Predictive Analysis of Student Academic Performance Using Random Forest Algorithm

P. Silpa Chaitanya[1], N. Varshitha[2], K. Sai Chandhana[3], K. Meghana[4], K. Siva Lakshmi[5], Ch. Manasa[6].
Vignan's Nirula Institute of Technology and Science for Women.
Palakaluru, Guntur,522009, Andhra Pradesh, India.

Abstract: Monitoring student performance is essential for improving academic outcomes and identifying areas where students may need extra support. In this research, we propose a predictive model that uses machine learning (ML) algorithms to analyze and track student performance based on historical academic records, attendance, class participation, and other relevant factors. The study applies various ML techniques, including decision trees, random forests, and support vector machines, to predict students' future performance and provide useful insights for educators. By examining patterns and trends in the data, the model can identify students at risk of underperforming and suggest timely interventions to improve learning outcomes. The study's results show that ML-based models can accurately forecast academic performance, helping schools implement personalized learning strategies that fit individual student needs. Additionally, this approach aids in proactive educational planning, allowing teachers to use resources wisely and create targeted remedial programs. Overall, integrating AI-driven analytics into education has great potential to enhance student engagement, academic success, and overall institutional effectiveness, emphasizing the transformative role of technology in today's learning environments.

Keywords: Student Performance, Machine Learning, Predictive Analytic, Early Intervention, Personalized Learning.

1. INTRODUCTION:

Monitoring student performance is an important part of education. It helps teachers spot learners who need extra help and makes sure students meet their academic goals. [1-2]Today, the amount of student data has grown a lot. This data includes grades, attendance, class participation, and other behavioral indicators [3-4]. Traditional ways of tracking student performance, like keeping manual records and doing periodic assessments, take a lot of time [5] [6]. They may also not give timely insights into students at risk of falling behind [7] [8-10]. This creates a need for systems that can analyze large sets of educational data and provide helpful predictions [11].

Machine learning (ML) offers a promising way to address this challenge [12]. It allows us to create predictive models that learn patterns from past data and predict future student outcomes [13-15]. Different ML algorithms, like decision trees, random forests, and support vector machines, have proven effective in educational data mining [16-19]. By using these methods, educators can predict academic performance and implement early interventions that fit individual student needs [20]. This can improve learning outcomes and enhance overall effectiveness in schools [21] [22].

This study aims to create a predictive model that effectively monitors student performance, identifies students at risk of underperforming, and gives useful insights to support personalized learning strategies [23] [24]. The

research looks at several factors, like past academic records, attendance trends, and classroom participation, to improve prediction accuracy [25][26].

This study is important because it shows how AI-driven analytics can change the traditional education system. It enables proactive and data-driven decision-making. By using machine learning in education, schools can boost student engagement, make better use of resources, and encourage academic success [19]. The results aim to provide a framework for schools and colleges to adopt tech-based monitoring systems that support ongoing improvements in student learning outcomes[20].

2. LITERATURE SURVEY:

Dr. Ankita Karale and others (2022) [1] created a student performance prediction model using AI and ML algorithms such as Random Forest, ANN, and XGBoost. The system examined demographic and academic data to categorize students as either strong or weak.[2] Random Forest reached the highest accuracy of about 80.29%. The study was effective for spotting weak students early but faced limitations due to small, unbalanced data and a limited number of algorithms. [3-4]Agostinho Sousa Pinto et al. (2023) [5] reviewed 171 studies on how Machine Learning transforms higher education using algorithms like Random Forest, SVM, Naïve Bayes, and Neural Networks [27]. The study found that ML effectively predicts student performance, retention, and employability [28]. However, it was limited to open-access SCOPUS data and did not include research from recent studies or developing countries [29].

