International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

CREDIT LOAN DEFAULT RISK PREDICTION USING IMBALANCED CLASSIFICATION TECHNIQUES AND BORROWER FINANCIAL PROFILE FEATURES

[1] N Bhargavi, [2] V Anirudh Sai, [3] N Shivanand, [4] B Prakash, [5] G Suraj
[1]AssistantProfessor, [2345]Students
Department of CSE(Data Science)
Siddhartha Institute of Technology & Sciences, Narapally

bhargavi.cse@siddhartha.co.in , 24TQ1A6765@siddhartha.co.in, 24TQ1A6734@siddhartha.co.in, 24TQ1A6749@siddhartha.co.in, 24TQ1A6718@siddhartha.co.in

Abstract

Accurate prediction of credit loan defaults is a crucial aspect of risk management in the financial sector, as it enables institutions to minimize losses and make informed lending decisions. However, building effective predictive models is challenging due to the inherent class imbalance in loan datasets, where non-defaulters significantly outnumber defaulters. This study presents a comprehensive machine learning framework designed to address this challenge and improve the accuracy of loan default prediction. The model utilizes key borrower financial features such as debt-to-income (DTI) ratio, credit utilization, payment history, and employment status to capture meaningful patterns related to credit risk.

To handle the imbalance in the dataset, advanced techniques including Synthetic Minority Over-sampling Technique (SMOTE), class weighting, and cost-sensitive learning are employed. These methods help balance the data distribution and enhance the model's ability to detect minority class instances, particularly high-risk borrowers. Multiple machine learning algorithms are implemented and evaluated using performance metrics suited for imbalanced data, such as Precision-Recall Area Under Curve (PR-AUC), F1-score, and recall for the minority class.

The experimental results demonstrate that combining relevant financial features with appropriate resampling and weighting strategies significantly improves predictive performance. The proposed framework effectively increases the detection rate of potential defaulters while maintaining overall model reliability. This approach supports financial institutions in reducing credit risk, optimizing loan approval processes, and strengthening decision-making in credit management systems.

I. Introduction

The extension of credit is a fundamental driver of the modern financial system, supporting economic growth, business expansion, and individual financial stability. For financial institutions, lending activities represent a major source of revenue; however, they also introduce significant risks, particularly the risk of borrower default. When borrowers fail to meet their repayment obligations, institutions face direct financial losses, increased operational costs related to recovery, and reduced liquidity. As a result, effective credit risk assessment prior to loan approval has become a

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

critical necessity, not only for regulatory compliance but also for maintaining financial stability and profitability.

Traditionally, credit risk evaluation relied on manual assessment by human underwriters or rigid rule-based systems that considered limited parameters. While these approaches provided a basic level of screening, they often lacked the ability to capture complex patterns and relationships within large datasets. With the advancement of technology, machine learning (ML) has emerged as a powerful solution for enhancing credit risk prediction. By analyzing extensive borrower data—including debt-to-income ratios, credit history, repayment behavior, and credit utilization—ML models can identify subtle indicators of risk and enable more accurate, data-driven decision-making.

Despite its advantages, applying machine learning to credit default prediction presents a major challenge in the form of class imbalance. In most real-world datasets, the number of non-defaulters significantly outweighs the number of defaulters. This imbalance can lead to biased models that favor the majority class, often resulting in misleadingly high accuracy while failing to correctly identify high-risk borrowers. This phenomenon, known as the accuracy paradox, limits the practical usefulness of traditional ML models in financial risk assessment.

II. Literature Survey

The study of credit risk prediction has evolved significantly from traditional statistical approaches to advanced machine learning techniques. Earlier methods such as Logistic Regression (LR), Linear Discriminant Analysis (LDA), and the Altman Z-score were widely used due to their simplicity and interpretability; however, they were limited in capturing complex, non-linear relationships present in modern financial data. With the advancement of computational capabilities, machine learning models such as Support Vector Machines (SVM) and Artificial Neural Networks (ANN) were introduced, offering improved predictive performance. More recently, ensemble learning methods like Random Forest, XGBoost, and LightGBM have gained prominence due to their ability to handle large-scale, high-dimensional datasets and mixed data types efficiently.

