

ROAD ACCIDENT SEVERITY PREDICTION USING DATA SCIENCE

¹D. ARUNA, ²GOIPARTHI LALITHA, ³VUSALA ISWARYA, ⁴JONNALAGADDA VIJAY KUMAR,
⁵PAMARTHI DEVANADH

¹Assistant Professor, ²³⁴⁵Students, Department of Computer Science and Engineering, SRI VASAVI
INSTITUTE OF ENGINEERING & TECHNOLOGY, NANDAMURU, ANDHRA PRADESH

ABSTRACT

Road accidents represent a major public safety concern worldwide, causing significant loss of life, injuries, and economic damage every year. Predicting the severity of road accidents before or immediately after they occur can help authorities implement preventive measures, improve emergency response, and enhance overall traffic management. The proposed system, Road Accident Severity Prediction Using Data Science, focuses on analyzing historical accident data and identifying patterns that influence accident severity. The system utilizes advanced data science and machine learning techniques to process multiple influencing factors such as weather conditions, road type, temperature, visibility, lighting conditions, and speed limits. Data preprocessing techniques including data cleaning, normalization, and categorical encoding are applied to prepare the dataset for effective modeling. Since accident datasets are often highly imbalanced, the Synthetic Minority Over-sampling Technique (SMOTE) is implemented to balance the dataset and improve prediction performance. Machine learning algorithms such as Random Forest, Decision Tree, and Logistic Regression are employed to build predictive models that classify accidents into different severity levels such as minor, serious, or fatal. Feature importance analysis helps identify the most influential factors contributing to severe accidents. The model is evaluated using performance metrics such as accuracy, precision, recall, and F1-score to ensure reliable predictions. The proposed system assists traffic management authorities, policymakers, and emergency services in understanding accident patterns and implementing preventive strategies. By integrating predictive analytics with traffic safety systems, the proposed framework contributes toward reducing accident-related fatalities and improving road safety planning. Overall, this data-driven approach demonstrates how intelligent predictive systems can support smarter transportation infrastructure and enhance public safety.

Keywords: Road Safety, Accident Severity Prediction, Machine Learning, Data Science, Traffic Analytics, PredictiveModeling, SMOTE.

Received: 06-02-2026

Accepted: 13-03-2026

Published: 20-03-2026

I INTRODUCTION

Road traffic accidents are one of the leading causes of death and serious injuries across the world, creating significant social and economic problems for developing and developed countries alike [1].

Rapid urbanization and the increasing number of vehicles on roads have significantly contributed to the growth of traffic congestion and accident occurrences [2]. Transportation safety has therefore become an important research area for engineers, policymakers, and data scientists who aim to

reduce accident risks and improve road infrastructure [3]. Traditional accident analysis methods mainly rely on descriptive statistics and historical reports, which often fail to provide predictive insights about accident severity [4]. With the advancement of data science and artificial intelligence technologies, researchers are increasingly exploring machine learning techniques to analyze complex traffic datasets and identify accident risk factors [5]. Machine learning algorithms can efficiently process large-scale datasets and detect patterns that are difficult to identify through manual analysis [6]. These techniques help researchers examine relationships between accident severity and contributing factors such as road conditions, driver behavior, weather patterns, and vehicle characteristics [7]. Predictive models have shown promising results in identifying accident-prone locations and understanding severity patterns in transportation systems [8]. Such predictive insights can assist government agencies and traffic authorities in designing effective road safety strategies [9]. The development of intelligent transportation systems has further enhanced the ability to collect and analyze traffic data using sensors, surveillance cameras, and digital monitoring platforms [10]. These technologies generate large amounts of structured and unstructured data that can be analyzed using modern data analytics techniques [11]. Researchers have shown that machine learning-based accident prediction models can significantly improve traffic safety planning and accident prevention measures [12]. Data-driven safety analysis provides more accurate insights compared to traditional statistical approaches [13]. The integration of predictive analytics into transportation management systems has therefore gained increasing attention in recent

years [14]. These predictive systems help shift road safety management from reactive accident response to proactive accident prevention [15].