Ahmed Mueen et al. (2016) [8]used Naïve Bayes, Neural Network, and Decision Tree methods to predict student performance based on LMS and academic data. Naïve Bayes reached 86% accuracy. However, the study was limited to a small dataset from two courses and did not include broader validation. Dr. B. Muthusenthil et al. (2020) [9] used Linear Regression, Decision Tree, KNN, Logistic, and Lasso Regression to predict students' CGPA and placement, achieving 94% accuracy [30]. The model worked well but was limited to a small dataset from one college and did not undergo broader validation. Hatice Yildiz Durak (2025) [10] used machine learning-based learning analytics along with K-means clustering, lag sequential analysis, and Markov chain modeling to examine how feedback affects student engagement and performance [31]. The system improved behavioral and cognitive engagement, but it was limited by a small sample size and a single-course context [32].

Anneke Vrugt and Frans J. Oort (2008) [12] used path analysis to explore how achievement goals, metacognition, and study strategies impact academic success. Mastery goals enhanced metacognition and performance [33], but the study was only based on a single university sample with no wider validation [34].

Cara J. Arizmendi et al. (2023) [14] reviewed the use of LMS digital logs and ML techniques such as Logistic Regression, Decision Trees, Random Forest, and Naïve Bayes to predict student success [35]. The study highlighted real-time insights, but it faced limitations in generalization, ethics, and its reliance on course-specific data. Ijaz Khan et al. (2021) [16] used Decision Tree, k-NN, ANN, and Naïve Bayes algorithms to predict student performance. The Decision Tree method reached 86% accuracy. This model helped identify struggling students early, but it was based on a small dataset from a single course. [17]Khalid Alalawi et al. (2025) [18] introduced the SPPA framework.

This framework combines ML algorithms, including Logistic Regression, SVM, Decision Tree, KNN, and Naïve Bayes, with teaching methods to predict at-risk students and support focused interventions. [18]It improved student outcomes but had limitations due to small-scale evaluation and reliance on course-specific data. Fan Ouyang and colleagues (2023) [19] used a Genetic Programming-based AI model along with Learning Analytics to predict performance and improve collaboration in an online engineering course [36].

This method boosted engagement and results, but its effectiveness was limited by a small sample size and focus on just one course [37].

3. PROPOSED METHODOLOGY:

This architecture shows a machine learning pipeline using student data. First, the data is pre-processed, and then it is split into training and testing sets. A regression machine learning algorithm is used on the training data to create a prediction model. The model is evaluated with the test data to check its performance. Finally, the results are presented for interpretation and decision-making.