A major challenge identified in the literature is the issue of class imbalance, where the number of non-defaulters significantly exceeds the number of defaulters. This imbalance leads to biased models that achieve high overall accuracy but fail to correctly identify high-risk borrowers, a problem known as the accuracy paradox. To address this, researchers have proposed various data-level and algorithm-level techniques. Data-level approaches include oversampling methods such as the Synthetic Minority Over-sampling Technique (SMOTE), which generates synthetic samples of the minority class, and undersampling techniques that reduce the majority class size. Hybrid methods like SMOTE-ENN further enhance performance by combining oversampling with noise removal.

In addition to resampling techniques, algorithm-level solutions such as cost-sensitive learning have been widely explored. These methods assign higher penalties to

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

misclassification of defaulters, ensuring that models prioritize the detection of high-risk cases. Ensemble techniques, particularly boosting algorithms like XGBoost and LightGBM, have shown superior performance when combined with class weighting strategies. Furthermore, feature engineering plays a crucial role in improving prediction accuracy. Key financial indicators such as debt-to-income ratio, credit utilization, and historical payment behavior are consistently identified as strong predictors of loan default.

Overall, the literature highlights that combining advanced machine learning models with effective imbalance-handling techniques and meaningful feature selection significantly enhances the performance of credit risk prediction systems. These advancements enable financial institutions to better identify potential defaulters, reduce financial losses, and improve decision-making in lending processes.

III. System Analysis

The credit loan default prediction system is designed to analyze borrower data and accurately identify high-risk applicants using machine learning techniques. The system focuses on improving decision-making in financial institutions by providing reliable predictions based on historical data. It evaluates multiple borrower attributes such as income, credit history, debt-to-income ratio, and repayment behavior. The system identifies patterns and relationships within the dataset to detect potential defaults. It addresses the challenge of class imbalance, which is common in financial datasets. Advanced preprocessing techniques are applied to clean and prepare the data for analysis. The system incorporates resampling and weighting strategies to improve model performance. It uses multiple machine learning algorithms for better prediction accuracy. Performance is evaluated using metrics suitable for imbalanced data. The system ensures scalability and adaptability for different datasets. It supports efficient risk assessment and minimizes financial loss. Overall, the system enhances the accuracy and reliability of credit risk prediction.

Existing System

The existing system for credit risk assessment primarily relies on traditional statistical models and manual evaluation processes. Financial institutions often use rule-based approaches or simple models such as logistic regression. These systems depend heavily on predefined rules and limited features. Data preprocessing and feature selection are often performed manually. The models are not capable of capturing complex patterns in borrower behavior. Existing systems struggle to handle large and diverse datasets efficiently. They do not effectively address the issue of class imbalance in credit data. As a result, predictions are often biased toward non-defaulters. Visualization and reporting capabilities are limited. There is minimal automation in model selection and evaluation. The systems require expert knowledge to interpret results. Overall, existing systems lack flexibility, scalability, and accuracy in modern data-driven environments.

Disadvantages of Existing System (Points)

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

- Relies on traditional and less accurate statistical models
- Cannot handle class imbalance effectively
- High dependency on manual data preprocessing
- Limited ability to capture complex data patterns
- Biased predictions toward majority class (non-defaulters)
- Requires expert knowledge for analysis and interpretation
- Poor scalability with large datasets
- Lack of automation in model selection and evaluation

Proposed System

The proposed system introduces a machine learning-based framework for accurate credit loan default prediction. It uses advanced algorithms to analyze borrower financial data and identify potential defaulters. The system incorporates data preprocessing techniques to handle missing values and inconsistencies. It applies imbalance handling methods such as SMOTE, class weighting, and cost-sensitive learning. These techniques improve the detection of minority class instances. The system evaluates multiple machine learning models to select the best-performing one. It uses performance metrics such as precision, recall, F1-score, and PR-AUC for evaluation. Feature engineering is applied to extract meaningful insights from borrower data. The system provides automated analysis and prediction outputs. It improves decision-making by identifying high-risk borrowers accurately. The framework is scalable and adaptable to different datasets. Overall, it enhances prediction accuracy and reduces financial risk.