Accident severity prediction requires the analysis of multiple environmental, road-related, and traffic-related factors that influence accident outcomes [16]. These factors include weather conditions, road geometry, visibility levels, lighting conditions, and vehicle speed limits [17]. Accident datasets collected from transportation departments often contain large numbers of variables that must be carefully processed before analysis [18]. Data preprocessing techniques such as cleaning, normalization, and encoding are necessary to transform raw accident data into a usable format for machine learning models [19]. One of the major challenges in accident severity prediction is the presence of imbalanced datasets, where severe accidents occur much less frequently than minor accidents [20]. This imbalance can lead to biased predictive models that perform poorly in detecting serious accident cases [21]. To address this issue, researchers have proposed resampling techniques such as the Synthetic Minority Over-sampling Technique (SMOTE) to balance datasets during model training [22]. Ensemble machine learning algorithms such as Random Forest have been widely used for accident severity prediction due to their high accuracy and robustness [23]. These algorithms combine multiple decision trees to improve prediction performance and reduce overfitting problems [24]. Feature importance analysis provided by ensemble models also helps identify key factors influencing accident severity [25]. Such insights can guide policymakers in implementing targeted road safety interventions [26]. Predictive accident severity models also assist emergency response teams in prioritizing accident

management strategies [27]. The integration of machine learning with transportation analytics therefore represents a promising approach to improving road safety systems [28]. Continued research in data-driven traffic safety analysis is expected to further enhance prediction accuracy and system efficiency [29]. As transportation networks continue to expand globally, intelligent accident prediction systems will play a critical role in developing safer and more sustainable mobility solutions [30].

II LITERATURE SURVEY

Early studies on road accident analysis primarily relied on statistical models such as logistic regression to identify factors affecting accident severity [1]. These models examined relationships between accident outcomes and variables such as driver characteristics, road conditions, and vehicle types [2]. Although statistical approaches provided valuable insights, they often struggled to capture complex nonlinear relationships present in traffic accident datasets [3]. To overcome these limitations, researchers began applying machine learning techniques to accident severity prediction problems [4]. Decision Tree algorithms became one of the earliest machine learning approaches used for analyzing accident data due to their interpretability and ability to handle categorical variables [5]. These models allow researchers to visualize decision paths that lead to different accident severity levels [6]. Several studies demonstrated that Decision Trees can effectively identify important accident risk factors such as road type, traffic density, and weather conditions [7]. However, single decision tree models are often sensitive to dataset variations and may suffer from overfitting problems [8]. To address this issue,

ensemble learning methods such as Random Forest were introduced for accident severity prediction tasks [9]. Random Forest combines multiple decision trees to generate more accurate and stable prediction results [10]. Research has shown that Random Forest models outperform many traditional classification algorithms in transportation safety analysis [11]. These models also provide feature importance rankings that help identify the most influential accident-related variables [12]. The ability to interpret model outputs makes ensemble learning particularly useful for traffic safety decision-making [13]. Machine learning-based accident analysis has therefore become a rapidly growing research field in intelligent transportation systems [14]. Many recent studies highlight the importance of integrating environmental and infrastructure factors in accident prediction models [15].

In addition to ensemble learning techniques, researchers have explored advanced machine learning and deep learning approaches for accident severity prediction [16]. Support Vector Machines have been applied to classify accident severity by separating data points using optimal hyperplanes in high-dimensional feature spaces [17]. These models are effective for handling complex classification tasks involving multiple accident variables [18]. Neural network models have also been applied to accident prediction problems due to their ability to capture nonlinear relationships in large datasets [19]. Deep learning architectures have further improved prediction performance by automatically learning hierarchical feature representations from traffic data [20]. However, these models require large datasets and extensive computational resources for effective training [21]. Another major challenge identified in accident

severity prediction research is the problem of class imbalance within accident datasets [22]. Severe and fatal accidents occur relatively less frequently compared to minor accidents, which can affect model training accuracy [23]. To address this issue, resampling techniques such as SMOTE have been widely adopted in machine learning research [24]. SMOTE generates synthetic samples for minority classes, allowing models to learn balanced patterns during training [25]. Studies have demonstrated that combining SMOTE with ensemble learning algorithms significantly improves accident severity prediction accuracy [26]. Researchers have also emphasized the importance of feature selection techniques to reduce dataset dimensionality and improve model efficiency [27]. Hybrid models integrating machine learning algorithms with feature optimization techniques have shown promising results in transportation safety studies [28]. These approaches enable more accurate identification of accident risk factors and severity patterns [29]. Overall, the literature indicates that data-driven predictive systems provide powerful tools for improving traffic safety and reducing accident-related fatalities [30].

