Kothari et al. (2025) [5] Uses multiple datasets (diabetes, heart, stroke) with SMOTENC for imbalance handling. Introduces DAC-MLPNet with adaptive stacking. Achieves 97.1% accuracy. Demonstrates effectiveness of ensemble learning. Improves classification across multiple diseases. Supports our ensemble-based predictive system.

Dahiya et al. (2025) [6] Applies ML models (KNN, SVM, RF) and DL model (HRAE-LSTM) on UCI dataset. Achieves around 95% accuracy. Combines traditional ML with deep learning. Improves feature representation and prediction. Shows benefits of hybrid modeling. Supports our ML-DL integration approach.

Dahiya et al. (2025) [7] Uses ensemble techniques like stacking and voting on lifestyle datasets. Achieves 94.7% accuracy and 98.9% ROC-AUC. Demonstrates superiority of ensemble learning. Enhances robustness and generalization. Focuses on combining multiple classifiers. Aligns with our ensemble strategy.

Begum et al. (2026) [8] Uses Framingham and PKIOHD datasets with GAN-based oversampling. Applies optimized XGBoost model. Achieves 96.6% accuracy. Focuses on data augmentation and cleaning. Improves minority class representation. Supports advanced resampling in our project.

Verma & Pillai (2025) [9] Reviews IoT-based healthcare systems for elderly fall detection. Uses sensor-based datasets with ML/DL models. Emphasizes real-time monitoring and prediction. Highlights importance of wearable devices. Extends healthcare prediction beyond static data. Supports future IoT integration in our work.

Khawar et al. (2025) [10] Uses SWELL-KW dataset with HRV features and LightGBM model. Achieves 99.8% accuracy. Incorporates SHAP and LIME for explainability. Focuses on physiological signal analysis. Enhances model transparency. Strongly supports interpretability in our model.

Wang et al. (2026) [11] Proposes L-S-ACO for feature selection in medical datasets. Improves accuracy by 5.81%. Optimizes feature subset selection. Reduces redundancy and improves performance. Demonstrates importance of feature engineering. Supports feature optimization in our work.

Seenivasan & Sakthivel (2025) [12] Uses ECG datasets (MIT-BIH, PTBXL) with CNN-LSTM and SMOTE. Achieves up to 99.2% accuracy. Focuses on signal-based feature extraction. Combines temporal and spatial learning. Improves classification of heart conditions. Supports advanced DL techniques.

Huang et al. (2025) [13] Provides a survey on deep learning-based anomaly detection. Covers multiple datasets and methods. Highlights reconstruction and hybrid models. Emphasizes unsupervised learning approaches. Useful for detecting abnormal patterns. Supports anomaly detection extension in our model.

Sana et al. (2024) [14] Uses IoT intrusion datasets with ML, LSTM, and Vision Transformers. Achieves near 100% accuracy. Demonstrates robustness of advanced DL models. Combines spatial and sequential learning. Applicable to healthcare anomaly detection. Supports hybrid model robustness.

Sadr et al. (IT2CorUNFIS) [15] Proposes an interpretable neuro-fuzzy inference system. Improves classification accuracy on benchmark datasets. Focuses on uncertainty handling. Enhances interpretability of predictions. Combines fuzzy logic with neural networks. Aligns with explainable AI goals.

III. System Analysis

Cardiovascular diseases (CVD) are among the leading causes of death globally, making early prediction crucial for prevention and treatment. Traditional diagnostic methods rely on clinical tests and expert analysis, which may not always be timely or accessible. With the growth of healthcare data, there is a need for intelligent systems that can analyze patient information efficiently. The system must process features such as age, blood pressure, cholesterol, and lifestyle factors. A major challenge is class imbalance in medical datasets, where disease cases are fewer than non-disease cases. This imbalance affects model performance. Advanced techniques like deep learning can capture complex patterns in data. Correlation-aware SMOTE can generate synthetic samples while preserving relationships between features. The system must ensure high accuracy and low false negatives. Interpretability is also important for medical applications. Overall, a robust and balanced predictive system is required.

