EFFICIENT HEART FAILURE PREDICTION USING MACHINE LEARNING

[1] P Akhila, [2] R Veera Tejaswini, [3] H Venkatesh, [4] P Vedhaarjun, [5] R Dhanush Sai Krishna
[1]AssistantProfessor, [2345]Students
Department of Computer Science and Technology
Siddhartha Institute of Technology & Sciences, Narapally

akhilapotu@siddhartha.org.in, 25TQ5A0516@siddhartha.co.in, 25TQ5A0515@siddhartha.co.in, 24TQ1A05J1@siddhartha.co.in, 24TQ1A05J2@siddhartha.co.in

Abstract

Heart failure is a critical cardiovascular disorder and a leading cause of mortality worldwide, necessitating early and accurate prediction for effective clinical intervention and improved patient outcomes. In this study, an optimized machine learning-based framework is proposed for the prediction and classification of heart failure using structured clinical data. The model is developed using the Extreme Gradient Boosting (XGBoost) algorithm, which is well-known for its robustness, scalability, and ability to handle complex nonlinear relationships among high-dimensional features. To address the issue of class imbalance commonly observed in medical datasets, the Synthetic Minority Over-sampling Technique (SMOTE) is employed to generate synthetic samples for the minority class, thereby ensuring balanced data distribution and improving model sensitivity.

Furthermore, hyperparameter tuning is carried out using GridSearchCV to determine the optimal combination of model parameters, enhancing predictive performance and minimizing error rates. The model is validated using Stratified K-Fold Cross-Validation, which ensures reliability, reduces overfitting, and improves generalization capability across unseen data. The dataset utilized in this study comprises key clinical attributes such as age, blood pressure, cholesterol levels, and other relevant medical indicators. Comprehensive data preprocessing techniques, including normalization and feature scaling, are applied to improve model efficiency and convergence.

I. Introduction

Heart failure is one of the most severe and life-threatening cardiovascular diseases, contributing significantly to global mortality and morbidity rates. According to recent medical studies, cardiovascular diseases account for a substantial proportion of deaths worldwide, with heart failure being a major contributor due to its complex nature and late-stage diagnosis. Early detection of heart failure plays a crucial role in improving patient survival rates, reducing healthcare costs, and enabling timely medical intervention. However, traditional diagnostic methods often rely on clinical expertise, medical imaging, and laboratory tests, which can be time-consuming, costly, and sometimes prone to human error. Therefore, there is an increasing need for intelligent, automated systems that can assist healthcare professionals in making accurate and efficient predictions.

In recent years, Machine Learning (ML) has emerged as a powerful tool in the field of healthcare analytics, offering data-driven solutions for disease prediction and diagnosis. ML algorithms can analyze large volumes of patient data, identify hidden

patterns, and generate predictive models with high accuracy. Various machine learning techniques such as Decision Trees, Support Vector Machines, Random Forests, and Logistic Regression have been widely used for heart disease prediction. However, these traditional models often face limitations in handling complex nonlinear relationships, high-dimensional datasets, and imbalanced class distributions, which are common challenges in medical datasets.

To overcome these limitations, advanced ensemble learning techniques such as Extreme Gradient Boosting (XGBoost) have gained significant attention. XGBoost is an optimized gradient boosting algorithm that provides high performance, scalability, and regularization capabilities, making it particularly suitable for medical prediction tasks. It effectively handles missing values, reduces overfitting, and captures intricate relationships between features, thereby improving predictive accuracy. Despite its advantages, the performance of XGBoost heavily depends on appropriate hyperparameter tuning and data preprocessing techniques.

II. Literature Survey

Zhang et al. |Ref [1]|, “Advancements in hybrid machine learning models for biomedical disease classification using integration of hyperparameter-tuning and feature selection” (DOI: 10.1007/s11831-025-10491-6) used public ECG datasets for heart disease classification. The study applied outlier detection, PCA for feature extraction, and EBRO-optimized Random Forest. The model achieved 93.56% accuracy. This work is highly relevant to our idea as it combines feature selection and optimization to improve prediction accuracy.

