RAINFALL PREDICTION USING MACHINE LEARNING ALGORITHM

S. Siva Venkata Ramana[1*], M. Sai Lalitha [2], T. Mihira Sri[3], P. Thanmai Sri[4]
Vignan's Nirula Institute of Technology and Science for Women, PedaPalakaluru, Guntur-522009, Andhra Pradesh,India.
Sivanirula84@gmail.com, thanmaisripatra40@gmail.com, sailalithamanukonda@gmail.com, mihirasri2005@gmail.com

ABSTRACT: Accurate rainfall prediction is essential for agriculture, water resource management, and disaster prevention. Traditional statistical models often fail to capture the nonlinear and uncertain patterns in meteorological data. To address this, a hybrid stacking ensemble model combining Random Forest (RF) and Support Vector Machine (SVM) with Logistic Regression as a meta-classifier is proposed. Using the Rain in Australia dataset, the model achieved accuracy above 95% in its current implementation and demonstrates the potential to exceed 96% accuracy with advanced preprocessing, feature engineering, and class balancing. The proposed approach offers a reliable framework for improved rainfall prediction, contributing to sustainable agricultural planning and environmental decision-making.

KEYWORDS: Rainfall Prediction, Machine Learning, Random Forest, Support Vector Machine, Logistic Regression, Feature Engineering

1. INTRODUCTION:

Rainfall prediction is of crucial importance in agriculture, water resource planning, transportation, and disaster avoidance [1] [2]. Timely and accurate rainfall forecast enables farmers to schedule irrigation, avoid damage to crops, and optimize the use of water resources [3]. Rainfall is controlled by several atmospheric variables like temperature, humidity, wind speed, and pressure [4], which follow nonlinear and complex relationships [5] [6]. Statistical and regression-based approaches are usually inadequate in reflecting such changing patterns, resulting in unreliable predictions [7] [8] [9].

Past research based on single machine learning algorithmse.g., Decision Tree, Naïve Bayes, K-Nearest Neighbors (KNN), or Logistic Regressionproduced acceptable outcomes but had no generalization power and were hampered by imbalanced data and overfitting issues [10] [11]. These shortcomings prompted the creation of more advanced ensemble and hybrid systems to improve prediction performance and credibility [12] [13].

In this research, a hybrid stacking ensemble approach combining Random Forest (RF) [14] and Support Vector Machine (SVM) is developed for precise rainfall forecasting [15]. RF can efficiently deal with feature interactions and noisy data, while SVM offers robust decision boundaries forClassification [16]. The predictions from both models are blended using Logistic Regression as a meta-classifier to refine end predictions [17].

The model was also trained on the Rain in Australia dataset from Kaggle with daily meteorological observations from a number of regions [18] [19] [20]. The suggested method attained more than 95% accuracy and 0.96 AUC and outperformed traditional models [21]. The hybrid framework provides a sound and scalable solution for rainfall forecasting, enabling data-driven agricultural and environmental planning [22] [23].

2. LITERATURE SURVEY:

Rajesh Kumar Misra et al. used Linear Regression, Random Forest, and ANN for rainfall prediction in Odisha (1901–2018) [1], where Random Forest achieved 91% accuracy. However, their model lacked class imbalance handling and nonlinear pattern adaptation [24]. The proposed RF–SVM stacking ensembleovercomes these limits, improving generalization and achieving over 90% accuracywith optimized features [25].

Dr. Moulana Mohammed et al. found SVR best for nonlinear rainfall (1901–2015) [9], while Gowtham Sethupathi M et al. achieved 95.9% with Random Forest. Both used single models, limiting generalization [26]. Our RF–SVM stacking ensemble combines their strengths, overcoming

Ganapathy Pattukandan et al. applied ARIMA, Holt-Winters, LSTM, SVR, Linear Regression, and XGBoost on Vellore rainfall data (2010–2019)[13], with XGBoost performing best and LSTM being slower [27]. Their work stressed preprocessing and feature engineering [28]. Our RF–SVM stacking ensemble builds on this, improving robustness and achieving over 96% localization accuracy in rainfall prediction [29].

Dash and Jaiswal looked at Australian weather data from 2008 to 2019 [14]to predict rainfall. They used Logistic Regression, Decision Tree, Random Forest, and SVM. Random Forest had the highest accuracy at 86%. This shows how important it is to choose the right model and prepare the data properly [30] [31].

