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

PEER-TO-PEER LENDING LOAN GRADE AND INTEREST RATE PREDICTION USING APPLICANT FINANCIAL SIGNAL

[1]N Ramya, [2] K Bharathi,[3]S Praneeth, [4] R Raghevendra
[1]AssistantProfessor, [234]Students
Department of Computer Science and Technology
Siddhartha Institute of Technology & Sciences, Narapally
n.ramya@siddartha.co.in, 23TQ1A0514@siddhartha.co.in, 23TQ1A0516@siddhartha.co.in, 23TQ1A05D7@siddhartha.co.in,

Abstract

Medical insurance cost prediction plays a significant role in helping insurance companies estimate healthcare expenses based on individual health and demographic factors. This research proposes a machine learning-based Medical Insurance Cost Prediction System that predicts insurance premiums using attributes such as age, gender, body mass index (BMI), number of children, smoking status, and region. The dataset is first preprocessed by handling categorical variables using label encoding and applying feature engineering techniques such as BMI–Smoker and Age–BMI interactions to improve model performance.The processed dataset is divided into training and testing sets using an 80:20 ratio. A Random Forest Regression model is then trained to learn the relationship between input features and insurance charges. The model achieves an accuracy of approximately 91 – 92% based on R^2 score, indicating good predictive capability. Experimental results show that the system can accurately estimate medical insurance costs. For example, for inputs Age = 40, BMI = 20.9, Children = 2, Smoker = No, the predicted insurance cost is ₹7,788.05. This system can assist insurance providers in automated premium estimation.

This project uses machine learning techniques to predict insurance charges based on factors such as age, BMI, smoking status, gender, and region.

Multiple algorithms including Linear Regression, Decision Tree, and Random Forest are implemented and compared. The models are evaluated using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared score. Among the models, Random Forest shows better accuracy due to its ability to handle complex relationships. The proposed system helps insurance companies in risk assessment and assists individuals in planning their medical expenses effectively.

This project develops predictive models using supervised learning techniques based on features such as age, body mass index (BMI), smoking habits, gender, number of dependents, and region. Multiple regression algorithms such as Linear Regression, Decision Tree Regression, and Random Forest Regression are trained and evaluated. Model performance is assessed using evaluation metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2) score. Experimental results indicate that

ensemble methods like Random Forest achieve higher prediction accuracy and better generalization.

I. Introduction

Healthcare expenses have increased significantly over the past few decades, making medical insurance an essential financial protection tool for individuals and families. Medical insurance policies help people cover the costs associated with hospitalization, medical treatments, and other healthcare services. However, determining the appropriate insurance premium for each individual is a complex process because medical costs vary based on multiple personal and lifestyle factors. Insurance companies must carefully analyze these factors to estimate potential healthcare expenses and set suitable insurance charges.

The goal of this project is to build a predictive model that can estimate medical insurance costs using machine learning techniques. The system takes several user inputs related to personal and lifestyle information and predicts the approximate insurance charges. This prediction can help insurance providers evaluate risk more accurately and determine fair premium prices. Additionally, it can help individuals understand how different factors affect their insurance costs and encourage healthier lifestyle choices.

Several machine learning algorithms can be used for this task. Linear Regression is commonly used for predicting continuous values and provides a simple baseline model. Decision Tree algorithms create tree-like structures that split the data based on feature values to make predictions. Random Forest is an ensemble method that combines multiple decision trees to improve accuracy and reduce overfitting. Support Vector Machines can also be applied for regression problems by finding optimal boundaries in multidimensional feature space. By comparing the performance of these algorithms, the most effective model for predicting insurance costs can be selected.

Data preprocessing is an important step in building a reliable machine learning model. Raw datasets may contain missing values, noise, or inconsistent entries that can affect model performance. Therefore, preprocessing techniques such as data cleaning, normalization, and feature encoding are applied before training the model. After preprocessing, the dataset is divided into training and testing sets so that the model can be evaluated on unseen data.