Wang et al. |Ref [2]|, “Development of a web platform for predicting fall risk in cardiovascular patients using machine learning” used the China Health and Retirement Longitudinal Study dataset (1784 samples). The method involved LASSO feature selection and LightGBM classifier with SHAP explainability. It achieved AUC of 0.839. This relates to our idea by demonstrating real-world ML deployment.

Ahmad et al. |Ref [3]|, “Interpretable machine learning framework for detection of cardiovascular diseases from real-time ECG signals” used 24,894 real-time ECG records from SKIMS hospital. The method combined 3PGA feature optimization with Gradient Boosting and SHAP/LIME explainability. It achieved 99.70% accuracy. This strongly supports our idea of interpretable ML models.

Kumar et al. |Ref [4]|, “Hybrid feature optimization and radial basis function networks for cardiovascular disease prediction” used clinical cardiovascular datasets (e.g., Cleveland-type data). The method applied Harris Hawks Optimization (HHS) + RBF Neural Network. It achieved 92.1% accuracy. This aligns with our hybrid optimization approach.

Singh et al. |Ref [5]|, “Quantum Neural Networks and Classical Machine Learning for Cardiovascular Disease Risk Prediction” reviewed multiple benchmark datasets across 12 studies. Methods included QNN and classical ML comparisons. Results showed QNN potential but limitations. This supports future enhancement in our project.

Demir et al. |Ref [6]|, "Hybrid metaheuristic optimized deep CNN for heart disease prediction" used a merged dataset (Statlog, Cleveland, Switzerland, Hungary – 1190 records). Methods included CNN optimized with GWO, WOA, AOA + SMOTE balancing. It achieved 97.42% accuracy. This strongly aligns with our hybrid deep learning idea.

Patel et al. |Ref [7]|, "HPML-CVD: hyperparameter tuned machine learning model" used the UCI Heart Disease dataset (303 samples). Methods included Chi-square feature selection + hyperparameter tuning of RF, SVM, KNN, DT. It achieved 93.41% accuracy (RF). This directly supports our ML approach.

Reddy et al. |Ref [8]|, "Adaptive Feature Weighting Framework" used Cleveland, Statlog, and combined UCI datasets. The method applied WFMM feature selection with multiple classifiers. It improved accuracy by 4–9%. This supports feature engineering in our model.

Chen et al. |Ref [9]|, "Explainable ML with data balancing for CVD prediction" used Heart Disease Classification (HDC) and Cardiovascular datasets. Methods included SMOTE + CatBoost + SHAP/LIME explainability. It achieved 99.44% accuracy. This aligns with our explainable AI approach.

Alqahtani et al. |Ref [10]|, "OptGPDCNN framework on IoT platform" used four benchmark cardiovascular datasets. Methods included deep CNN, CapsNet, EAFT feature selection, MCSA optimization, GAN-based balancing. It achieved 99.56% accuracy. This supports real-time intelligent systems.

Sharma et al. |Ref [11]|, "Mayfly algorithm for cardiovascular classification" used five real-time cardiovascular datasets. Methods included Mayfly optimization + RF, SVM, KNN classifiers. It achieved 90–95% accuracy. This supports optimization-based feature selection.

Garcia et al. |Ref [12]|, "Two-tier classification framework" used real-world cardiovascular datasets. Methods included CNN + CatBoost + LightGBM with metaheuristic optimization. It achieved ~92% accuracy. This aligns with hybrid model architecture.

Verma et al. |Ref [13]|, "Explainable ML approaches using UCI dataset" used the UCI Heart Disease dataset. Methods included CatBoost, XGBoost with oversampling and SHAP explainability. It showed superior performance of ensemble models. This supports our approach.

Ali et al. |Ref [14]|, "Stacking ensemble model for heart disease diagnosis" used a public heart disease dataset. Methods included DT, KNN, GNB + Logistic Regression stacking + SMOTE + Optuna tuning. It achieved 99.61% accuracy. This validates ensemble learning.