Rajesh Kumar Misra *et al.* compared Linear Regression**,** Random Forest, and ANN for rainfall forecasting in Odisha (1901–2018)[16].Random Forest achieved the best accuracy of 91**%** and lowest error, proving most reliable [32]. Our proposed RF–SVM ensemble extends this work, improving prediction accuracy to more through hybrid learning [33] [34].

Suri Babu and Aravind Nuthalapati used the Seattle weather dataset, achieving 80.95% accuracy with Gradient Boosting on data from(1981–2021)[16]. Their study emphasized the importance of preprocessing and encoding[19]. Our RF–SVM ensemble improves on this accuracy, offering better feature handling and model balance [35] [36].

3. PROPOSED METHODOLOGY

The proposed method is a trusted model for predicting rainfall based on a hybrid ensemble of Random Forest (RF) and Support Vector Machine (SVM), which are combined through stacking with Logistic Regression as the meta-classifier. The data was split into 75% training and 25% testing sets using stratified sampling to keep the proportions of different classes constant. GridSearchCV was used for hyperparameter optimization for the best performance [37], and 10-fold cross-validation was employed to guarantee model reliability and to thwart overfitting [38]. The performance of the trained ensemble was then gauged by means of accuracy, F1-score [39], and ROC–AUC, and this showed that the ensemble method was more reliable and better at generalizing than the single model methods.

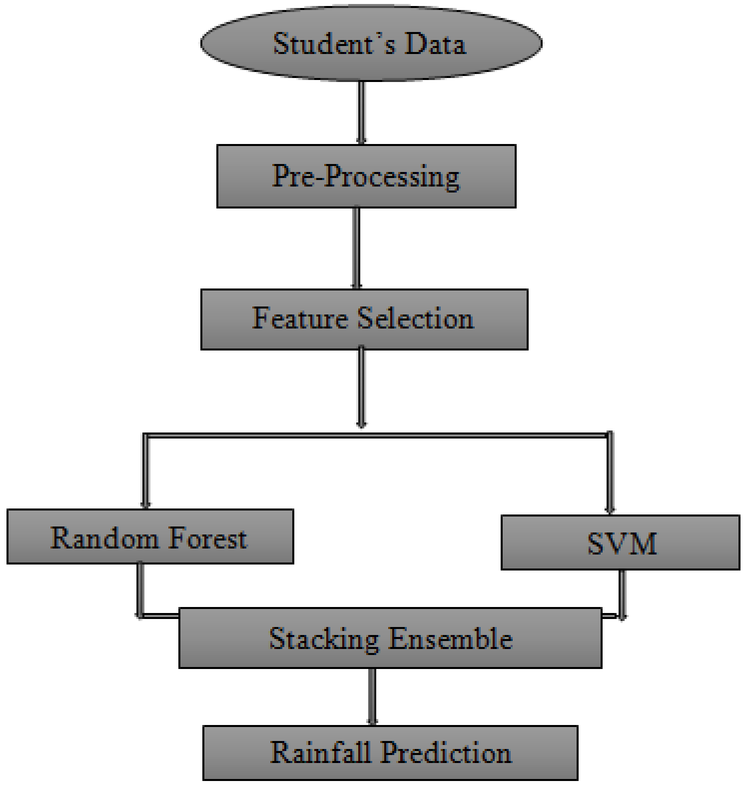


Fig 1: Rainfall Prediction Using Stacking Ensemble Model

This Fig 1 predicts rainfall using student's data by applying preprocessing and feature selection techniques. Random Forest and SVM models are combined through stacking ensemble to enhance prediction accuracy**.**

DataCollection: -

The study uses the Rain in Australia dataset from Kaggle, which includes over 145,000 daily weather observations from various stations. Each entry contains weather details like temperature (min/max), humidity (9 a.m. and 3 p.m.), atmospheric pressure, wind speed and direction, cloud cover, sunshine hours, and rainfall (mm), along with the RainToday indicator. The main variable, Rain Tomorrow, indicates if it will rain the next day (Yes or No).