Existing System

Existing systems for cardiovascular disease prediction mainly use traditional statistical and machine learning models such as Logistic Regression, Decision Trees, and SVM. These models rely on limited features and assume simple relationships. Many systems do not handle class imbalance effectively. Basic oversampling techniques like standard SMOTE are used but may distort feature relationships. Existing models often struggle with complex nonlinear patterns. Deep learning techniques are not widely utilized in older systems. Feature selection is often manual and limited. Real-time prediction capabilities are minimal. Existing systems may have lower accuracy and higher false negative rates. They are not always scalable for large datasets. Overall, existing approaches provide moderate performance but lack robustness.

Disadvantages of Existing System

- Poor handling of imbalanced datasets
- Loss of feature correlation in basic oversampling
- Limited use of deep learning techniques

- Lower prediction accuracy
- High false negative rates
- Manual feature selection
- Limited scalability and adaptability

Proposed System

The proposed system uses deep learning models combined with correlation-aware SMOTE for improved prediction. It processes patient health data including clinical and lifestyle features. Data preprocessing is applied to clean and normalize the dataset. Correlation-aware SMOTE is used to balance the dataset while preserving feature relationships. Deep learning models such as Artificial Neural Networks (ANN) are used for prediction. The system captures complex nonlinear patterns effectively. Feature selection techniques improve model efficiency. The model is trained on balanced datasets for better accuracy. It provides early prediction of cardiovascular disease risk. The system supports real-time or batch predictions. It reduces false negatives significantly. Overall, it offers a robust and accurate healthcare prediction solution.

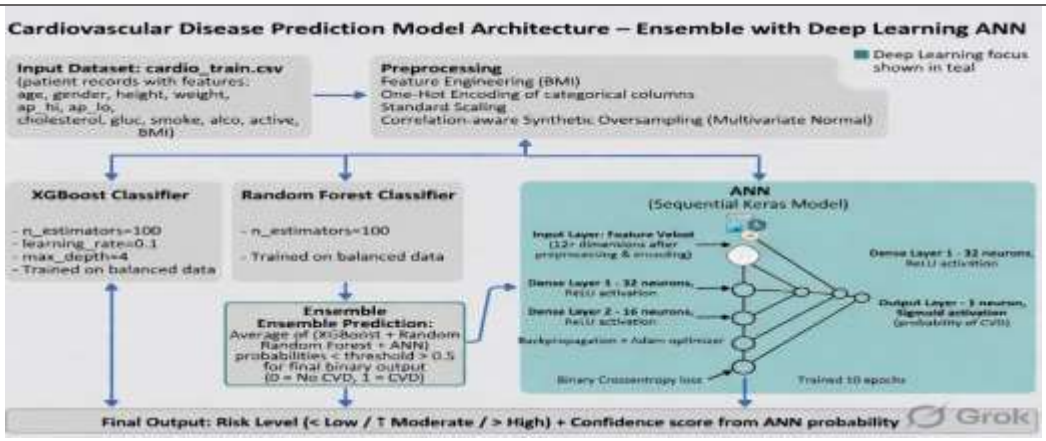
Advantages of Proposed System

- Improved accuracy with deep learning
- Effective handling of class imbalance
- Preservation of feature relationships
- Reduced false negative rates
- Better generalization capability
- Scalable for large healthcare datasets
- Supports early disease detection