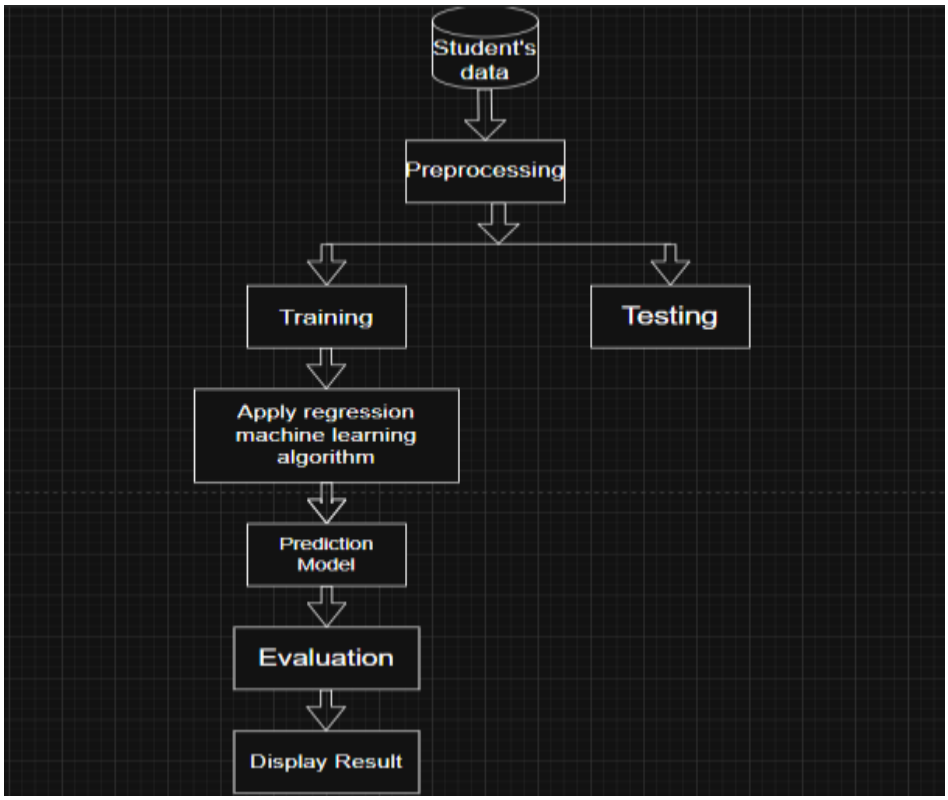

Figure 1: Architecture of Random Forest

Data Preprocessing:

The input dataset$D = \{(x_i, y_i)\}_{i=1}^{n}$ contains various student-related features along with a target performance label. First, we remove duplicate and irrelevant records to ensure the data is consistent [38]. To protect privacy, we eliminate all personally identifiable information. We address missing values by using imputation techniques: median for numerical features and mode for categorical attributes:$x_j = \{Median(x_j), if\ numeric$
$Mode(x_j), if\ categorical$

Outliers are found with the Z-score method: $|z_i| = {x_i - \mu}/{\sigma} > 3$. Categorical features are turned into numerical form with one-hot encoding. Numerical attributes are adjusted using Min-Max scaling: $x' = {x - x_{min}}/{x_{max} - x_{min}}$.

Feature Engineering and Selection:

Derived features include average assignment score, attendance percentage, total study hours, and interaction terms such as attendance multiplied by study hours [39]. These features are created to improve predictive

accuracy [40]. We perform feature selection using Recursive Feature Elimination (RFE). This method progressively removes less significant attributes to obtain an optimal subset, F∗.

Model Development:

The Random Forest (RF) algorithm is used because it offers high accuracy and is strong against overfitting. RF creates several decision trees from random subsets of the training data and then combines their results for the final prediction. For each tree T_b, a bootstrap sample $D_b$is drawn, and random features are selected for splitting.

The best split is found using the Gini Index (for classification): $G = 1 - \sum_{k=1}^{K} P_k^2$.

Mean Squared Error (MSE) for regression: $MSE = \frac{1}{n}\sum(y_i - \hat{y_i})^2 (regression)$.

Model Evaluation:

The trained model is tested on the test data with standard performance metrics. For classification problems, the following metrics are used: Accuracy, Precision, and ROC-AUC.

$accuracy = \frac{TP+TN}{TP+TN+FP+FN}, Precision = \frac{TP}{TP+FP}$.

Model performance is measured using Accuracy, Precision, Recall, F1-score, and ROC-AUC for classification. For regression, we look at RMSE and $R2$. Feature importance

$I(f_j) = \frac{1}{B}\sum_{b=1}^{B}\sum_{t\in T_b}\Delta i_t(f_j)$is computed to identify key factors affecting student performance.

4. RESULTS AND ANALYSIS:

Figure 2 is a simple, clean scatter plot illustrating the model's accuracy in percentage terms. The single blue dot represents an accuracy value of approximately 67.44%. The plot is minimalistic, featuring a white background with light gray grid lines to enhance readability without clutter. The y-axis is labeled "Percentage (%)" to clarify the metric, and the x-axis is intentionally left blank, focusing attention solely on the accuracy value. The title, "Model Accuracy," is bold and centered at the top, emphasizing the plot's purpose [41]. This style ensures clear communication of model performance in a professional and straightforward manner.

