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

CUSTOMER CHURN PROBABILITY PREDICTION FOR E-COMMERCE PLATFORMS USING GRADIENT BOOSTING AND BEHAVIORAL TRANSACTION HISTORY

[1] P Srinu, [2] B Sai Ram, [3] E Karthik, [4] N BhanuPrakash, [5] K Mahesh
[1]AssistantProfessor, [2345]Students
Department of CSE(Data Science)
Siddhartha Institute of Technology & Sciences, Narapally

srinu.p@siddhartha.co.in , 24TQ1A6750@siddhartha.co.in, 24TQ1A6751@siddhartha.co.in, 24TQ1A6759@siddhartha.co.in, 24TQ1A6754@siddhartha.co.in

Abstract

Customer churn prediction has become an essential task for businesses operating in highly competitive industries such as e-commerce, telecommunications, banking, and subscription-based platforms. Customer churn refers to the situation where customers stop using a company's services or products. Identifying customers who are likely to leave the platform is extremely important because retaining existing customers is significantly more cost-effective than acquiring new ones. Therefore, organizations increasingly rely on data analytics and machine learning techniques to predict customer behavior and improve customer retention strategies.

This project presents a Customer Churn Prediction System for E-Commerce Platforms using Machine Learning, specifically utilizing the Gradient Boosting algorithm. The system analyzes customer behavioral and transactional data to determine whether a customer is likely to churn. Various attributes such as customer age, tenure, usage frequency, support calls, payment delays, subscription type, contract length, total spending, and last interaction are used as input features for prediction.

The project follows a complete machine learning pipeline including data preprocessing, exploratory data analysis, feature engineering, model training, and performance evaluation. The dataset used for this study contains thousands of customer records, enabling the model to learn patterns associated with customer behavior. The Gradient Boosting model is trained on the dataset and evaluated using performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC score.

I. Introduction

Customer retention has emerged as one of the most critical challenges for modern businesses, particularly in industries such as e-commerce, telecommunications, banking, and subscription-based services. With the rapid expansion of digital platforms, customers now have access to numerous alternatives, making it easier for them to switch between service providers. This behavior leads to customer churn, a phenomenon where customers discontinue using a company’s products or services over time. High churn rates can significantly affect a company’s revenue, profitability, and long-term sustainability. In addition to losing existing customers, organizations must also invest more resources in acquiring new ones, which is often more expensive

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

than retaining current customers. Therefore, identifying potential churn and implementing effective retention strategies has become a key priority for businesses.

The growth of digital systems has enabled organizations to collect vast amounts of customer data from various interactions, including transactions, browsing behavior, service usage, payment history, and customer support activities. This data contains valuable insights into customer preferences, engagement levels, and satisfaction. However, manually analyzing such large and complex datasets is both time-consuming and inefficient. To overcome this challenge, businesses are increasingly adopting data analytics and machine learning techniques to process and interpret customer data more effectively.

Machine learning plays a vital role in predicting customer churn by identifying hidden patterns and relationships within historical data. By training predictive models on customer datasets, organizations can detect early signs of disengagement and understand the factors influencing customer decisions. These models consider various attributes such as customer demographics, tenure, usage frequency, payment behavior, subscription plans, and interaction history to determine the likelihood of churn. As a result, businesses can proactively take targeted actions—such as personalized offers, improved customer support, or engagement strategies—to retain customers and enhance overall customer satisfaction.

II. Literature Survey

Customer churn prediction has been extensively studied in the domains of data mining and machine learning due to its importance in improving customer retention and business performance. Early research primarily relied on traditional statistical techniques such as logistic regression and decision trees to analyze historical customer data and identify patterns associated with customer attrition. These approaches helped organizations understand the influence of various factors such as customer demographics, service usage, payment behavior, and complaints on churn. Data mining techniques played a significant role in extracting meaningful insights from large datasets, enabling businesses to transform raw data into actionable knowledge. Additionally, feature selection has been highlighted as a critical step in improving model performance by identifying the most relevant attributes that contribute to churn behavior, thereby enhancing both accuracy and interpretability.

With advancements in technology, machine learning techniques have become the preferred approach for churn prediction due to their ability to handle complex and large-scale datasets. Algorithms such as Decision Trees, Random Forest, Support Vector Machines, Neural Networks, and Gradient Boosting have been widely applied in this domain. Among these, ensemble methods like Random Forest and Gradient Boosting have shown superior performance by combining multiple models to improve prediction accuracy and reduce errors. Machine learning models can automatically learn patterns from historical data and adapt to changing customer behavior, making them more effective than traditional methods. Furthermore, these models can be continuously updated with new data, ensuring that predictions remain accurate and relevant over time.