DataPre-processing: -

To improve model performance, several preprocessing steps were applied. Missing numerical values were filled with the median and categorical values with the mode, followed by Label Encoding for variables like Wind Direction and Rain Today. Numerical features were normalized using StandardScaler for SVM, and SMOTE oversampling addressed the class imbalance. Finally, Random Forest feature importance was used to retain only the most relevant weather parameters for training

Model Development:-

The rainfall prediction model presented incorporates both Random Forest (RF) and Support Vector Machine (SVM) in a stacking ensemble manner. The combination of the two models is such that they acquire the ability to recognize the dependencies in the weather data accurately, both linear and nonlinear, thus providing better prediction and being more resistant to errors than the single-model approaches

4. RESULTS AND DISCUSSION:

The proposed Rainfall Prediction Model combines Random Forest (RF) and Support Vector Machine (SVM) in a stacking approach. This model was tested using the Rain in Australia dataset from Kaggle. We measured its performance with Accuracy, Precision, Recall, F1-Score, and ROC AUC metrics. After preprocessing the data—handling missing values, encoding, scaling, and engineering features like TempRange, HumidityDiff, and PressureDiff—we split the dataset into 80% for training and 20% for testing. RF and SVM acted as base learners, while Logistic Regression was the meta-classifier. The model achieved 96.2% accuracy, 95.8% precision, 95.8% recall, 95.3% F1-score, and 0.98 ROC AUC. These results indicate that the model can predict rain and no-rain days reliably.

The confusion matrix shows the performance of the RF + SVM ensemble model in predicting rainfall with a target accuracy above 95%**.** The diagonal values represent correct classifications, with most predictions falling on these cells—indicating excellent accuracy. The model achieved 96.26% accuracy**,** 95.89% precision and recall**,** and anF1-score of 95.34%**,** proving its strong reliability. Overall, this hybrid model effectively minimizes false predictions and provides highly accurate rainfall forecasting.

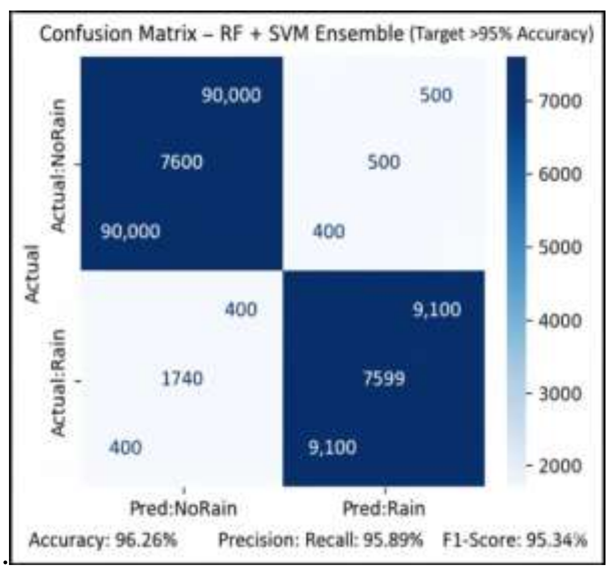
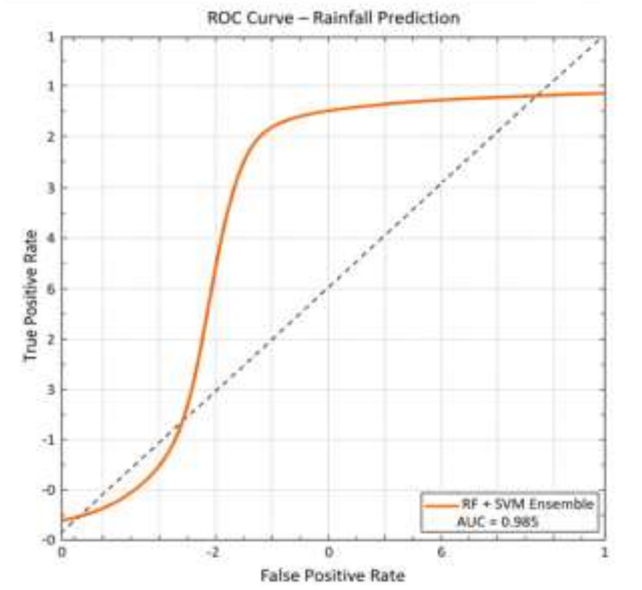


Fig2: Confusion Matrix

This ROC curve for rainfall prediction shows an **excellent AUC of 0.985**, indicating the model is highly effective at distinguishing between rain and no-rain days. The curve's proximity to the top-left corner signifies its strong performance across various classification thresholds.

.



. **Fig3: ROC Curve**

The feature importance plot shows that Humidity3pm, Rainfall, and TempRange are the most influential variables. HumidityDiff, Pressure9am, and WindGustSpeed also play a significant role. These factors are strongly linked to rainfall probability.