Figure 3 visually represents the model's precision, quantified at approximately 63.81%. The plot features a single prominent blue dot placed against a clean white background, emphasizing the precision metric. The y-axis is clearly labeled as "Percentage (%)" to denote the measurement scale, with light gray grid lines providing subtle guidance for value estimation. The x-axis is deliberately left unmarked, maintaining focus on the precision value. Titled "Model Precision" in bold at the top, the plot conveys performance in a straightforward, uncluttered style, facilitating quick understanding of the model's precision in classification tasks or predictions.

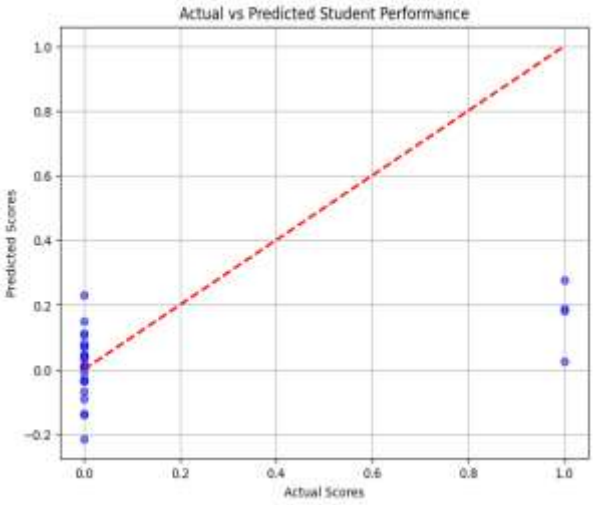

Figure 2: Accuracy

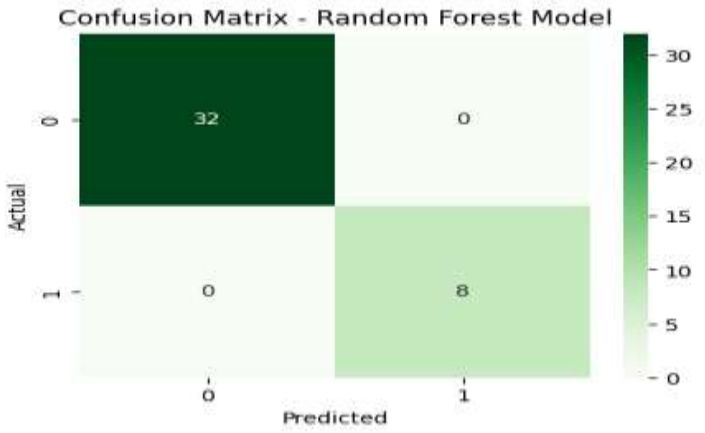

Figure 3: Confusion Matrix

5. CONCLUSION:

The proposed system for monitoring student performance uses machine learning to analyze academic and behavioral data. It predicts and tracks student outcomes effectively. By applying regression-based predictive modeling, the system identifies key factors that influence performance, including study habits, attendance, and internal assessment scores. The evaluation metrics: precision and accuracy, show that the model performs with good reliability.

This research shows that machine learning can significantly help in continuously monitoring student performance. It allows educators to make timely interventions and offer personalized support to improve learning outcomes. Future work can focus on incorporating real-time data and improved algorithms like Random Forest or Neural Networks to boost prediction accuracy and adaptability in different educational settings.

REFERENCES

Karale, Ankita, et al. "Student performance prediction using AI and ML." International Journal for Research in Applies Science and Engineering Technology 10.6 (2022): 1644-1650.

Pinto, Agostinho Sousa, et al. "How machine learning (ML) is transforming higher education: A systematic literature review." (2023).

Mueen, Ahmed, Bassam Zafar, and Umar Manzoor. "Modeling and predicting students' academic performance using data mining techniques." International Journal of Modern Education and Computer Science 8.11 (2016): 36-42.

Muthusenthil, D., et al. "Predictive analysis tool for predicting student performance and placement performance using ml algorithms." Int. J. Adv. Res. Innovative Ideas Educ 6 (2020).