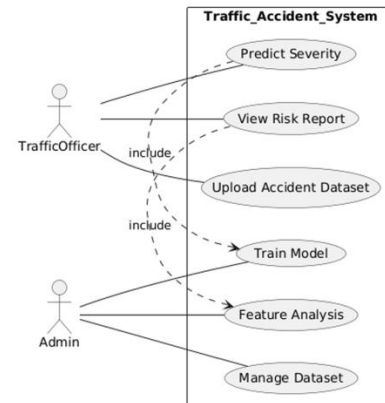
III METHODOLOGY

The proposed Road Accident Severity Prediction system follows a structured data science methodology consisting of data collection, preprocessing, feature engineering, model training, and evaluation stages. Initially, accident data is collected from reliable transportation datasets containing attributes such as weather conditions, road type, temperature, visibility, speed limits, lighting conditions, and traffic environment. The

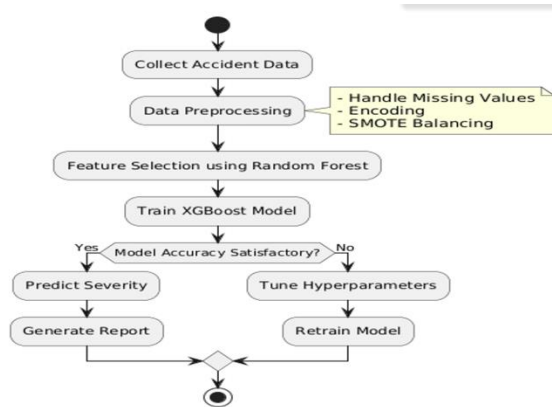
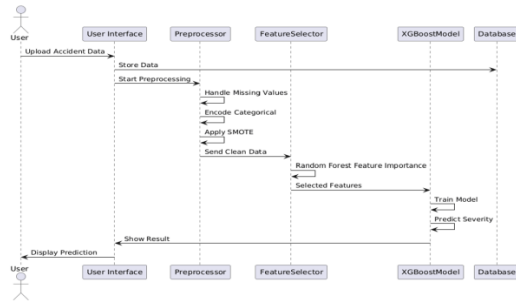
collected data is then subjected to preprocessing to ensure quality and consistency. Data cleaning techniques are applied to remove missing values, handle outliers, and correct inconsistencies in the dataset. Categorical variables such as weather conditions and road types are transformed using encoding techniques, while numerical features are normalized to maintain uniform data distribution. Since accident datasets often contain imbalanced class distributions, the Synthetic Minority Over-sampling Technique (SMOTE) is implemented to generate additional samples for minority classes such as fatal accidents. This ensures balanced model training and improves prediction accuracy. After preprocessing, feature selection techniques are used to identify the most influential attributes contributing to accident severity. Machine learning algorithms including Decision Tree, Random Forest, and Logistic Regression are then applied to train predictive models using the processed dataset. Random Forest is particularly useful because it combines multiple decision trees and reduces overfitting while providing feature importance analysis. The trained models are evaluated using performance metrics such as accuracy, precision, recall, and F1-score to assess prediction reliability. Cross-validation techniques are also applied to ensure model stability and prevent overfitting. The final predictive model classifies accidents into different severity categories such as minor, serious, and fatal based on input conditions. This systematic methodology enables the system to generate reliable predictions and supports traffic authorities in implementing preventive safety measures and improving accident response strategies.

IV SYSTEM DESIGN

The system design of the Road Accident Severity Prediction model follows a modular architecture that ensures efficient data processing, predictive analysis, and result generation. The system consists of several key components including data collection module, preprocessing module, feature engineering module, machine learning model module, and prediction output module. The data collection module gathers accident datasets from transportation authorities or publicly available traffic safety databases. These datasets contain detailed accident-related information such as environmental conditions, road characteristics, vehicle information, and accident outcomes. Once the data is collected, it is stored in a structured database where it can be accessed for further processing. The preprocessing module plays a crucial role in improving data quality by handling missing values, removing duplicate records, and correcting inconsistencies. Data transformation techniques are applied to convert categorical attributes into numerical representations that can be processed by machine learning algorithms. Additionally, normalization techniques are used to scale numerical features, ensuring that all variables contribute equally during model training.