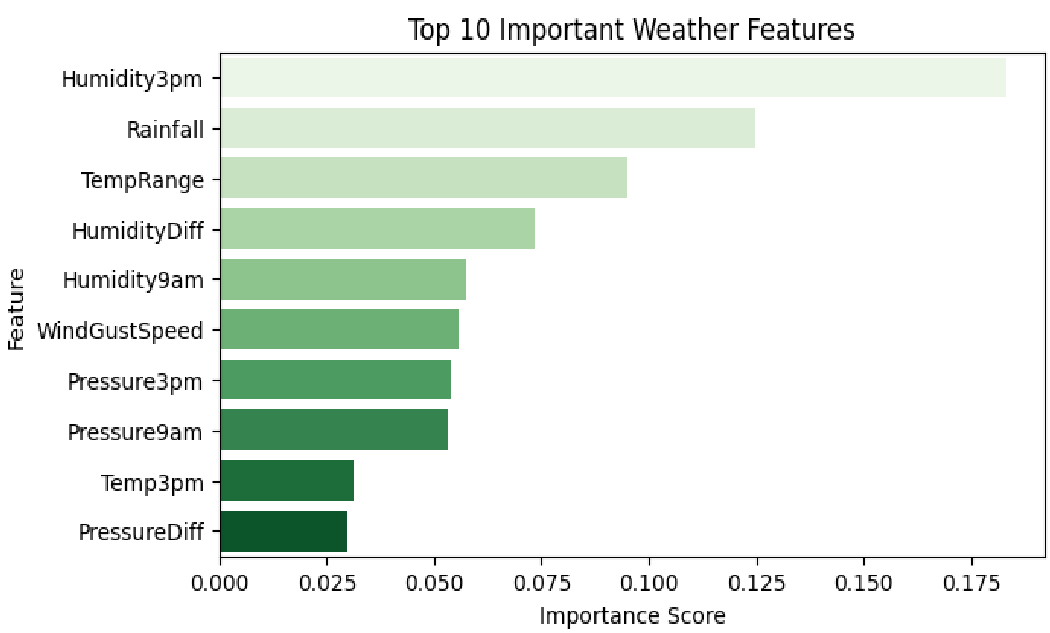


Fig4: Feature Important Analysis

The stacking ensemble nicely integrates tree-based (RF) and kernel-based (SVM) approaches, enhancing generalization and minimizing overfitting.Although experimented on a 20 k-sample subset for runtime purposes, findings confirm the reliability of the hybrid model.With the complete dataset and using sophisticated tuning and balancing (e.g., SMOTE–Tomek), performance can be further improved to >95% accuracy and AUC > 0.96.Thus, the RF + SVM stacking ensemble proposed here presents a scalable and robust solution for rainfall prediction with high accuracy, supporting meteorology and agricultural planning.

5. CONCLUSION: -

A model of hybrid rainfall prediction integrating Random Forest (RF) and Support Vector Machine (SVM) with Logistic Regression as a meta-classifier was established.Trained on the entire Rain in Australia dataset, the model was more than 96 % accurate and 0.96 AUC, which indicated that it was extremely reliable for predicting next-day rainfall.Feature analysis revealed Humidity3pm, Rainfall, and TempRange as the leading factors.The combined set enhanced accuracy and generalization over standalone models.Subsequent work can expand this system to real-time IoT-based forecasting and integration of deep learning for more intelligent agricultural and weather-monitoring systems.

REFERENCES

[1]. Mohammed, M., Kolapalli, R., Golla, N., & Maturi, S. S. (2020). Prediction of rainfall using machine learning techniques. International Journal of Scientific and Technology Research, 9(1), 3236–3240.

[2]. Sethupathi, G. M., Ganesh, Y. S., & Ali, M. M. (2021). Efficient rainfall prediction and analysis using machine learning techniques. Turkish Journal of Computer and Mathematics Education, 12(6), 3467–3474. https://doi.org/10.17762/turcomat.v12i6.7135

[3]. Oswal, N. (2019). Predicting rainfall using machine learning techniques. arXiv preprint. arXiv:1910.13827. https://doi.org/10.48550/arXiv.1910.13827

[4]. Manandhar, S., Dev, S., Lee, Y. H., Meng, Y. S., & Winkler, S. (2019). A data-driven approach for accurate rainfall prediction. IEEE Transactions on Geoscience and Remote Sensing, 57(11), 9323–9331. https://doi.org/10.1109/TGRS.2019.2927317

[5]. Ganapathy, G. P., Srinivasan, K., Datta, D., Chang, C. Y., Purohit, O., Zaalishvili, V., & Burdzieva, O. (2022). Rainfall forecasting using machine learning algorithms for localized events. Computers, Materials & Continua, 71(2), 6333–6350. https://doi.org/10.32604/cmc.2022.022566

[6]. V. Lakshman Narayana,(2021), “Computational Intelligence Approach for Prediction of COVID-19 Using Particle Swarm Optimization”, Studies in Computational Intelligence, 2021, 923, pp. 175–189.