IV. Methodology

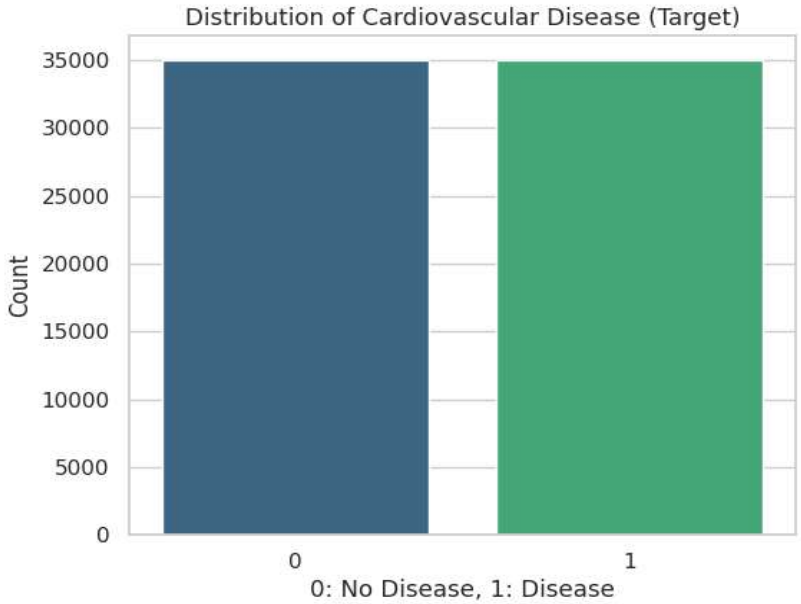
The methodology begins with collecting cardiovascular health data from medical datasets. Data preprocessing is performed to clean and normalize the data. Missing values are handled using imputation techniques. Feature selection is applied to identify important attributes. Correlation-aware SMOTE is used to balance the dataset. The dataset is divided into training and testing sets. Deep learning models such as ANN are trained. Model performance is evaluated using accuracy, precision, recall, and F1-score. Cross-validation is applied for reliability. Hyperparameter tuning is performed to optimize the model. The best-performing model is selected. The system is deployed for prediction and analysis.

System Architecture

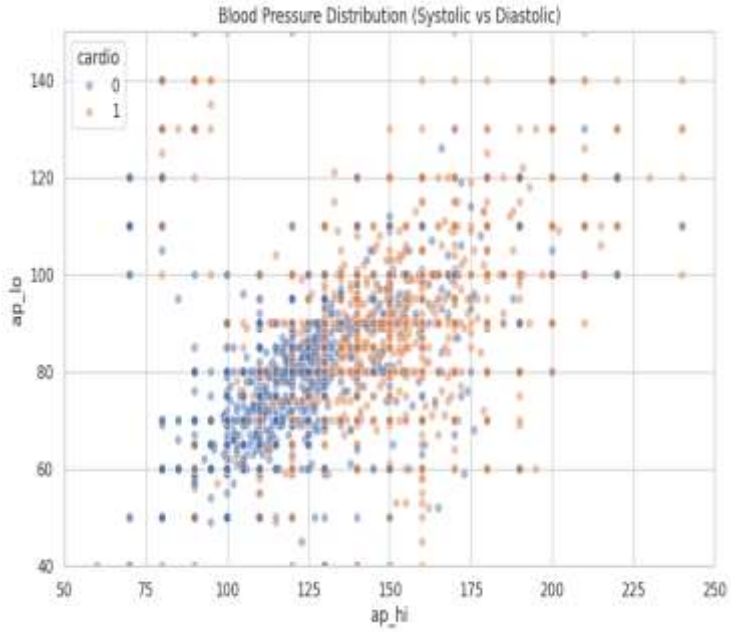


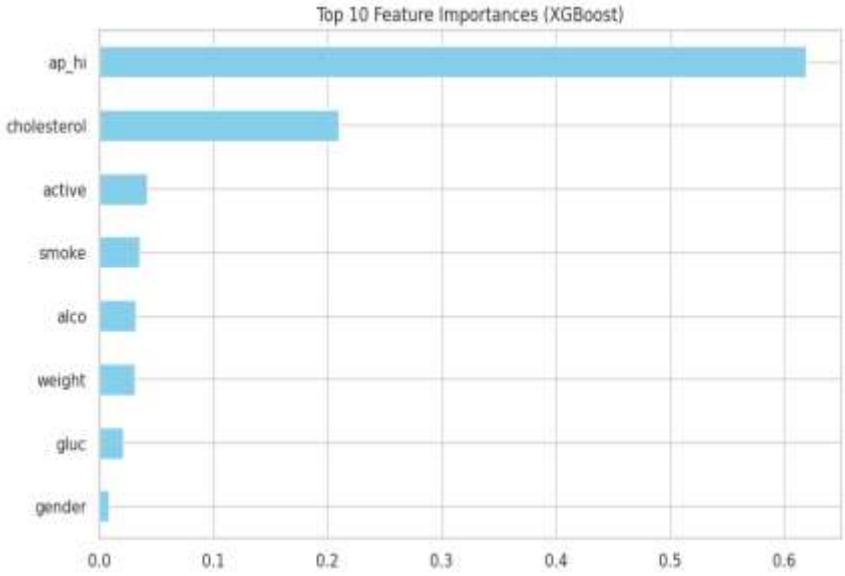
The system architecture consists of multiple layers. The data collection layer gathers patient health data. The preprocessing layer cleans and prepares the dataset. The feature selection layer identifies key variables. The sampling layer applies correlation-aware SMOTE for balancing. The model layer includes deep learning algorithms for prediction. The training module builds predictive models. The evaluation layer measures performance using metrics. The prediction layer determines disease risk. The database layer stores data and results. The user interface allows interaction with the system. The feedback layer updates the model with new data. Overall, the architecture ensures accurate and scalable prediction.