Advantages of Proposed System (Points)

- Handles class imbalance using advanced techniques like SMOTE
- Improves accuracy in detecting loan defaulters
- Automates data preprocessing and analysis
- Supports multiple machine learning algorithms
- Uses appropriate evaluation metrics for better performance
- Reduces financial risk for institutions
- Enhances decision-making with reliable predictions
- Scalable for large and complex datasets
- Minimizes human intervention and errors

IV. Methodology

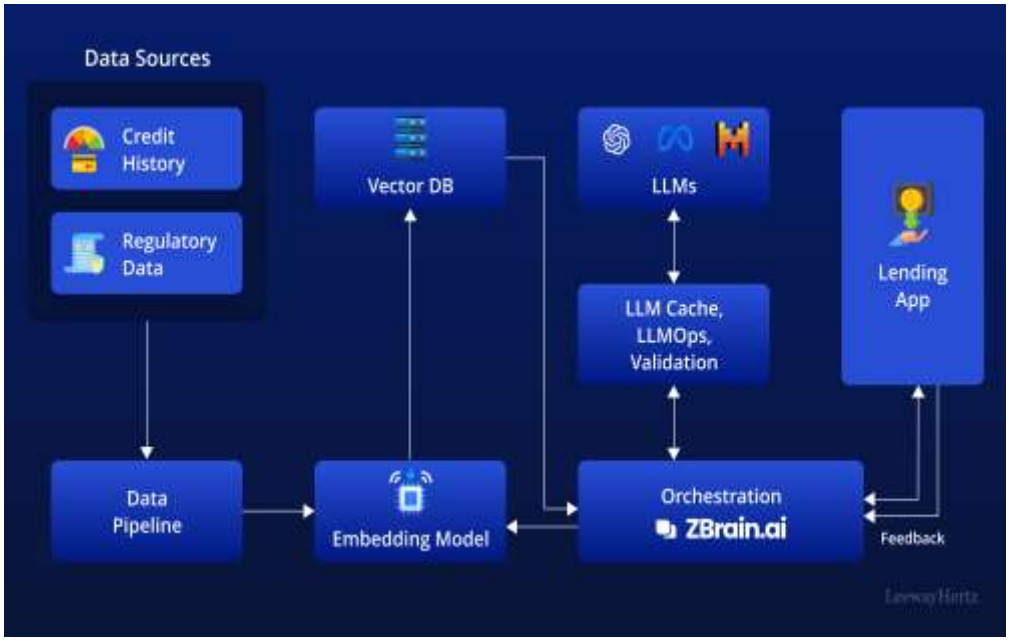
The system follows a structured methodology for credit default prediction. Initially, the dataset is collected from reliable financial sources. Data preprocessing is performed to handle missing values and inconsistencies. Feature selection is carried out to identify relevant attributes. The dataset is then analyzed for class imbalance. Techniques such as SMOTE and class weighting are applied to balance the data. The processed data is split into training and testing sets. Multiple machine learning algorithms are trained on the dataset. Model performance is evaluated using metrics like precision, recall, and F1-score. The best-performing model is selected for

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

prediction. The system generates predictions for new borrower data. Finally, results are analyzed to support decision-making in credit risk management.

System Architecture

The system architecture consists of several interconnected components. The Data Input Module collects borrower data from various sources. The Preprocessing Module cleans and prepares the data for analysis. The Feature Engineering Module extracts important features from the dataset. The Imbalance Handling Module applies techniques like SMOTE and class weighting. The Model Training Module trains multiple machine learning algorithms. The Evaluation Module assesses model performance using suitable metrics. The Prediction Module generates credit default predictions. The Reporting Module presents results in a structured format. A storage layer is used to manage datasets and outputs. The backend is implemented using Python and machine learning libraries. The system is designed to ensure scalability, efficiency, and accuracy.



V. Result and Output

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

VI. Conclusion

Accurate prediction of credit loan defaults is a critical requirement for effective risk management in the financial sector, and this project demonstrates the limitations of traditional approaches in addressing this challenge. Conventional rule-based systems and basic machine learning models often fail due to the inherent class imbalance in credit data, where the majority of borrowers are non-defaulters. These systems tend to prioritize overall accuracy, leading to the accuracy paradox, where models appear effective but fail to identify high-risk borrowers, ultimately increasing financial risk.

To overcome these challenges, this project developed a specialized, end-to-end machine learning framework tailored for imbalanced financial datasets. By focusing on detailed borrower financial attributes such as debt-to-income ratio, credit utilization, and historical payment behavior, the system was able to capture deeper insights into borrower risk profiles. This feature-driven approach enabled the model to move beyond traditional scoring techniques and provide more precise and meaningful predictions.