Lopez et al. |Ref [15]|, "Grid, Random, Bayesian optimization comparison" used UCI ML repository datasets. Methods compared Grid Search, Random Search, Bayesian

Optimization. It showed optimization improves accuracy significantly. This guides our tuning strategy.

III. System Analysis

The system is designed to predict the likelihood of heart failure using machine learning techniques applied to clinical and patient data. Heart conditions such as Heart Failure are a major cause of mortality worldwide, requiring early and accurate diagnosis. The system analyzes various health parameters including age, blood pressure, cholesterol levels, heart rate, and medical history. Data preprocessing techniques are applied to handle missing values and normalize features. Feature selection is used to identify the most relevant attributes affecting heart health. Machine learning models are trained to recognize patterns associated with heart failure risk. The system aims to provide early warnings and assist healthcare professionals in decision-making. It reduces reliance on manual diagnosis and improves efficiency. The system is scalable and can handle large healthcare datasets. It supports real-time prediction and monitoring. Overall, it enhances patient care through intelligent analysis and prediction.

Existing System

Existing systems for heart failure prediction rely mainly on traditional clinical assessments and manual diagnosis by doctors. These methods include physical examinations, ECG reports, and laboratory tests. While effective, they are time-consuming and depend heavily on expert interpretation. There is limited use of automated tools in many healthcare setups. Existing approaches may not effectively analyze large datasets or detect hidden patterns. They often lack predictive capabilities and focus only on diagnosis after symptoms appear. Data integration from multiple sources is limited. Traditional methods may lead to delayed diagnosis. Existing systems are not scalable for large populations. They also lack real-time monitoring capabilities. As a result, early detection and prevention are limited. Overall, these systems are less efficient and more resource-intensive.

Disadvantages of Existing System

- Time-consuming diagnosis process
- High dependency on medical experts
- Limited predictive capability
- Delayed detection of heart conditions
- Inefficient handling of large datasets
- Lack of automation
- No real-time monitoring
- Higher cost of diagnosis
- Limited scalability

Proposed System

The proposed system introduces a machine learning-based approach for efficient heart failure prediction. It collects patient data from medical records and wearable devices. Data preprocessing techniques are applied to clean and normalize the dataset. Feature

selection methods identify the most important health indicators. Machine learning models such as Random Forest, SVM, and Logistic Regression are used for prediction. The system provides accurate risk assessment for heart failure. It supports early detection and preventive healthcare. The solution enables automated and real-time analysis. It is scalable and suitable for large healthcare systems. The system can be integrated with hospital management systems. It reduces manual effort and improves decision-making. Overall, it enhances healthcare efficiency and patient outcomes.

Advantages of Proposed System

- Early detection of heart failure risk
- High prediction accuracy
- Automated analysis
- Efficient handling of large datasets
- Supports real-time monitoring
- Reduces healthcare costs
- Assists doctors in decision-making
- Scalable and reliable
- Faster diagnosis process

IV. Methodology

The methodology begins with collecting patient health data from medical datasets. Data preprocessing is performed to handle missing values and normalize features. Feature selection techniques are applied to identify important variables. The dataset is split into training and testing sets. Machine learning models such as Logistic Regression, SVM, and Random Forest are trained. Hyperparameter tuning is used to optimize model performance. The models are evaluated using accuracy, precision, recall, and F1-score. Cross-validation techniques ensure reliability. The best-performing model is selected for deployment. The system predicts heart failure risk for new patients. Continuous learning improves accuracy over time. The system supports real-time data integration.