The feature engineering and model development modules form the core of the system design. Feature engineering identifies important variables that significantly influence accident severity, enabling the model to focus on the most relevant factors. The system incorporates SMOTE to address class imbalance in accident data, ensuring that minority classes such as fatal accidents are adequately represented during training. After preprocessing and feature engineering, the dataset is divided into training and testing subsets. Machine learning algorithms such as Random Forest, Decision Tree, and Logistic Regression are applied to build predictive models capable of classifying accident severity levels. Random Forest is selected as the primary model due to its robustness, high accuracy, and ability to analyze feature importance. The prediction module uses the trained model to evaluate new input data and classify the expected severity level of a potential accident scenario. The results are presented through a user interface or dashboard, allowing traffic authorities to interpret predictions easily. This system design enables efficient analysis of accident-related data and supports proactive traffic safety management strategies.



V PROPOSED SYSTEM

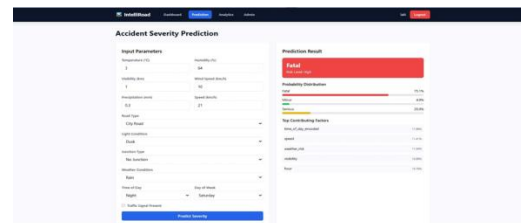
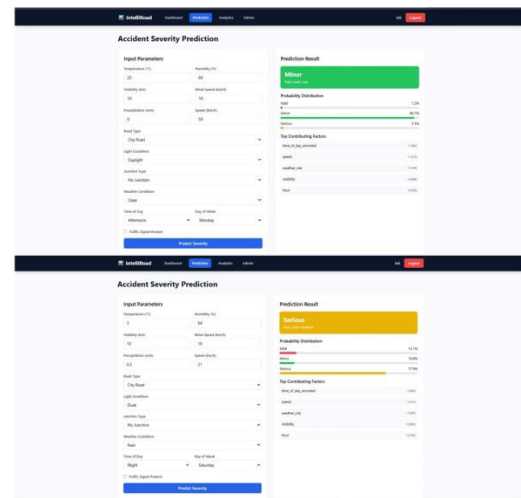
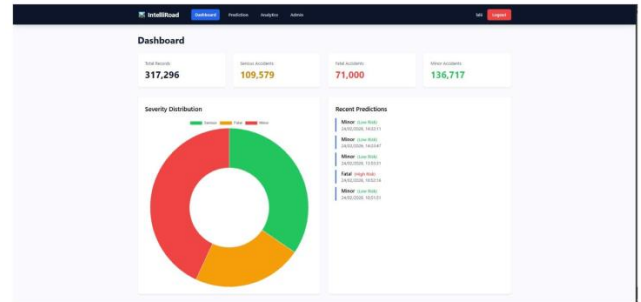
The proposed system introduces an intelligent data-driven framework designed to predict road accident severity using machine learning and advanced data analysis techniques. Unlike traditional traffic safety systems that rely mainly on historical statistics and manual analysis, the proposed system integrates automated predictive modeling to analyze accident risk factors and estimate potential severity levels. The system utilizes a comprehensive accident dataset containing environmental, road, and traffic-related variables. These variables include weather conditions, road surface type, temperature, lighting conditions, speed limits, and visibility. The system processes these features through a structured data pipeline that ensures efficient preprocessing and feature extraction. By applying normalization and encoding techniques, the system prepares the dataset for effective machine learning model training. One of the major improvements

introduced in the proposed system is the use of SMOTE to handle imbalanced accident datasets. Since fatal accidents occur less frequently compared to minor accidents, traditional models may produce biased predictions. The SMOTE technique generates synthetic samples for minority classes, enabling the model to learn balanced patterns and improve classification performance.