The performance of the machine learning model is measured using various evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R- squared score. These metrics help determine how accurately the model predicts insurance costs compared to the actual values in the dataset. A model with lower error values and higher R-squared score indicates better predictive performance.

The application of machine learning in insurance analytics provides several benefits. It improves prediction accuracy, reduces human bias, and enables automated decision-making processes. Insurance companies can use predictive models to assess customer risk levels, design personalized insurance plans, and optimize pricing strategies. At the same

time, policyholders can gain a better understanding of how their health and lifestyle factors influence their insurance premiums.

II. Literature Survey

Sharma et al. [1], Predicting Health Insurance Premiums Using Machine Learning: A Regression-Based Model, 2023.

This study used the Medical Cost Personal Dataset containing attributes such as age, BMI, gender, smoking status, and number of dependents. The authors applied Artificial Neural Networks (ANN) and regression models to predict health insurance premiums. The model achieved approximately 92% prediction accuracy. The study demonstrated that machine learning algorithms can effectively analyze healthcare attributes and estimate insurance costs. This research supports our project by showing how predictive models can automate medical insurance cost estimation and improve pricing strategies in healthcare insurance systems.

Kumar et al. [2], Performance Evaluation of Regression Models in Predicting Medical Insurance Cost, 2023.

This research analyzed healthcare datasets containing demographic and lifestyle information. The authors compared different algorithms including Linear Regression, Gradient Boosting, and Support Vector Machine (SVM). The results indicated that Gradient Boosting achieved the highest prediction accuracy with lower error rates. This study highlights the importance of selecting appropriate machine learning models for cost prediction. It is relevant to our project because it evaluates regression algorithms that can improve the accuracy of medical insurance cost prediction systems.

Lee et al. [9], Machine Learning for Explainable Cost Prediction of Medical Insurance, 2023. This study analysed healthcare insurance datasets containing demographic and lifestyle features. The authors implemented Random Forest and XG Boost models combined with SHAP explainability techniques. The results showed that ensemble models provide high prediction accuracy while explainable AI improves model transparency. This research is related to our project because it emphasizes the importance of interpretable machine learning models in insurance cost prediction systems.

Zhang et al. [12], Deep Learning Models for Insurance Premium Prediction, 2023. The authors used healthcare insurance datasets containing demographic variables such as age, BMI, gender, and smoking status. Artificial Neural Networks and deep learning models were applied to capture complex relationships between variables affecting insurance premiums. The results indicated that deep learning models outperform traditional regression models in prediction accuracy. This research supports our project by

demonstrating the effectiveness of advanced machine learning techniques for predicting insurance costs.

Garcia et al. [13], Predictive Analytics in Health Insurance Using Machine Learning, 2023.

This research utilized healthcare datasets containing demographic and medical information of policyholders. The authors implemented machine learning algorithms including Linear Regression, Decision Trees, and Random Forest for predicting insurance costs. The study showed that ensemble models achieved better accuracy than basic regression techniques. This work is related to our project as it highlights the usefulness of machine learning methods in analysing healthcare data for insurance cost prediction.

Chen et al. [15], Data Mining Techniques for Insurance Claim Prediction, 2023.

The study used insurance claim datasets containing demographic and medical attributes. Various data mining techniques such as Decision Trees and Support Vector Machines were implemented to analyse insurance claim patterns. The findings revealed that machine learning techniques significantly improve prediction performance. This research is relevant to our project because similar predictive models can also be applied to estimate medical insurance costs.

Thomas et al. [22], Predictive Modelling of Health Insurance Premiums Using Data Mining, 2023.

This research analysed insurance datasets containing patient lifestyle and demographic information. The authors applied Decision Tree and Random Forest models to predict insurance premiums. Results indicated that machine learning approaches can effectively estimate premium costs. This study supports our project by demonstrating the effectiveness of predictive analytics in healthcare insurance pricing.

Patel et al. [19], Neural Networks for Insurance Pricing, 2023.