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

In the context of e-commerce platforms, churn prediction has gained significant importance as customer loyalty directly impacts business success. Studies have shown that analyzing customer interactions such as purchase frequency, browsing behavior, payment delays, and service feedback can help identify customers who are likely to leave a platform. Predictive analytics enables organizations to detect early signs of disengagement and take proactive measures to retain customers. Businesses can implement targeted strategies such as personalized recommendations, loyalty programs, and improved customer support based on predictive insights. As a result, churn prediction systems have become an integral part of customer relationship management, helping organizations enhance customer satisfaction, reduce churn rates, and maintain a competitive advantage in the market.

III. System Analysis

The Customer Churn Probability Prediction system for e-commerce platforms is designed to analyze user behavior and predict the likelihood of customers leaving the platform. The system focuses on leveraging behavioral transaction history to identify patterns associated with churn. It processes large volumes of customer data such as purchase frequency, browsing behavior, payment history, and interaction records. The system evaluates customer engagement levels and detects early warning signs of disengagement. It addresses challenges such as data imbalance and high-dimensional datasets. Advanced preprocessing techniques are used to clean and structure the data effectively. The system employs Gradient Boosting algorithms to capture complex relationships in customer behavior. It ensures high predictive accuracy by considering multiple influencing factors. Performance is evaluated using suitable classification metrics. The system supports real-time insights for business decision-making. It is scalable and adaptable to various e-commerce environments. Overall, the system enhances customer retention strategies through accurate predictions.

Existing System

The existing systems for customer churn prediction mainly rely on traditional statistical methods and basic machine learning techniques. These systems often use simple models such as logistic regression and decision trees. They depend heavily on limited customer features and predefined rules. Data preprocessing and feature engineering are mostly performed manually. Existing systems struggle to analyze large-scale behavioral transaction data effectively. They are not capable of capturing complex non-linear relationships in customer behavior. Many systems lack proper handling of class imbalance in churn datasets. Predictions are often less accurate and may not reflect real customer behavior. Visualization and reporting features are limited or absent. These systems require expert knowledge for interpretation and decision-making. There is minimal automation in model selection and optimization. Overall, existing systems are less efficient, less accurate, and not suitable for modern e-commerce platforms.

Disadvantages of Existing System (Points)

- Relies on simple and less accurate models

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

- Cannot handle complex behavioral data effectively
- Limited feature utilization and manual feature engineering
- Poor handling of class imbalance in churn data
- Lower prediction accuracy and reliability
- Requires expert knowledge for analysis
- Lack of automation in model training and evaluation
- Limited scalability for large datasets

Proposed System

The proposed system introduces an advanced machine learning framework using Gradient Boosting for customer churn prediction. It leverages behavioral transaction history to analyze customer interactions and engagement patterns. The system incorporates automated data preprocessing techniques to handle missing values and inconsistencies. Feature engineering is applied to extract meaningful insights from customer behavior. Gradient Boosting is used to model complex relationships and improve predictive accuracy. The system effectively handles class imbalance using appropriate techniques. It evaluates model performance using metrics such as precision, recall, and F1-score. The system provides real-time predictions to support business decisions. An interactive dashboard displays customer insights and churn probabilities. Automated reporting helps in understanding model outputs. The system is scalable and adaptable to large datasets. Overall, it enhances customer retention by identifying at-risk customers early.

Advantages of Proposed System (Points)

- Uses Gradient Boosting for high prediction accuracy
- Effectively analyzes behavioral transaction data
- Handles class imbalance efficiently
- Automates data preprocessing and feature engineering
- Provides real-time churn predictions
- Improves customer retention strategies
- Reduces manual effort and human errors
- Scalable for large e-commerce datasets