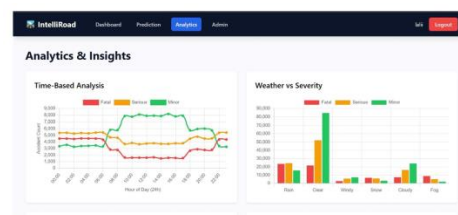
The predictive engine of the proposed system is based on ensemble machine learning techniques, particularly the Random Forest algorithm. Random Forest combines multiple decision trees to generate a robust predictive model capable of handling complex relationships between accident variables. This approach improves prediction accuracy and reduces the risk of overfitting. Additionally, feature importance analysis is performed to identify the most influential factors affecting accident severity, allowing traffic authorities to understand key risk contributors. The trained model is capable of classifying accident severity into categories such as minor, serious, and fatal based on the provided input conditions. The system also includes a visualization module that displays prediction results and important contributing factors through charts or dashboards, making it easier for decision-makers to interpret the analysis. By integrating machine learning prediction with real-world traffic data, the proposed system provides valuable insights that can support road safety planning, infrastructure improvements, and targeted traffic regulations. Ultimately, the proposed system contributes toward reducing accident-related fatalities and improving transportation safety through intelligent data-driven decision support.

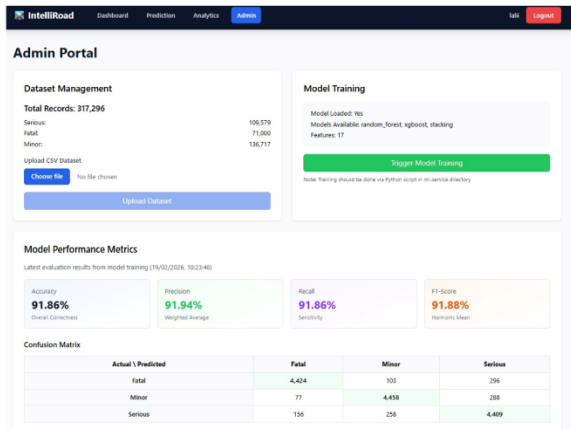
VI RESULTS & DISCUSSION

The experimental results demonstrate the effectiveness of the proposed Road Accident Severity Prediction system in accurately classifying accident severity levels. The dataset was divided into training and testing subsets to evaluate the performance of different machine learning models. Among the evaluated algorithms, the Random Forest model achieved the highest accuracy due to its ensemble learning capability and ability to handle complex feature relationships. The implementation of SMOTE significantly improved model performance by balancing the dataset and reducing bias toward majority classes. Evaluation metrics such as precision, recall, and F1-score indicated that the model effectively identified severe accident cases while maintaining overall prediction reliability. Feature importance analysis revealed that weather conditions, visibility, road type, and speed limits were among the most influential factors affecting accident severity. The results confirm that machine learning-based predictive models can provide valuable insights into accident patterns and support proactive road safety management strategies.



48





VII CONCLUSION

Road accident severity prediction is an important research area that can significantly contribute to improving traffic safety and reducing accident-related fatalities. The proposed Road Accident Severity Prediction system utilizes advanced data science and machine learning techniques to analyze complex accident datasets and predict the severity level of potential accidents. By incorporating data preprocessing, feature engineering, and predictive modeling techniques, the system effectively transforms raw accident data into meaningful insights. The use of SMOTE for handling imbalanced datasets improves model fairness and ensures that severe accident cases are accurately identified. Among the evaluated algorithms, Random Forest proved to be the most effective model due to its robustness, high accuracy, and ability to analyze feature importance. The predictive system helps identify critical factors such as weather conditions, road type, visibility, and speed limits that significantly influence accident severity. These insights can assist transportation authorities, policymakers, and safety planners in implementing targeted road safety measures and improving traffic management

strategies. Furthermore, the integration of predictive analytics into transportation systems enables proactive accident prevention rather than reactive response. Future improvements may include integrating real-time traffic data, IoT sensors, and deep learning models to further enhance prediction accuracy and system efficiency. Overall, the proposed data-driven framework demonstrates how intelligent predictive systems can support smarter transportation infrastructure, reduce accident risks, and improve public safety. The adoption of such technologies can play a vital role in developing safer roads and sustainable transportation systems worldwide.