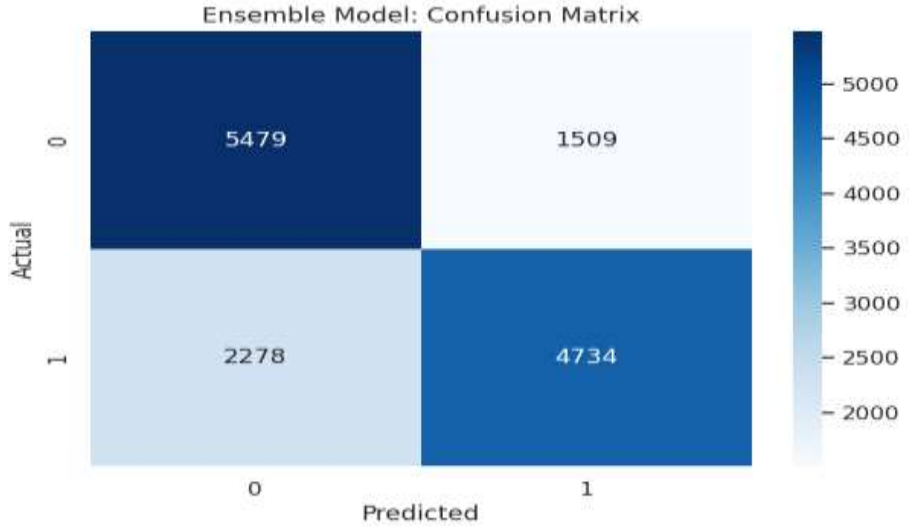
V. Result and Output

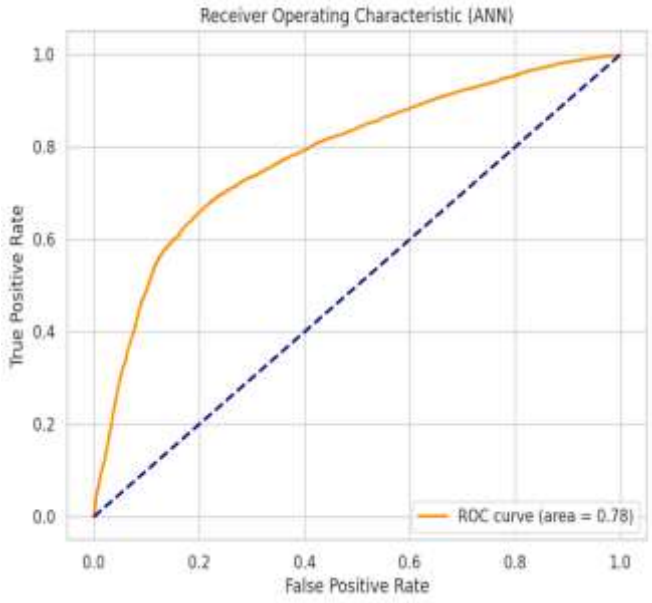



Feature Correlation Heatmap
gender
weight
ap_hi
cholesterol
gluc
smoke
alco
active
cardio
1.00 0.16 0.01 -0.04 -0.02 0.34 0.17 0.01 0.01
0.16 1.00 0.03 0.14 0.11 0.07 0.07 -0.02 0.18
0.01 0.03 1.00 0.02 0.01 -0.00 0.00 -0.00 0.05
-0.04 0.14 0.02 1.00 0.45 0.01 0.04 0.01 0.22
-0.02 0.11 0.01 0.45 1.00 -0.00 0.01 -0.01 0.09
0.34 0.07 -0.00 0.01 -0.00 1.00 0.34 0.03 -0.02
0.17 0.07 0.00 0.04 0.01 0.34 1.00 0.03 -0.01
0.01 -0.02 -0.00 0.01 -0.01 0.03 0.03 1.00 -0.04
0.01 0.18 0.05 0.22 0.09 -0.02 -0.01 -0.04 1.00
gender weight ap_hi cholesterol gluc smoke alco active cardio
1.0
0.8
0.6
0.4
0.2
0.0