A key strength of the system lies in the integration of advanced imbalance handling techniques. Methods such as SMOTE, class weighting, and cost-sensitive learning were applied to ensure that the model effectively prioritized the minority class, i.e., potential defaulters. Additionally, the use of powerful ensemble models like Random Forest and XGBoost improved the model's ability to capture complex, non-linear relationships within the data. The adoption of evaluation metrics such as Precision-Recall AUC (PR-AUC) and recall further ensured that the model's performance was aligned with real-world financial objectives, particularly minimizing costly false negatives.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

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.
12. Gaddam, S. (2023). Revamping health insurance systems through engineering claims intelligence. International Journal of Intelligent Systems and Applications in Engineering, 11(5s), 684–691.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

13. Reddy, S. K. R. Developing a Modular AI Framework to Enhance Scalability and Personalization in Next-Generation Reward Platforms.
14. Poojari, R. INTELLIGENT SYSTEMS+B108 AND APPLICATIONS IN ENGINEERING.
15. Purmani, S. S. R. (2024). Aligning IT investment decisions with overall business strategy from an enterprise program management perspective, focusing on the integration of IT leadership in strategic decision-making processes. International Journal of Communication Networks and Information Security, 16(5), 1213–1219
16. Viswanathan, V., Polagani, S. S., Agarwal, R., Akula, S., Dey, S., & Kashyap, R. (2025, September). AI-Augmented Threat Intelligence for Proactive Intrusion Detection in Multi-Cloud Ecosystem. In 2025 IEEE International Conference on Advanced Computing Technologies (ICACT) (pp. 567-572). IEEE.
17. Mudusu, S. K. (2026, February 9). AI-augmented data quality engineering. InfoWorld (Foundry Expert Contributor Network).
18. DEVARASETTY, N. (2023). SCALABLE DATA ENGINEERING APPROACHES FOR AI-DRIVEN INDUSTRIAL IOT APPLICATIONS. INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH AND MANAGEMENT, 11(06), 954-968.
19. Gajula, S. (2025, December). Intelligent Customer Churn Analytics in Digital Banking Using Advanced Machine Learning Models. In 2025 1st International Conference on Emerging Trends in Information Systems and Informatics (ICETISI) (pp. 1-6). IEEE.
20. Maturi, S. Y. (2024). Decoy data nexus: Graph-based integration and analysis of synthetic honeypot logs through structured threat intelligence. International Journal of Computational and Experimental Science and Engineering (IJCESEN), 10(4), 4255–4261. https://doi.org/10.22399/ijcesen.5010
21. Pavan Kumar Adabala. (2026). Smart Retail Fuel Systems: IoT-Enabled Solutions for Loss Prevention and Environmental Safety. Computer Fraud and Security, 868–875. https://doi.org/10.52710/cfs.995
22. Srikanth Kavuri. (2022). Large Language Model (LLM)-Based Automation for Software Test Script Generation. Computer Fraud and Security. https://doi.org/10.52710/cfs.836
23. Gummadi, V. P. K., Chilamkurthi, L. S., & Kavuri, S. (2026). Service Level Objective (SLO) Observability with Splunk and Dynatrace in Microservices. 2026 International Conference on Artificial Intelligence, Systems, and Emerging Technologies (ICAISET), 1–4. https://doi.org/10.1109/icaiset66439.2026.11541542
24. Gummadi, V. P. K., Chilamkurthi, L. S., & Kavuri, S. (2026). Securing APIs in Government Clouds and Runtime Fabric Using FIPS-Enabled MuleSoft. 2026 International Conference on Artificial Intelligence, Systems, and Emerging Technologies (ICAISET), 1–6. https://doi.org/10.1109/icaiset66439.2026.11542099
25. Kumar Gummadi, V. P., Chilamkurthi, L. S., & Kavuri, S. (2026). Distributed Platform Architecture and API-Led Integration. 2026 International Conference on Artificial Intelligence, Systems, and Emerging Technologies (ICAISET), 1–6. https://doi.org/10.1109/icaiset66439.2026.11541787