System Architecture

The system architecture for efficient heart failure prediction is designed as a layered pipeline that processes patient data and generates risk predictions. The input layer collects health data such as age, blood pressure, cholesterol, and medical history related to Heart Failure. This data is passed to the preprocessing layer, where missing values are handled, noise is removed, and features are normalized for consistency. The feature selection module then identifies the most relevant attributes that significantly impact heart health. The processed data is fed into the machine learning module, where models such as Logistic Regression, SVM, or Random Forest analyze patterns and predict risk levels. The evaluation module validates model performance using metrics like accuracy and recall. The output layer displays prediction results, indicating the likelihood of heart failure. A database layer stores patient records, model outputs, and historical data for future reference. The system also includes a user interface for doctors and healthcare professionals to interact with predictions. It supports integration with hospital management systems and wearable devices. This architecture ensures scalability, real-time processing, and accurate decision support.

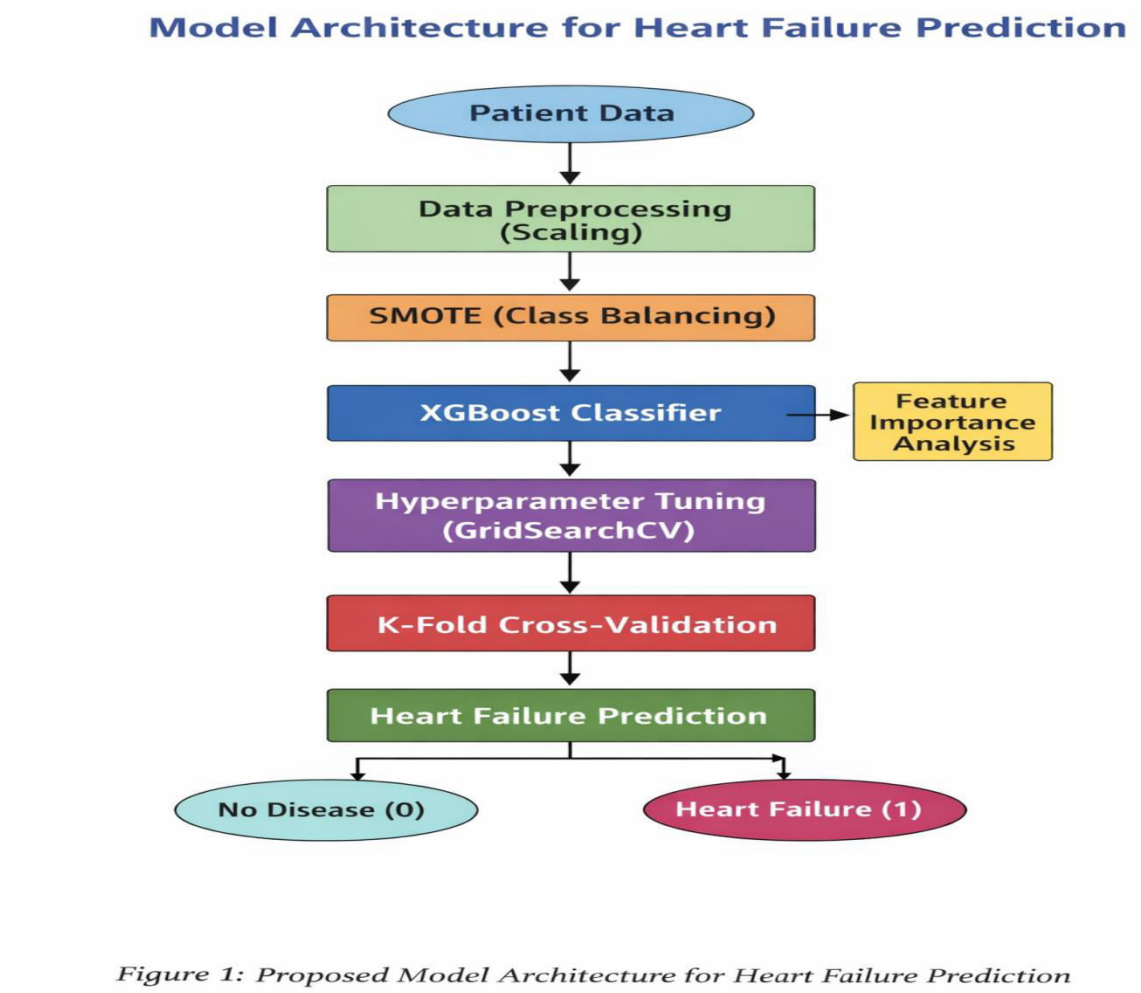


Figure 1: Proposed Model Architecture for Heart Failure Prediction

V. Result and Output

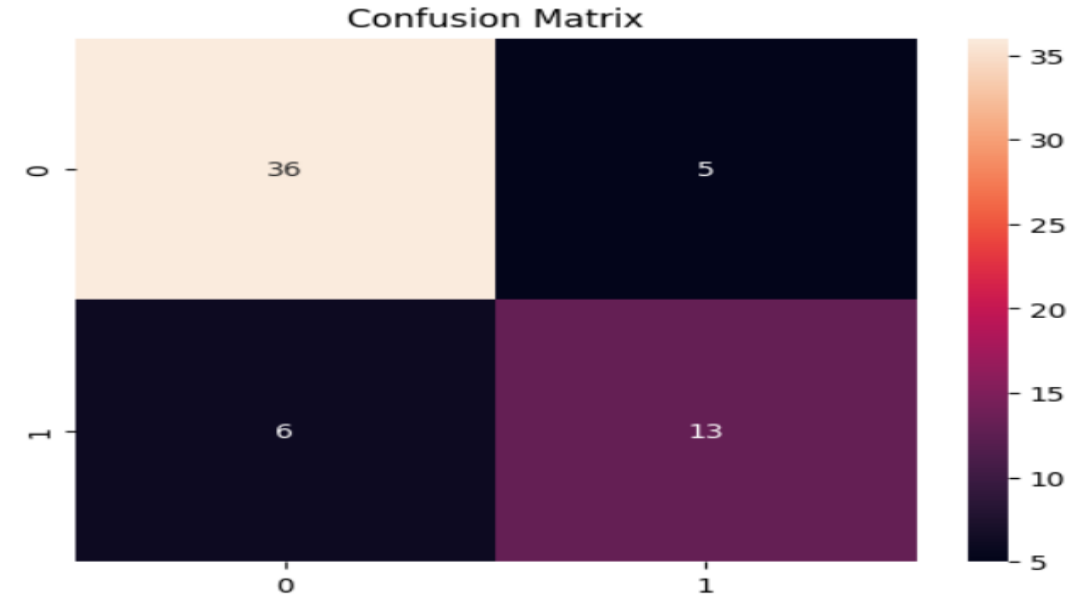
```
# Output result
print("\n🩺 RESULT:")
if prediction == 1:
    print("⚠️ High Risk of Heart Failure")
else:
    print("✅ Low Risk / No Heart Failure")

print(f"Probability: {probability:.2f}")
```

```
Enter age: 45
Enter anaemia: 0
Enter creatinine_phosphokinase: 0.6
Enter diabetes: 0
Enter ejection_fraction: 60
Enter high_blood_pressure: 120
Enter platelets: 150000
Enter serum_creatinine: 0.6
Enter serum_sodium: 135
Enter sex: 1
Enter smoking: 0
Enter time: 1

🩺 RESULT:
✅ Low Risk / No Heart Failure
Probability: 0.14
```