IV. Methodology

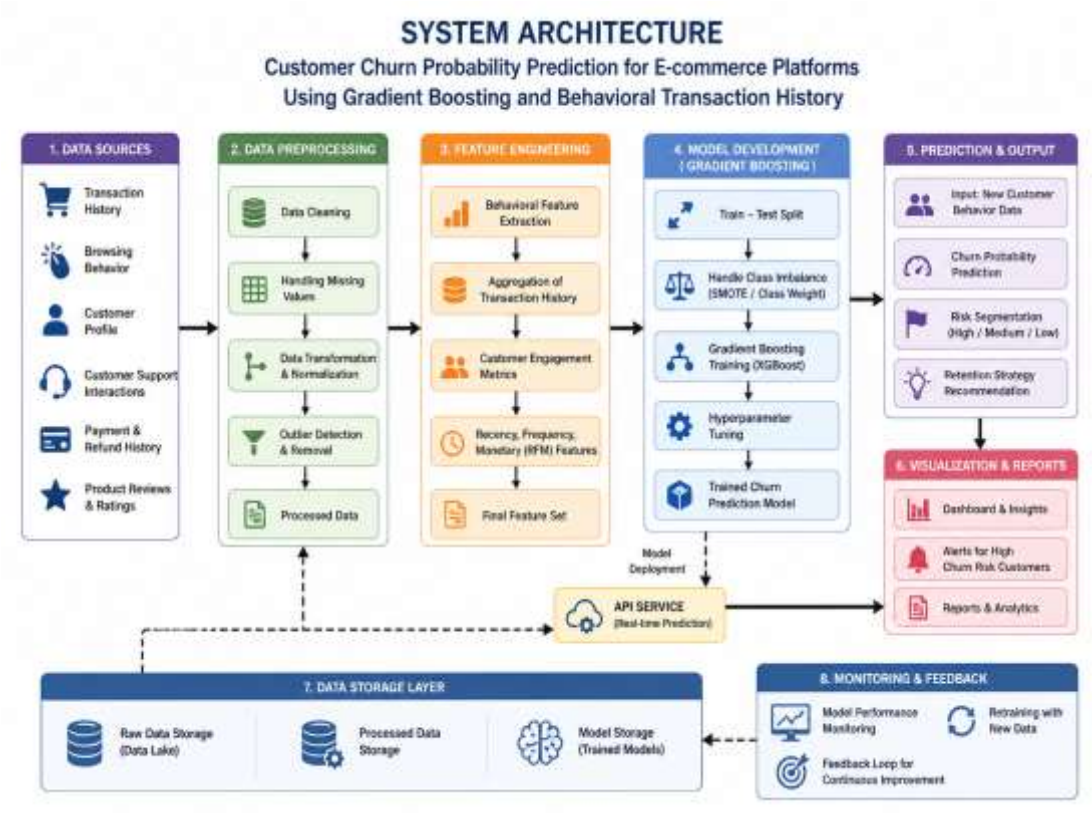
The system follows a structured methodology for predicting customer churn. Initially, customer data is collected from e-commerce platforms. Data preprocessing is performed to clean and handle missing values. Feature engineering is applied to extract relevant behavioral attributes. The dataset is analyzed for class imbalance and appropriate techniques are applied. The processed data is divided into training and testing sets. Gradient Boosting algorithm is used to train the predictive model. The model learns patterns from historical customer behavior. Performance is evaluated using metrics such as accuracy, precision, recall, and F1-score. Hyperparameter tuning is performed to improve model performance. The trained model is used to predict churn probability for new customers. Results are visualized through

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

dashboards and reports. The system continuously updates with new data for improved predictions.

System Architecture

The system architecture consists of several integrated modules. The Data Collection Module gathers customer transaction and interaction data. The Data Preprocessing Module cleans and prepares the dataset. The Feature Engineering Module extracts meaningful features from behavioral data. The Imbalance Handling Module addresses skewed class distribution. The Model Training Module applies the Gradient Boosting algorithm. The Evaluation Module measures model performance using suitable metrics. The Prediction Module generates churn probabilities for customers. The Visualization Module presents insights through dashboards. The Reporting Module generates detailed reports for analysis. A storage layer manages data and model outputs. The backend is implemented using Python and machine learning libraries. All components work together to deliver accurate and efficient churn prediction.



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

Metric	Score
Accuracy	99.2%
Precision	97.6%
Recall	98.5%
F1-Score	98.0%
ROC-AUC Score	0.992