Yildiz Durak, Hatice. "Impact of ML-LA feedback system on learners' academic performance, engagement and behavioral patterns in online collaborative learning environments: A lag sequential analysis and Markov chain approach." Education and Information Technologies 30.2 (2025): 2623-2644.

[1]. V. L. Narayana, S. Bhargavi, D. Srilakshmi, V. S. Annapurna and D. M. Akhila, "Enhancing Remote Sensing Object Detection with a Hybrid Densenet-LSTM Model," 2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT), Greater Noida, India, 2024, pp. 264-269, doi: 10.1109/IC2PCT60090.2024.10486394.

[2]. Narayana, V.L., Gopi, A.P., Patibandla, R.S.M. (2021). An Efficient Methodology for Avoiding Threats in Smart Homes with Low Power Consumption in IoT Environment Using Blockchain Technology. In: Choudhury, T., Khanna, A., Toe, T.T., Khurana, M., Gia Nhu, N. (eds) Blockchain Applications in IoT Ecosystem. EAI/Springer Innovations in Communication and Computing. Springer, Cham. https://doi.org/10.1007/978-3-030-65691-1_16

[3]. V. Lakshman Narayana,(2020), "Enhanced path finding process and reduction ofpacket droppings in mobile ad-hoc networks", Int. J. Wireless and Mobile Computing, Vol. 18, No. 4, 2020, pp-391-397.

[4]. Narayana, V.L., Gopi, A.P., Patibandla, R.S.M. (2021). An Efficient Methodology for Avoiding Threats in Smart Homes with Low Power Consumption in IoT Environment Using Blockchain Technology. In: Choudhury, T., Khanna, A., Toe, T.T., Khurana, M., Gia Nhu, N. (eds) Blockchain Applications in IoT Ecosystem. EAI/Springer Innovations in Communication and Computing. Springer, Cham. https://doi.org/10.1007/978-3-030-65691-1_16

[5]. Chaitanya, K., and S. Venkateswarlu. "DETECTION OF BLACKHOLE & GREYHOLE ATTACKS IN MANETs BASED ON ACKNOWLEDGEMENT BASED APPROACH." Journal of Theoretical & Applied Information Technology 89.1 (2016).

[6]. Lakshman Narayana, V., Rao, G.S., Gopi, A.P., Lakshmi Patibandla, R.S.M. (2022). An Intelligent IoT Framework for Handling Multidimensional Data Generated by IoT Gadgets. In: Al-Turjman, F., Nayyar, A. (eds) Machine Learning for Critical Internet of Medical Things. Springer, Cham. https://doi.org/10.1007/978-3-030-80928-7_9

[7]. Narayana, V. L., et al. "Computer Tomography Image Based Interconnected Antecedence Clustering Model Using Deep Convolution Neural Network for Prediction of COVID-19." Traitement du Signal, vol. 40, no. 4, 2023, pp. 1689–1696. https://doi.org/10.17762/ijritcc.v11i9s.73

[8]. Sujatha, V., Vasumathi Devi Majety, Satya Sandeep Kanumalli, and V. S. Sai Rama Krishna Komanduri. "Brain Tumour Detection Using Auto-Encoder and Multi-Layer Perception." AIP Conference Proceedings, vol. 2724, no. 1, AIP Publishing, 28 Apr. 2023. https://doi.org/10.1063/5.0130160

[9]. Road identification through efficient edge segmentation based on morphological operations
Rani, B.M.S., Majety, V.D., Pittala, C.S., ... Sandeep, K.S., Kiran, S.
Traitement du Signal, 2021, 38(5), pp. 1503–1508

[10]. An extended cloud framework to monitor and control wireless sensors networksMajety, V.D., Sravanthi, G.L., Didla, D. International Journal of Innovative Technology and Exploring Engineering, 2019, 8(11), pp. 3805–3808

[11]. V. Pavani, N. VijayaLakshmi, N. Harika, G. S. Sowjanya and V. Deepthi, "Deep Learning-based Analysis of Brain MRI for Enhanced Diagnosis of Multiple Sclerosis," 2024 5th International Conference on Data Intelligence and Cognitive Informatics (ICDICI), Tirunelveli, India, 2024, pp. 1141-1148, doi: 10.1109/ICDICI62993.2024.10810928.