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

26. Gajula, S., Bondhala, S., & Margam, M. (2026). Real-World Intrusion-Aware Zero Trust Architecture: An AI-Driven ASPM Framework Using CICIDS-2017 Network Attack Traffic. 2026 IEEE 5th International Conference on AI in Cybersecurity (ICAIC), 1–7. https://doi.org/10.1109/icaic67076.2026.11395835
27. Gaddam, S. (2024). Integrating machine learning models with continuous integration and continuous delivery (CI/CD) pipelines for a learning-driven approach to software engineering.
28. Reddy, S. K. R. Developing a Modular AI Framework to Enhance Scalability and Personalization in Next-Generation Reward Platforms.
29. Poojari, R. INTELLIGENT SYSTEMS+B108 AND APPLICATIONS IN ENGINEERING.
30. Santthosh Saai Reddy Purmani. (2026). Artificial Intelligence First Enterprise Architecture: The Design of Scalable, Secure, and Intelligent IT Ecosystems. American Journal of AI Cyber Computing Management, 6(1(2)), 1–8. https://doi.org/10.64751/ajaccm.2026.v6.n1(2).pp1-8
31. Viswanathan, V. (2023). AI-Augmented Decision Intelligence for Enterprise Systems: Integrating Cognitive Analytics for Resource and Talent Optimization.
32. Mudusu, S. (2025). Health Insurance Fraud Detection: The Role Of Advanced It Systems In Preventing And Identifying Fraud. International Journal, 16(1), 3769-3777
33. Viswanathan, V. Generative AI for Smarter Workforce Planning and Enterprise Resource Decisions.
34. Mudusu, S. K. (2025, December 22). Cognitive data architecture: Designing self-optimizing frameworks for scalable AI systems. CIO (Foundry Expert Contributor Network).
35. Agrawal, A. M., Gajula, S., Shinde, R. P., Shah, H., & Ghosh, H. (2025, July). Machine Translation for Long Sequences with Enhanced Attention Mechanisms. In 2025 5th International Conference on Electrical, Computer and Energy Technologies (ICECET) (pp. 1-6). IEEE.
36. Maturi, S. Y. (2021). Blockbond hardening: Securing pooled-hash protocols against traffic tampering, MITM hash-rate hijacking, and template coercion. International Journal of Communication Networks and Information Security, 13(3), 718–728.
37. Sikder, M. Z., Shakil, M. A. I., Ahad, A., Karim, M. F., Intakhab, B., & Islam, D. A. (2025, June). Microwave-Based Detection of Early-Stage Renal Cell Carcinoma Using UHF Range Antenna. In 2025 International Conference on Computer Systems and Technologies (CompSysTech) (pp. 1-6). IEEE.
38. Manoharan, D. (2024). Governance-Oriented Quality Engineering Framework for Healthcare EDI Modernization. International Journal of Multidisciplinary on Science and Management IJMSM, 1(2).
39. Ravishankara, M. (2026, February). PlotChain: Deterministic Checkpointed Evaluation of Multimodal LLMs on Engineering Plot Reading. In SoutheastCon 2026 (pp. 1-8). IEEE.
40. Doragacharla, V. R. (2026). Building Real-Time Pricing Systems for Modern Retail. Available at SSRN 6451760.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

41. Adabala, P. K. (2024). Utilizing predictive analytics to improve efficiency and decision-making in ERP-connected supply chains. International Journal of Intelligent Systems and Applications in Engineering, 12(22s), 2465
42. Venkata Ramana, P. (2024). AI-driven predictive analytics in ERP systems for proactive supply chain optimization. International Journal of Research in Information Technology and Computing, 8(4).
43. Kavuri, S. (2026). An Explainable Machine Learning Framework for Predicting Software Defects in Large-Scale Software Systems. 2026 IEEE 5th International Conference on AI in Cybersecurity (ICAIC), 1–6. https://doi.org/10.1109/icaic67076.2026.11395777
44. Srikanth Kavuri. (2025). AI-DRIVEN TEST AUTOMATION FRAMEWORKS: ENHANCING EFFICIENCY AND ACCURACY IN SOFTWARE QUALITY ASSURANCE. International Journal of Applied Mathematics, 38(10s), 699–710. https://doi.org/10.12732/ijam.v38i10s.990
45. Venkata Pavan Kumar Gummadi. (2023). MuleSoft Batch Processing: High-Volume Streaming Architecture. Computer Fraud and Security, 50–57. https://doi.org/10.52710/cfs.886
46. Venkata Pavan Kumar Gummadi. (2026). Infrastructure Optimization Techniques for Enterprise Integration Platforms: A Comprehensive Analysis. Computer Fraud and Security, 37–44. https://doi.org/10.52710/cfs.875