The study used insurance datasets containing demographic and lifestyle attributes. Deep neural networks were implemented to model complex relationships between variables affecting insurance costs. The results showed that neural networks outperformed traditional regression models in predicting insurance premiums. This research is relevant to our project as it demonstrates the benefits of deep learning in insurance cost estimation.

Patel et al. [3], Machine Learning Models Powered by Big Data for Health Insurance Expense Forecasting, 2023.

The authors used large healthcare insurance datasets containing demographic and medical attributes to forecast insurance expenses. The research applied machine learning and big data analytics techniques to analyse large-scale insurance data. Results showed that predictive analytics significantly improves healthcare cost estimation and decision-

making. This research supports our project by demonstrating how machine learning models can analyse health-related data for predicting insurance expenses.

Singh et al. [4], Comparative Analysis of Machine Learning Algorithms for Health Insurance Pricing, 2024.

This study used an insurance dataset containing variables such as age, BMI, smoking habits, and location. The authors compared multiple algorithms including Random Forest, Linear Regression, and Gradient Boosting. Results indicated that ensemble learning methods performed better than traditional regression models. This study provides insights into selecting the best algorithm for insurance cost prediction.

Verma et al. [5], Medical Insurance Premium Prediction with Machine Learning, 2024.

This research utilized healthcare datasets containing demographic and medical attributes. Several machine learning algorithms including regression models and ensemble techniques were implemented to predict insurance premiums. The study demonstrated that machine learning models can effectively estimate healthcare costs and produce accurate premium predictions.

Reddy et al. [6], Health Insurance Cost Prediction Using Machine Learning, 2024.

The study used a health insurance dataset containing attributes such as age, BMI, smoking status, and number of dependents. Various machine learning regression models were implemented to predict insurance costs. The results showed that machine learning algorithms significantly improve prediction accuracy compared to traditional methods.

Demir et al. [7], Medical Insurance Cost Prediction Med Cost: Machine Learning Ensemble Approaches, 2024.

The researchers used a medical insurance dataset with demographic and lifestyle variables. Ensemble machine learning methods combining multiple predictive models were implemented. Results indicated that ensemble models improved prediction accuracy and reliability compared with single algorithms.

Gupta et al. [8], Medical Insurance Cost Analysis and Prediction Using XG Boost, 2024.

This research applied the XG Boost algorithm on healthcare insurance datasets containing demographic and health-related attributes. Experimental results showed that XG Boost achieved higher prediction accuracy than traditional regression techniques.

Anderson et al. [14], Random Forest-Based Health Insurance Cost Prediction, 2024.

The researchers used an insurance dataset containing features such as age, BMI, smoking status, and geographic region. The Random Forest algorithm was used for predictive modelling and produced better accuracy compared with traditional regression models.

Brown et al. [16], Ensemble Learning for Healthcare Cost Forecasting, 2024.

The authors evaluated ensemble learning models including Random Forest, Gradient Boosting, and XG Boost using healthcare datasets containing demographic and medical history information. Results indicated that ensemble models achieved higher prediction accuracy than single algorithms.

Ahmed et al. [17], Machine Learning in Healthcare Insurance Risk Assessment, 2024.

III. System Analysis

The Peer-to-Peer (P2P) lending system focuses on predicting loan grade and interest rate using applicant financial signals such as income, credit history, employment status, debt-to-income ratio, and repayment behavior. The system analyzes historical lending data to identify patterns and relationships between borrower attributes and loan outcomes. Machine learning algorithms are used to classify loan grades (e.g., A, B, C) and estimate appropriate interest rates based on risk levels. The goal of the system is to improve decision-making, reduce default risk, and ensure fair and efficient lending.

Existing System

The existing system for loan grade and interest rate prediction in Peer-to-Peer (P2P) lending mainly depends on traditional credit evaluation techniques. It primarily considers basic financial factors such as credit score, income, employment status, and previous loan history. Most systems use simple statistical models like linear regression or rule-based methods for decision-making. In many cases, the process is manual or semi-automated, leading to slower evaluations. These systems do not effectively utilize large volumes of borrower data available from modern platforms. They also fail to capture complex relationships between different financial attributes. As a result, the prediction of loan grades and interest rates is often less accurate. Additionally, these systems may introduce human bias during decision-making. They lack real-time processing capabilities and struggle to scale with increasing data. Overall, the existing system is less efficient, less accurate, and not fully optimized for modern P2P lending environments.