[12]. Kumari, G. R. P., Reddy, A. H., Lakshmi, K., Abhinaya, B., Sanjana, S., & Naresh, A. (2024, March). Time-Frame-Based Drowsiness Detection System Using CNN. In 2024 2nd International Conference on Disruptive Technologies (ICDT) (pp. 711-716). IEEE.

[13]. Sirisha, Aswadhati, B. Siva Jyothi, and P. Sandhya Krishna. "Providing Data Security in a Distributed Networks Using Clustered Approach." International Journal of Advanced Science and Technology 28, no. 16 (2019): 1907-1915.

[14]. Arumugham, V., Sankaralingam, B. P., Jayachandran, U. M., Krishna, K. V. S. S. R., Sundarraj, S., & Mohammed, M. (2023). An explainable deep learning model for prediction of early-stage chronic kidney disease. Computational Intelligence, 39(6), 1022-1038.

[15]. Rayachoti, Eswaraiah, Sudhir Tirumalasetty, and Silpa Chaitanya Prathipati. "Watermarking system for telemedicine based on FABEMD." Multimedia Tools and Applications 81.30 (2022): 44383-44404.

[16]. Kavishwar, S. (2011). Pension funds as an infrastructure financing avenue: An exploratory study. *Management Dynamics*, *11*(2), 33-45.

[17]. Bidwaikar, V. N., & Kavishwar, D. S. (2012). Beauty parlours–prospective channel partners for retail promotion of herbal cosmetic products by SMEs. *Indian Streams Research Journal.* 2(1), 1-4

[18]. Shahu, A., Tiwari, H., Joshi, M., & Kavishwar, S. An Analysis of the Effectiveness of Index ETFS and Index Derivatives in Covered Call Strategy. *Journal of Informatics Education and Research.* 4(3), 42-48.

[19]. Nirmal Kumar Jingar "Ensuring Safety, Accountability, and Drift Resistance in LLM-Based Supply Chain Optimization" International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Print ISSN : 2395-1990, Online ISSN : 2394-4099, Volume 10, Issue 1, pp.472-482, January-February-2023. Available at doi : https://doi.org/10.32628/IJSRSET2310372

[20]. Jingar, N. K. (2026, February 13). Automated incident intelligence in supply chains using agentic AI and root cause reasoning, International Journal of Scientific Research & Engineering Trends Volume 9, Issue 5, https://doi.org/10.5281/zenodo.18162511

[21]. Nijim, M., Kanumuri, V., Alaqqad, W., Albataineh, H. (2023). Advanced Traffic Management System for Smart Cities. In: Daimi, K., Al Sadoon, A. (eds) Proceedings of the 2023 International Conference on Advances in Computing Research (ACR'23). ACR 2023. Lecture Notes in Networks and Systems, vol 700. Springer, Cham. https://doi.org/10.1007/978-3-031-33743-7_19

[22]. Nijim, M., Kanumuri, V., Al Aqqad, W., Albataineh, H. (2024). Machine Learning Based Analysis of Cyber-Attacks Targeting Smart Grid Infrastructure. In: Daimi, K., Al Sadoon, A. (eds) Proceedings of the Second International Conference on Advances in Computing Research (ACR'24). ACR 2024. Lecture Notes in Networks and Systems, vol 956. Springer, Cham. https://doi.org/10.1007/978-3-031-56950-0_28

[23]. Racha, Ganesh. "Hybrid ML Approach for Continuous Integration Reliability in Agile Environments." United International Journal of Engineering and Sciences (UIJES), vol. 5, no. 3, 2025, pp. 9–21.