REFERENCES

1. Abdel-Aty, M., & Pande, A. (2007). Crash data analysis. *Accident Analysis & Prevention*, 39(3), 476–482.
2. Chang, L., & Wang, H. (2006). Analysis of traffic injury severity. *Accident Analysis & Prevention*, 38(3), 475–484.
3. Lord, D., & Mannering, F. (2010). Statistical analysis of crash-frequency data. *Transportation Research Part A*, 44(5), 291–305.
4. Savolainen, P., Mannering, F., Lord, D., & Quddus, M. (2011). Crash severity modeling. *Accident Analysis & Prevention*, 43(5), 1666–1676.
5. Washington, S., Karlaftis, M., & Mannering, F. (2010). *Statistical and econometric methods for transportation data analysis*. CRC Press.
6. Zeng, Q., Wen, H., Huang, H., & Abdel-Aty, M. (2017). Bayesian models for crash

- severity. *Accident Analysis & Prevention*, 99, 206–215.
7. Chen, C., Zhang, G., Qian, Z., & Tarefder, R. (2016). Decision tree analysis of crash severity. *Transportation Research Record*, 2583, 74–82.
8. Li, Z., Liu, P., Wang, W., & Xu, C. (2018). Traffic crash severity analysis. *Journal of Safety Research*, 65, 115–122.
9. Ye, X., Pendyala, R., & Washington, S. (2008). Crash injury severity analysis. *Accident Analysis & Prevention*, 40(4), 1616–1622.
10. Sohn, S., & Shin, H. (2001). Pattern recognition for road accidents. *Expert Systems with Applications*, 21(2), 89–97.
11. Anderson, T. (2009). Machine learning for crash prediction. *Transportation Research Board*.
12. Kotsiantis, S. (2007). Supervised machine learning review. *Informatica*, 31(3), 249–268.
13. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
14. Quinlan, J. (1993). C4.5: Programs for machine learning. Morgan Kaufmann.
15. Bishop, C. (2006). *Pattern recognition and machine learning*. Springer.
16. Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, W. (2002). SMOTE technique. *Journal of Artificial Intelligence Research*, 16, 321–357.
17. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
18. Han, J., Kamber, M., & Pei, J. (2012). *Data mining concepts and techniques*. Elsevier.
19. Tan, P., Steinbach, M., & Kumar, V. (2018). *Introduction to data mining*. Pearson.
20. Witten, I., Frank, E., & Hall, M. (2016). *Data mining practical machine learning tools*. Morgan Kaufmann.
21. Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer.
22. Zhou, Z. (2012). *Ensemble methods in machine learning*. Springer.
23. Aggarwal, C. (2015). *Data mining: The textbook*. Springer.
24. Bishop, C. (2006). *Pattern recognition techniques*. Springer.
25. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *Introduction to statistical learning*. Springer.
26. Murphy, K. (2012). *Machine learning: A probabilistic perspective*. MIT Press.
27. Géron, A. (2019). *Hands-on machine learning with Scikit-Learn*. O'Reilly.
28. Raschka, S., & Mirjalili, V. (2019). *Python machine learning*. Packt.
29. Pedregosa, F. et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