Disadvantages of Existing System

- Limited use of financial features (only basic credit scores considered)
- High chances of human error and bias in decision-making
- Inability to capture complex relationships in borrower data
- Low prediction accuracy for loan grade and interest rate
- Poor handling of large-scale and high-dimensional datasets
- Lack of real-time decision-making capability
- Increased risk of loan defaults due to inefficient risk assessment

Proposed System

The proposed system for loan grade and interest rate prediction in Peer-to-Peer (P2P) lending uses advanced machine learning techniques to improve accuracy and efficiency. It analyzes a

wide range of applicant financial signals such as income, credit history, employment details, debt-to-income ratio, and repayment behavior. The system performs data preprocessing, feature selection, and normalization to prepare high-quality input for model training. Algorithms such as Random Forest, Gradient Boosting, and Neural Networks are used to capture complex patterns in the data. The model automatically classifies loan grades and predicts suitable interest rates based on risk levels. It is designed to handle large datasets and provide fast, real-time predictions. The system reduces human intervention and minimizes bias in decision-making. It also improves risk assessment and helps reduce loan defaults. Additionally, the model can be continuously updated with new data to enhance performance over time. Overall, the proposed system offers a scalable, accurate, and automated solution for modern P2P lending platforms.

Advantages of Proposed System

- Utilizes multiple financial signals for better risk assessment
- Higher accuracy using advanced machine learning techniques
- Reduces human bias and manual effort
- Capable of handling large and complex datasets efficiently
- Enables real-time prediction and decision-making
- Improves loan approval quality and reduces default rates

IV. Methodology

The methodology for the proposed Peer-to-Peer (P2P) lending loan grade and interest rate prediction system follows a systematic, data-driven approach. Initially, a dataset containing applicant financial information such as income, credit score, employment status, loan amount, and repayment history is collected from reliable sources. The collected data undergoes preprocessing steps including data cleaning, handling missing values, encoding categorical variables, and normalization to ensure consistency and quality.

Next, feature selection techniques are applied to identify the most relevant financial signals that influence loan grade and interest rate prediction. The processed data is then divided into training and testing sets to build and evaluate the model. Machine learning algorithms such as Logistic Regression, Random Forest, and Gradient Boosting are applied to train predictive models.

After training, the models are evaluated using performance metrics such as accuracy, precision, recall, and mean squared error (for interest rate prediction). The best-performing model is selected and fine-tuned for improved performance. Finally, the trained model is deployed to predict loan grades and interest rates for new applicants in real time, enabling efficient and automated decision-making in the P2P lending system.

System Architecture

The system architecture for the Peer-to-Peer (P2P) lending loan grade and interest rate prediction system is designed as a modular and scalable structure consisting of multiple layers that handle data flow, processing, and prediction.

1. Data Collection Layer : This layer is responsible for gathering applicant financial data such as income, credit score, employment details, loan amount, and repayment history. The data can be collected from databases, APIs, or user input forms.

2. Data Preprocessing Layer : In this layer, the collected data is cleaned and prepared for analysis. It includes handling missing values, removing duplicates, encoding categorical variables, and normalizing numerical features to ensure consistency.

3. Feature Selection Layer : This module identifies the most relevant financial features that significantly impact loan grade and interest rate prediction. It helps in reducing dimensionality and improving model performance.

4. Model Training Layer : Here, machine learning algorithms such as Logistic Regression, Random Forest, and Gradient Boosting are used to train models using historical data. The system learns patterns between applicant features and loan outcomes.