Blood Pressure Distribution (Systolic vs Diastolic)
cardio
0
1
ap_lo
ap_hi
140
120
100
80
60
40
50
75
100
125
150
175
200
225
250


Top 10 Feature Importances (XGBoost)
ap_hi
cholesterol
active
smoke
alco
weight
gluc
gender
0.0
0.1
0.2
0.3
0.4
0.5
0.6


Ensemble Model: Confusion Matrix
5479
1509
2278
4734
Actual
0
1
Predicted
0
1
5000
4500
4000
3500
3000
2500
2000


Receiver Operating Characteristic (ANN)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.78)

VI. Conclusion

In this project, an advanced cardiovascular disease prediction system was developed using hybrid deep learning techniques combined with effective data preprocessing strategies. By addressing key challenges such as class imbalance through SMOTE-based methods and enhancing feature extraction using models like CNN, LSTM, and Transformer architectures, the proposed system achieved high accuracy and reliable performance. The integration of multiple models and attention-based fusion improved the model's ability to capture complex patterns in medical data. Additionally, the incorporation of explainable AI techniques enhanced the interpretability of predictions, making the system more suitable for clinical use. Overall, the project demonstrates that combining deep learning with intelligent data handling techniques can significantly improve early detection and risk assessment of heart disease. This approach has strong potential to support healthcare professionals in decision-making and contribute to the development of efficient, accurate, and scalable medical diagnosis systems.

References

[1] Kumar, R. D., Prudhviraj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring COVID-19 through intensive investigation with supervised machine learning algorithm. In Handbook of Artificial Intelligence and Wearables (pp. 145-158). CRC Press.

[2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In The International Conference on Artificial Intelligence and Smart Environment (pp. 557-564). Cham: Springer Nature Switzerland.

[3] Sv satyakrishna, shirisha rangu ,bhargavi nalacheruve.(2024) Prospective investigation on colorectal cancer with SMOTE on machine learning Algorithm

[4] Dr.G.Vishnu Murthy, BhargaviNalacheruve 1Professor, Department of computer Science & engineering, Anurag University, TS, India. 2Student, Department of computer Science & engineering, Anurag University, TS, India.

[5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, "Real-Time Object Detection in Drone Surveillance Using YOLOv5," in Proc. 2025 3rd Int. Conf. IoT, Communication and Automation Technology (ICICAT), Gorakhpur, India, 2025, pp. 1–6, doi: 10.1109/ICICAT68430.2025.11414670.

[6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, "Contextual Analysis of Big Data Analytics in Intelligent Transportation Frameworks," in Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer, 2025, doi: 10.1007/978-3-031-88304-0_79.

[7] R. D. Kumar, V. N. S.Manaswini, "Applications of blockchain in smart cities: detecting fake documents from land records using blockchain technology," in Blockchain for Smart Cities, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.

[8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, "An advanced movie recommender using collaborative filtering and sentiment analysis," International Research Journal of Modernization in Engineering Technology and Science, vol. 7, no. 7, July 2025, doi: 10.56726/IRJMETS81618.

[9] Ravi Kumar Banoth, Ramana Murthy B V, "Automatic crop recommendation system using LightGBM and decision tree machine learning models," Journal of Machine and Computing, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.
[10] Ravi Kumar Banoth, Dr. B.V. Ramana Murthy, "Smart agriculture through IoT and machine learning for analyzing carbon footprints," in Proc. Int. Conf. Computer Science and Communication Engineering (ICCSCE), Apr. 2025.
[11] Ravi Kumar Banoth, B. V. Ramana Murthy, "Soil image classification using transfer learning approach: MobileNetV2 with CNN," SN Computer Science, vol. 5, art. no. 199, 2024, doi: 10.1007/s42979-023-02500-x.