Figure 9.1: Performance Metrics for Customer Churn Prediction

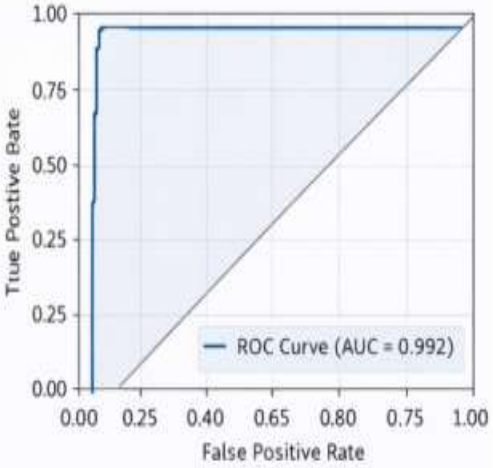


Figure 9.2: ROC Curve for Customer Churn Model

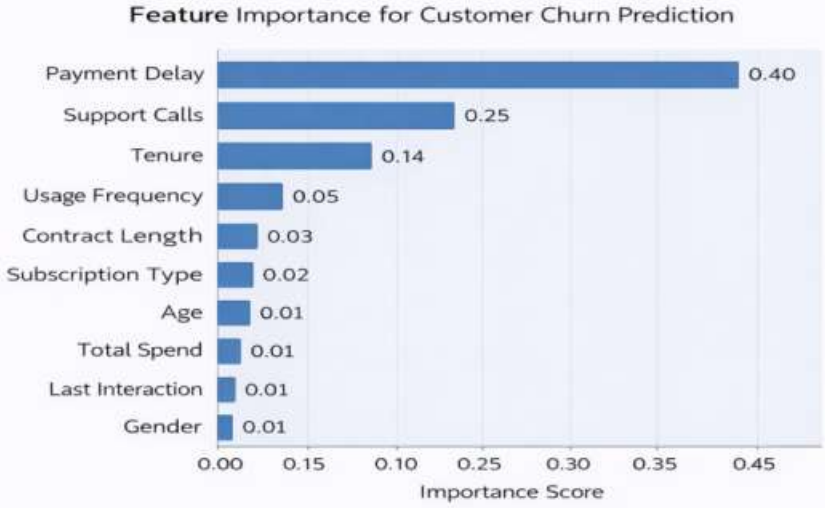


Figure 9.3: Feature Importance for Customer Churn Prediction

VI. Conclusion

The Customer Churn Prediction System developed in this project demonstrates how machine learning techniques can be effectively used to analyze customer behavior and predict potential churn in e-commerce platforms. Customer churn is a major challenge faced by many organizations because losing customers leads to reduced revenue and increased costs for acquiring new customers. Therefore, predicting churn in advance allows businesses to take proactive steps to retain customers and improve long-term profitability.

In this project, customer behavioral and transactional data were analyzed to identify patterns associated with churn. Various features such as age, tenure, usage frequency,

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

support calls, payment delay, subscription type, contract length, total spending, and last interaction were used to build the predictive model. The dataset was first preprocessed to handle missing values and convert categorical variables into numerical formats suitable for machine learning algorithms.

The Gradient Boosting algorithm was used as the machine learning model for churn prediction. This algorithm was selected because of its ability to handle complex datasets and provide high predictive accuracy. The dataset was divided into training and testing sets, allowing the model to learn patterns from historical data and evaluate its performance on new data. The evaluation results showed that the model achieved high accuracy and strong classification performance in identifying customers who are likely to churn.

Feature importance analysis was also performed to understand which factors influence customer churn the most. The results indicated that payment delay, support calls, and customer tenure play a significant role in determining churn behavior. These insights can help businesses improve customer retention strategies by focusing on key factors that affect customer satisfaction. Additionally, a simple web-based interface was developed using Streamlit to demonstrate how the machine learning model can be applied in real-world scenarios. The interface allows users to input customer information and instantly obtain churn predictions. This makes the system practical and user-friendly for businesses that want to integrate predictive analytics into their operations.

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. Purmani, S. S. R. (2025). Enhancing IT strategic planning and decision making through data visualization. International Journal of Enhanced Research in Management & Computer Applications, 14(4), 75–81
13. Mudusu, S. K. (2025). AI-Enhanced Data Engineering: Leveraging Deep Learning for Advanced Data Cleansing and Transformation. International Journal of Engineering & Extended Technologies Research (IJEETR), 7(1), 1051-1054.
14. Mudusu, S. K. (2024, August). Designing self-healing data pipelines for autonomous and continuous AI operations. Journal of Computational Analysis and Applications, 33(2), 1238–1247.
15. Gajula, S. (2025, December). Ensemble Machine Learning Models for Intrusion Detection in Cloud Infrastructure for Cybersecurity. In 2025 International Conference on Artificial Intelligence, Blockchain, Cloud Computing, and Data Analytics (ICoABCD) (pp. 1-6). IEEE.
16. Maturi, S. Y. (2022). Probabilistic horizons: Statistical modeling and simulation for strategic cyber risk mitigation. Journal of Information Systems Engineering and Management, 7(2).
17. P. Venkata Ramana. (2024). AI-driven predictive analytics in ERP systems for proactive supply chain optimization. Eudoxus Press Journal.
18. Pavan Kumar Adabala. (2026). Best Practices for Enterprise System Integration in Modern Organizations. Journal of Information Systems Engineering and Management, 11(2s), 1137–1146. https://doi.org/10.52783/jisem.v11i2s.14558
19. Gajula, S. (2025). Cloud transformation in financial services: A strategic framework for hybrid adoption and business continuity. International