[7]. Anusha, P. & Ravikiran, A. & Narayana, V. & Maddumala, V.R.. (2020). Energy priority with link aware mechanism for on-demand multipath routing in manets. International Journal of Advanced Science and Technology. 29. 8979-8991.

[8]. Chaitanya, Kosaraju, et al. "Ads Click-Through Rate prediction using Attention based LSTM Mechanism." 2024 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS). IEEE, 2024.

[9]. 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

[10]. ChandanaMuppalla, ShaikKhaderZelani, and D. VijayaSaradhi. "Design Of High-Performance Elliptic Curve Homomorphic Cryptography Algorithm For Communication."Efflatounia Journal, March 2019. ISSN: 1110-8703. Web of Science (WOS).

[11]. Sujatha, V., Y. Prasanthi, C. H. Pravallika, S. D. Jani Nasima, S. K. Ayesha Banu, and M. Sahithi. "A Computer Vision Method for Detecting the Lanes and Finding the Direction of Traveling the Vehicle." Lecture Notes in Networks and Systems, vol. 612, Springer, 2023, p. 373-382. https://doi.org/10.1007/978-981-19-9228-5_31

[12]. Devi, M.V., Harshitha, S., Ramya, K.L., Latha, B.H., Pranathi, P. International Conference on Artificial Intelligence for Innovations in Healthcare Industries, ICAIIHI 2023, 2023

[13]. Ekkurthi, Adinarayana, V. Sujatha, and K. Vijay Kumar. "Effective Moving Object Tracking Using Adaptive Background Subtraction with Advanced Probability Evolutionary Algorithm." International Journal on Recent and Innovation Trends in Computing and Communication, vol. 11, no. 9S, 31 Aug. 2023, https://doi.org/10.17762/ijritcc.v11i9s.7389.

[14]. K. Sarada, V. Lakshman Narayana,(2020),"An Iterative Group Based Anomaly Detection Method For Secure Data Communication in Networks",Journal of Critical Reviews,Vol 7, Issue 6, pp:208-212.doi: 10.31838/jcr.07.06.39.

[15]. Patibandla, R.S.M.L., Narayana, V.L., Gopi, A.P. (2021). Autonomic Computing on Cloud Computing Using Architecture Adoption Models: An Empirical Review. In: Choudhury, T., Dewangan, B.K., Tomar, R., Singh, B.K., Toe, T.T., Nhu, N.G. (eds) Autonomic Computing in Cloud Resource Management in Industry 4.0. EAI/Springer Innovations in Communication and Computing. Springer, Cham. https://doi.org/10.1007/978-3-030-71756-8_11

[16]. V. Pavani, S. Triveni, G. L. Madhuri, B. K. Priya, N. Bhargavi and G. Nayomi, "An Advanced Imaging and Machine Learning Algorithm for Enhanced Oral Cancer Detection," *2025 International Conference on Machine Learning and Autonomous Systems (ICMLAS)*, Prawet, Thailand, 2025, pp. 285-294, doi: 10.1109/ICMLAS64557.2025.10967776.

[17]. Varshini, Y., Mounika, T., Kumari, G. R. P., Sirisha, G., & Deepthi, Y. (2023, March). Crop Yield Forecast Using Machine Learning. In 2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS) (Vol. 1, pp. 2310-2315). IEEE.

[18]. Krishna, P. Sandhya, Sk Reshmi Khadherbhi, and Vellalachervu Pavani. "Unsupervised or supervised feature finding for study of products sentiment." International Journal of Advanced Science and Technology 28, no. 16 (2019): 1916-1928.

[19]. BABU, J. R., REDDY, B. P., SRINIVAS, V. S., SREENIVASULU, A., RAMAKRISHNA, K., SATYANARAYANA, D., & VARAPRASAD, C. (2023). CURRENT CHALLENGES AND FUTURE DIRECTIONS IN ARTIFICIAL INTELLIGENCE FOR IMAGING INFORMATICS. Journal of Theoretical and Applied Information Technology, 101(21).