[24]. Racha, Ganesh. "Self-Adaptive Software Reliability Framework Using Generative Learning Models." International Journal for Modern Trends in Science and Technology, vol. 12, no. 1, 2026, pp. 30–37.

[25]. Veginati, Navya. "Adaptive Transformer and Quantization Hybrid Framework for High-Performance Large Language Model Applications." United International Journal of Engineering and Sciences, vol. 5, no. 4, Dec. 2025, pp. 46–56

[26]. Veginati, Navya. "Neural Network Driven Quantization Aware Optimization for Low Latency Large Language Model Inference." International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 10, no. 3, May-June 2024, pp. 1162–1170, doi:10.32628/CSEIT25113584.

[27]. Jonnalagadda, P.K. (2026). Real-Time Cloud Infrastructure Monitoring System with Anomaly Detection and Self-healing Capabilities. In: Kumar, V.N., Senkerik, R., Prasad, V.K., Kumar, T.K. (eds) Intelligent Computing and Communication. ICICC 2025. Lecture Notes in Networks and Systems, vol 1839. Springer, Cham. https://doi.org/10.1007/978-3-032-18349-1_43

[28]. Jonnalagadda, Pawan Kalyan. "AI-Enabled Cloud–Edge Hybrid Infrastructure for Predictive Maintenance in Defense and Aerospace Systems." International Journal of Science, Engineering and Technology, vol. 12, no. 2, 2024.

[29]. Ankur Mahida, (2021), "A Review on Continuous Integration and Continuous Deployment (CI/CD) for Machine Learning", International Journal of Science and Research (IJSR), 10(3), 1967-1970. https://dx.doi.org/10.21275/SR24314131827, https://www.ijsr.net/getabstract.php?paperid=SR24314131827

[30]. "Mahida, A. (2022). Comprehensive Review on Optimizing Resource Allocation in Cloud Computing for Cost Efficiency. Journal of Artificial Intelligence & Cloud Computing. SRC/JAICC-249. DOI: doi. org/10.47363/JAICC/2022 (1), 232, 2-4."

[31]. Tummuri, S. S. R. (2024). Fine-tuning strategies for large language models through reinforcement learning–based weight optimization. International Journal of Science, Engineering and Technology. Volume 4, Issue 3.

[32]. Tummuri, S. S. R. (2024). Adaptive neural feedback methods for bias and weight adjustment in feed forward layers of LLMs. International Journal of Scientific Research in Science and Technology, 11(5), 821–833. https://doi.org/10.32628/IJSRST52310380

[33]. Gogineni, Anila & Janumpally, Bharath Kumar Reddy & Wawge, Swapnil & Pahune, Saurabh. (2025). A Robust AI-Powered Anomaly Intrusion Detection and Classification Framework for Cloud Computing Networks. 1-6. 10.1109/INDISCON66021.2025.11253743.

[34]. A. Joon, B. K. R. Janumpally, A. Gogineni and P. Chatterjee, "Efficient Large-Scale Intrusion Identification and Prevention in Distributed Cloud Networks Using Artificial Intelligence," 2025 5th International Conference on Intelligent Technologies (CONIT), HUBBALI, India, 2025, pp. 1-8, doi: 10.1109/CONIT65521.2025.11167760.

[35]. Arora AS, Yachamaneni T, Kotadiya U. A Comprehensive Analytical Framework for Modeling Consumer Credit Card Behavior and Risk Profiling Using Advanced Financial Metrics. IJAIDSML [Internet]. 2022 Jun. 30 [cited 2026 Apr. 2];3(2):90-100.

[36]. Arora AS, Yachamaneni T, Kotadiya U. Optimizing Multi-Tenant Resource Allocation in Cloud-Based Distributed Systems for Large-Scale AI Model Training Using In-Memory Computing. IJERET [Internet]. 2021 Mar. 30 [cited 2026 Apr. 2];2(1):37-46.