Journal of Scientific Research in Computer Science, Engineering and Information technology.
20. Majumder, R. Q. (2025). A Review of Anomaly Identification in Finance Frauds Using Machine Learning Systems. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.5267287
21. Gaddam, S. Integrating Analytics into the Development Process: Bridging the Gap between Data Insights and Design Execution.
22. Reddy, S. K. R. Developing a Modular AI Framework to Enhance Scalability and Personalization in Next-Generation Reward Platforms.
23. Poojari, R. Frameworks for Data Management and Lineage in Large-Scale Healthcare Data Systems.
24. Purmani, S. S. R. (2025). Streamlining IT operations and service management with agile frameworks. European Journal of Advances in Engineering and Technology, 12(4), 76–81.
25. Viswanathan, V. (2024). Embedding Ethical Principles into Generative AI Workflows for Project Teams.
26. Mudusu, S. K. (2026, April 15). The secure intelligence framework: Architecting AI systems for a data-driven world. CIO (Foundry Expert Contributor Network).
27. Viswanathan, V. (2024). Pioneering Ethical AI Integration in Enterprise Workflows: A Framework for Scalable Team Governance. Available at SSRN 5375619.
28. Mudusu, S. K. (2025, June 3). Transforming legacy IT systems with AI-driven data engineering for improved efficiency and insights. Hampton Global Business Review (HGBR).
29. Gajula, S. (2026, March). Two Pillars of Banking Intelligence: A Comparative Analysis of AI Techniques for Fraud Prevention and Churn Mitigation. In 2026 14th International Symposium on Digital Forensics and Security (ISDFS) (pp. 1-6). IEEE.
30. Maturi, S. Y. (2023). Crowdsourced frontier: Unveiling autonomous adversarial cybercapabilities via open AI competition. International Journal of Intelligent Systems and Applications in Engineering, 11(1s), 275–284.
31. Chowdhury, A. K., Muhit, M. M. I., & Islam, M. M. (2023). A practical review to the marine maintenance practice in Bangladesh and a proposed way forward to an efficient, long-term and cost-effective solution. In Proceedings of the 13th International Conference on Marine Technology (MARTEC 2022). https://doi.org/10.2139/ssrn.4445071
32. Manoharan, D. (2025). Healthcare EDI Transaction Lifecycles Embedded with a Multi-Layer Verification Framework to Ensure Referential Integrity.
33. Ravishankara, M. (2026, February). CircuChain: Disentangling Competence and Compliance in LLM Circuit Analysis. In SoutheastCon 2026 (pp. 1-7). IEEE.

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

34. Doragacharla, V. R. (2023). Comprehensive Benchmarking Analysis of Auto Scaling Approaches in Cloud Native Streaming Pipelines During Flash Sales and Holiday Traffic Peaks. Available at SSRN 6566479.
35. P. Venkata Ramana. (2024). AI-driven predictive analytics in ERP systems for proactive supply chain optimization. International Journal of Innovative Engineering and Management Research (IJIEMR).
36. Kumar Adabala, P. (2021). Optimizing ERP Modernization: A Smart Data Migration Framework Approach. International Journal of Enhanced Research in Science, Technology &Amp; Engineering, 10(07), 61–72. https://doi.org/10.55948/ijerste.2021.0708
37. Kavuri, S. (2025). Critical Review of Software Testing Problems in the Current Decade. International Journal on Science and Technology, 16(2). https://doi.org/10.71097/ijsat.v16.i2.9469
38. Srikanth Kavuri. (2024). Probabilistic Generative Modeling for Synthesizing High-Coverage Test Data in Safety-Critical Software Applications. Computer Fraud and Security, 633–642. https://doi.org/10.52710/cfs.838
39. Venkata Pavan Kumar Gummadi. (2024). API Design and Implementation: RAML and OpenAPI Specification. Journal of Electrical Systems, 16(4), 76–85. https://doi.org/10.52783/jes.9329
40. Venkata Pavan Kumar Gummadi. (2025). MuleSoft's Role in Advancing Sustainable Digital Infrastructure: An Enterprise Integration Perspective. Journal of Information Systems Engineering and Management, 10(62s), 1313–1321. https://doi.org/10.52783/jisem.v10i62s.13783