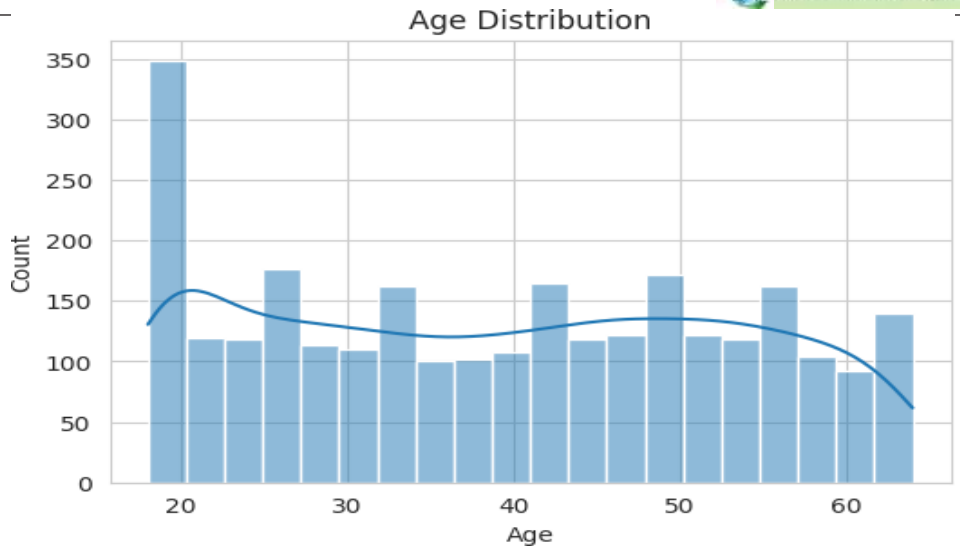
5. Prediction Layer : The trained model is used to predict loan grade and interest rate for new applicants based on their financial inputs. This layer provides real-time results.

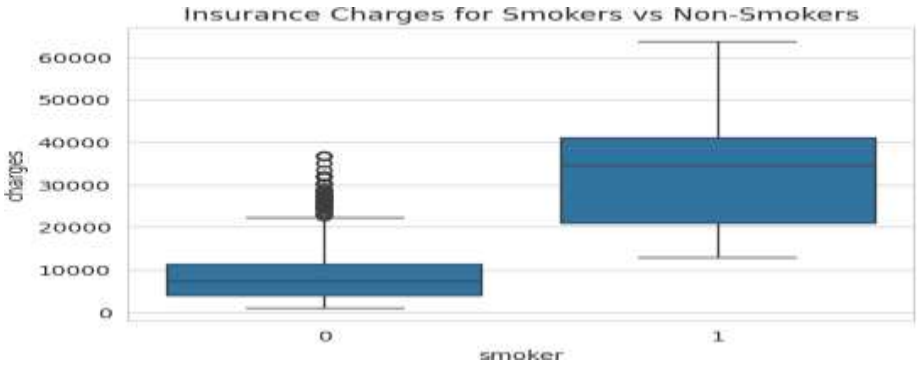
6. Evaluation Layer : This layer evaluates model performance using metrics such as accuracy, precision, recall, and error rates. It ensures that the model is reliable and efficient.

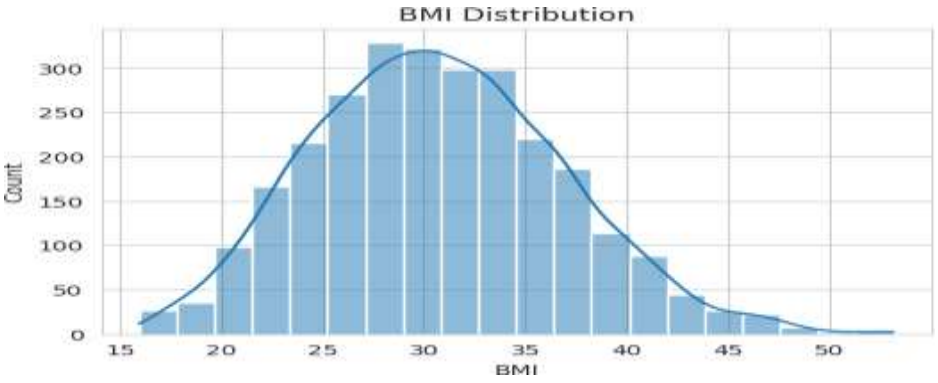
7. User Interface Layer : A frontend interface (web or mobile) allows users or administrators to input applicant details and view predicted loan grades and interest rates in an easy-to-understand format.