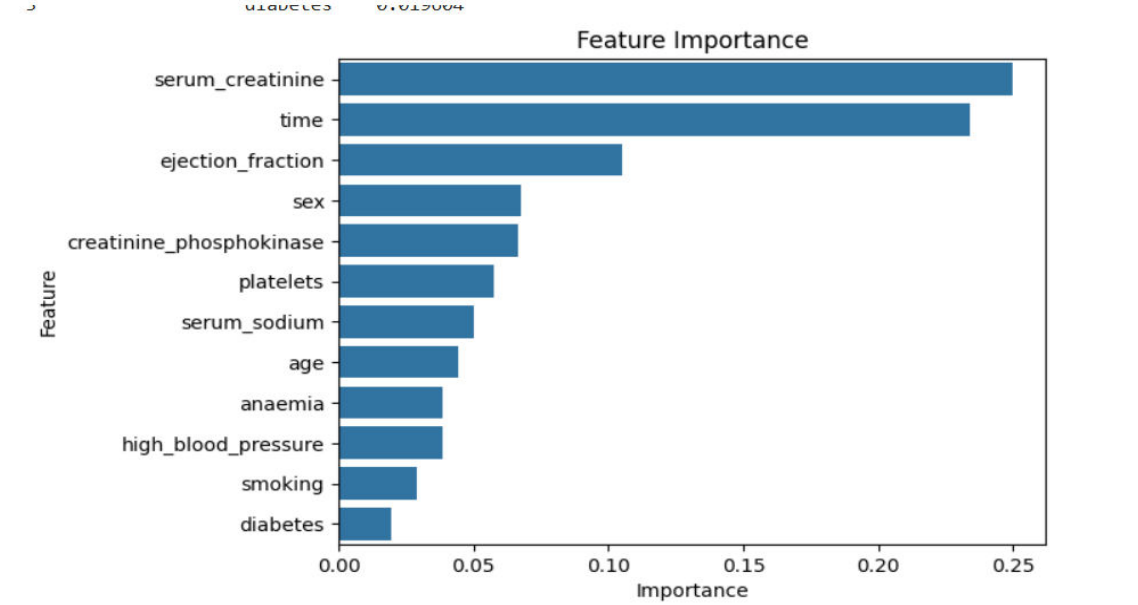
```
Enter age: 29
Enter anaemia: 1
Enter creatinine_phosphokinase: 1.3
Enter diabetes: 1
Enter ejection_fraction: 45
Enter high_blood_pressure: 120
Enter platelets: 420000
Enter serum_creatinine: 200
Enter serum_sodium: 132
Enter sex: 0
Enter smoking: 0
Enter time: 10

🩺 RESULT:
⚠️ High Risk of Heart Failure
Probability: 0.94
/usr/local/lib/python3.12/dist-packages/sklearn/utils/validation.py:2739: UserWarning: X does not have valid feature
  warnings.warn(
```

```
    Accuracy: 0.8166666666666667
...  ROC-AUC: 0.8202824133504493

    Classification Report:
                  precision    recall  f1-score   support

               0       0.86      0.88      0.87        41
               1       0.72      0.68      0.70        19

        accuracy                           0.82        60
       macro avg       0.79      0.78      0.79        60
    weighted avg       0.81      0.82      0.82        60
```

```
    accuracy                           0.82        60
   macro avg       0.79      0.78      0.79        60
weighted avg       0.81      0.82      0.82        60
```



```
Top Features:
                      Feature  Importance
7           serum_creatinine    0.249598
11                      time    0.234244
4          ejection_fraction    0.104878
9                         sex    0.067662
2   creatinine_phosphokinase    0.066257
6                  platelets    0.057502
8               serum_sodium    0.050197
0                        age    0.044197
1                    anaemia    0.038587
5         high_blood_pressure    0.038545
10                    smoking    0.028728
3                   diabetes    0.019604
```



VI. Conclusion

The developed machine learning pipeline presents an effective approach for predicting heart failure outcomes using structured clinical data. By combining preprocessing techniques such as feature scaling and class balancing with SMOTE, the model ensures improved data quality and representation. The use of XGBoost, along with hyperparameter tuning through GridSearchCV and Stratified K-Fold cross-validation, significantly enhances predictive performance and robustness. Furthermore, the evaluation metrics and visualization tools provide a comprehensive understanding of model behavior and performance.

Despite its strengths, the model has certain limitations, including the lack of advanced feature engineering and limited interpretability. Incorporating domain knowledge and explainable AI techniques could further improve its applicability in clinical settings. Overall, this work serves as a strong foundation for heart disease prediction systems and demonstrates the importance of combining data preprocessing, model

optimization, and evaluation strategies. Future improvements can focus on enhancing interpretability, reducing computational cost, and validating the model on diverse datasets to ensure reliability and scalability in real-world healthcare applications.

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.