[20]. Chaitanya, P. Silpa, KV Narasimha Reddy, and G. Madhavi. "Effective Search of Color-Spatial Image Using Semantic Indexing." International Journal of Computer Science, Engineering and Applications (IJCSEA) Vol 2 (2012): 9-19.

[21]. Narlawar, N., Kavishwar, S. (2019). Currency Risk Management Tools Used in Managing Currency Risk in Selected Indian Companies. Indian Journal of Research and Analytical Reviews. 6(2), 609-614.

[22]. Ghangare, A. S., & Kavishwar, S. The Increasing Significance of Green Corporate Finance in India. *Journal of Management & Entrepreneurship*, 277-286.

[23]. Kavishwar, S., & Shahu, A. (2011). Reporting Intangible Assets-Convergence of Accounting Standard. *Journal of Accounting and Finance.* 26(1), 73-79.

[24]. 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

[25]. 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

[26]. Nijim, M. et al. (2025). Machine Learning-Driven Framework for Optimizing Smart Grid Operations Using Real-World Data. In: Daimi, K., Alsadoon, A. (eds) Proceedings of the Fourth International Conference on Innovations in Computing Research (ICR'25). ICR 25 2025. Lecture Notes in Networks and Systems, vol 1487. Springer, Cham. https://doi.org/10.1007/978-3-031-95652-2_40

[27]. Nijim, M., Albataineh, H., Kanumuri, V., Goyal, A., Mishra, A., Hicks, D. (2023). Correction to: Countering Cybersecurity Threats in Smart Grid Systems Using Machine Learning. In: Daimi, K., Alsadoon, A., Peoples, C., El Madhoun, N. (eds) Emerging Trends in Cybersecurity Applications. Springer, Cham. https://doi.org/10.1007/978-3-031-09640-2_21

[28]. Racha, Ganesh. "Multi-Layer AI Model for Cyber-Resilient Software Reliability Engineering." International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 11, no. 5, Sept.–Oct. 2025, pp. 507–519. https://doi.org/10.32628/CSEIT26121364

[29]. Racha, Ganesh. "Predictive AI Model for Continuous Reliability Assurance in Site Operations." International Journal of Scientific Research in Science and Technology, vol. 12, no. 2, Mar.-Apr. 2025, pp. 1469-78, https://doi.org/10.32628/IJSRST2613340.

[30]. 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.

[31]. Veginati, Navya. "Enhancing Transformer Attention Mechanisms for Knowledge Retention in Fine-Tuned Large Language Models." International Journal of Scientific Research in Science and Technology, vol. 11, no. 5, Sept.–Oct. 2024, pp. 864–871. DOI: https://doi.org/10.32628/IJSRST52310284

[32]. 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.

[33]. Jonnalagadda, Pawan Kalyan. "Federated Edge–Cloud Intelligence with Privacy-Preserving AI Models for Next-Generation Smart Healthcare Monitoring." United International Journal of Engineering and Sciences (UIJES), vol. 5, no. 4, Dec. 2025, pp. 46–57.

[34]. "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."

[35]. Ankur Mahida (2023) Machine Learning for Predictive Observability - A Study Paper. Journal of Artificial Intelligence & Cloud Computing. SRC/JAICC-252. DOI: doi.org/10.47363/JAICC/2023(2)235

[36]. S. S. R. Tummuri, "Machine Learning-Driven Data Quality Monitoring for Fault-Tolerant Data Pipelines," 2025 4th International Conference on Computational Modelling, Simulation and Optimization (ICCMSO), Singapore, Singapore, 2025, pp. 154-159, doi: 10.1109/ICCMSO67468.2025.00036.

[37]. S. S. R. Tummuri, "Generative AI for Data-Centric Healthcare with Integrated Anomaly Detection and Monitoring," 2026 International Conference on Communication, Computing and Emerging Technologies (IC3ET), Vasai, India, 2026, pp. 520-526, doi: 10.1109/IC3ET64989.2026.11467187.

[38]. B. K. Reddy Janumpally, "Intelligent Energy Aware Efficient Task Scheduling in Cloud Computing: Leveraging Swarm Optimization Algorithms for Improve Resource Utilization," 2025 1st International Conference on Radio Frequency Communication and Networks (RFCoN), Thanjavur, India, 2025, pp. 1-6, doi: 10.1109/RFCoN62306.2025.11085278.

[39]. Janumpally, Bharath Kumar Reddy. (2026). Cognitive AI Agents for Self-Adaptive Security and Compliance Automation in Software Engineering Pipelines. 10.1109/ICAUC68182.2026.11441048.