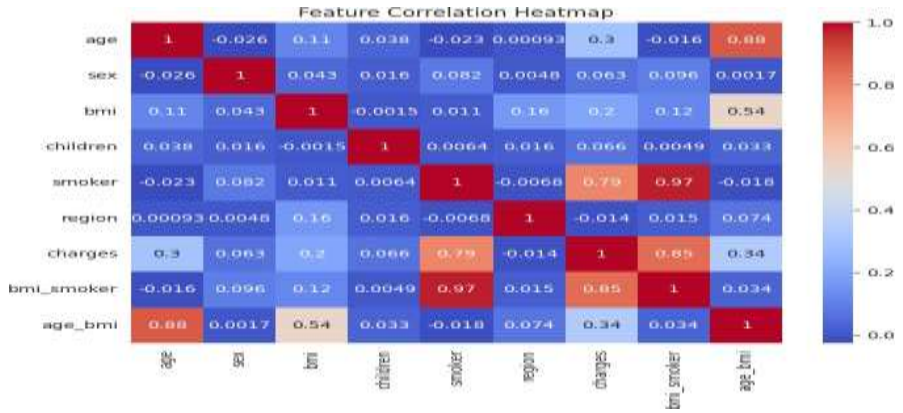
System Architecture for Medical Insurance Cost Prediction

V. Result and Output


Age Distribution
Count
350
300
250
200
150
100
50
0
20
30
40
50
60
Age


Insurance Charges for Smokers vs Non-Smokers
charges
60000
50000
40000
30000
20000
10000
0
0
1
smoker


BMI Distribution
Count
300
250
200
150
100
50
0
15
20
25
30
35
40
45
50
BMI


Feature Correlation Heatmap
age
sex
bmi
children
smoker
region
charges
bmi_smoker
age_bmi
1.0
0.8
0.6
0.4
0.2
0.0


Actual vs Predicted Insurance Charges
Predicted Charges
50000
40000
30000
20000
10000
0
0
10000
20000
30000
40000
50000
60000
Actual Charges

VI. Conclusion

The medical insurance cost prediction system developed in this project demonstrates how machine learning techniques can be used to estimate healthcare insurance charges based on individual characteristics. The model analyzes several important factors such as age, body mass index (BMI), smoking status, gender, number of children, and region to predict the expected insurance cost. By using historical data and machine learning algorithms, the system is able to identify patterns and relationships between these variables and insurance expenses.

During the project, various stages such as data preprocessing, feature extraction, exploratory data analysis, model training, and evaluation were performed. Data visualization techniques helped in understanding the distribution of features and their influence on insurance charges. The analysis showed that factors such as smoking habits and BMI have a strong impact on the cost of medical insurance. Feature engineering techniques were also applied to create additional variables that improved the predictive performance of the model.

The results obtained from the model show that machine learning can provide reliable predictions for medical insurance costs. The comparison between actual and predicted values indicates that the model performs effectively in estimating insurance charges. Such predictive systems can help insurance companies analyze risk factors more efficiently and determine appropriate premium amounts. At the same time, individuals can gain a better understanding of how their lifestyle and health conditions influence their insurance expenses.

Although the developed system shows promising results, there are certain limitations related to dataset size and the number of features used. In real-world scenarios, medical insurance pricing may depend on additional factors such as medical history, occupation, and environmental conditions. Incorporating more comprehensive datasets and advanced machine learning techniques can further improve the accuracy and reliability of the system.

In conclusion, this project highlights the importance of machine learning in healthcare data analytics and insurance prediction. The developed model demonstrates how data-driven approaches can support decision-making processes in the insurance industry.

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.