30. Biau, G., & Scornet, E. (2016). Random forest theory. *Test*, 25(2), 197–227.
31. Mahesh Ganji. (2025). Enhancing Oracle Cloud HR Reporting Through AI-Driven Automation. *Journal of Science & Technology*, 10(6), 28–36. <https://doi.org/10.46243/jst.2025.v10.i06.p28-36>
32. Todupunuri, A. (2025). THE ROLE OF AGENTIC AI AND GENERATIVE AI IN TRANSFORMING MODERN BANKING SERVICES. *American Journal of AI Cyber Computing Management*, 5(3), 85–93. <https://doi.org/10.64751/ajaccm.2025.v5.n3.pp85-93>
33. Todupunuri, A. . (2024). Artificial Intelligence Ethics: Investigating Ethical Frameworks, Bias Mitigation, and Transparency in AI Systems to Ensure Responsible Deployment and Use of AI Technologies. *International Journal of Innovative Research in Science, Engineering and Technology*, 13(09), 1–14. <https://doi.org/10.15680/ijirset.2024.1309002>
34. Sushma Babburi. (2025). Token-Based Data Accounting System For Transparent Model Training And Cost Allocation. *American Journal of AI Cyber Computing Management*, 5(4), 463–474. <https://doi.org/10.64751/ajaccm.2025.v5.n4.pp463-474>
35. Snigdha Gaddam. (2025). SOFTWARE STACK PREPARED FOR AI TRANSITIONING FROM MODULES TO MODELS. *American Journal of AI Cyber Computing Management*, 5(4), 451–462. <https://doi.org/10.64751/ajaccm.2025.v5.n4.pp451-462>
36. Gaddam, S. INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING.
37. Bajarang Bhagwat, V. (2023). Optimizing Payroll to General Ledger Reconciliation: Identifying Discrepancies and Enhancing Financial Accuracy. *JOURNAL OF ADVANCE AND FUTURE RESEARCH*, 1(4). <https://doi.org/10.56975/jafr.v1i4.501636>
38. Srinivasa Kalyan Immadi. (2025). Harnessing Artificial Intelligence In Oracle Hcm: Revolutionising Workforce Management With Automation And Predictive Analytics. *International Journal of Data Science and IoT Management System*, 4(4), 7–13. <https://doi.org/10.64751/ijdim.2025.v4.n4.pp7-13>
39. S. M. K. P. (2025). Cryptography in iOS: A Study of Secure Data Storage and Communication Techniques. *International Journal on Science and Technology*, 16(1). <https://doi.org/10.71097/ijst.v16.i1.1403>
40. Suhasnadh Reddy Veluru, Sai Teja Erukude, and Viswa Chaitanya Marella. 2025. Multimodal Detection of Fake Reviews using BERT and ResNet-50. In 2025 4th International Conference on Innovative Mechanisms for Industry Applications (ICIMIA). IEEE, 877–882.
41. Cyril, H. P. (2025). Event-Driven Provisioning Architectures For Modern Telecom Networks: Overcoming Legacy Limitations And Enabling Autonomous 6g Operations. *International Journal of Advanced Research in Computer Science*, 16(6), 75–82. <https://doi.org/10.26483/ijarcs.v16i6.7389>
42. Jay Bharat Mehta. (2025). AUTONOMOUS PATCH VALIDATION FOR ZERO-DAY EXPLOITS IN ENTERPRISE CLOUDS. *International Journal of Applied Mathematics*, 38(4s), 1270–1285. <https://doi.org/10.12732/ijam.v38i4s.685>
43. Reddy, S. K. (2025). Hyperpersonalization driven by AI is expected to be at the Lead in shaping the future of loyalty rewards.

- Journal of Emerging Technologies and Innovative Research.
44. Reddy, S. K. R. (2021). Strengthening the Security of Loyalty Reward Systems: An In-Depth Analysis of Emerging Cyber Threats and Protection Mechanisms. *Journal of Computational Analysis and Applications*, 29(6).
45. Poojari, R. (2026). Privacy-Preserving Generative AI in Healthcare Systems Using Federated Learning Approaches. *International Journal of Data Science and IoT Management System*, 5(1), 78-88.
46. Uday Kumar Kalae. (2025). AN AUTOMATED SYSTEM FOR MANAGING HIGH-AVAILABILITY CLOUD INFRASTRUCTURE THROUGH INFRASTRUCTURE-ASCODE (IAC) PRACTICES. *American Journal of AI Cyber Computing Management*, 5(2), 42–50. <https://doi.org/10.64751/ajaccm.2025.v5.n2.pp42-50>
47. Saikumar, B. (2024). Optimizing Crew Scheduling and Absence Management using Microservices: Enhancing Reliability and Efficiency in Crew Management Systems. *International Journal of Enhanced Research in Management & Computer Applications*, 13(11), 50–55. <https://doi.org/10.55948/ijermca.2024.0116>
48. Saikumar, B. (2023). Enhancing Client Engagement through AI-Driven Real-Time Reporting and Automated Alerts. *International Journal of Enhanced Research in Science, Technology & Engineering*, 12(11), 111–117. <https://doi.org/10.55948/ijerste.2023.1115>