IJEAST

INTERNATIONAL JOURNAL
OF ENGINEERING APPLIED SCIENCE
AND TECHNOLOGY

VOLUME : 10 ISSUE : 11 Print / Issue Publication Date: 18-Apr-2026

ISSN : 2455-2143

DOI : 10.33564/IJEAST.2026.v10i11.001

Indexed In

Google SCHOLAR OPEN ACCESS doi Crossref

WWW.IJEAST.COM

editor@ijeast.com

International Journal of Engineering Applied Sciences and Technology, 2026
Vol. 10, Issue 11, ISSN No. 2455-2143, Pages 01-05
Published Online March 2026 in IJEAST (http://www.ijeast.com)

RAINFALL PREDICTION SYSTEM USING MACHINE LEARNING ALGORITHM

Ramathilagarajan N, Dr. M. Ganesan
Dept. of Computer Science & Engineering
Sri Manakula Vinayagar Engineering College, Puducherry

Abstract: **Accurate rainfall prediction is a critical challenge in meteorological science, carrying significant implications for agricultural planning, disaster risk management, and water resource allocation. Conventional numerical weather prediction (NWP) models are computationally prohibitive and inadequate for localized, real-time forecasting. This paper presents a robust ensemble machine learning framework for binary rainfall classificationRain versus No Raintrained on historical meteorological records spanning diverse Indian geographic regions from 2000 to 2025. Four classifiers are systematically evaluated: Logistic Regression, Decision Tree, Multi-Layer Perceptron (MLP), and Random Forest. Feature selection is conducted via Mutual Information scoring, isolating five critical predictors: Maximum Temperature, Wind Speed, Elevation, Latitude, and Longitude. Experimental results demonstrate that the Random Forest Classifier achieves 86.87% accuracy with a ROC-AUC of 0.943, outperforming all competing models. The trained model is deployed through a Tkinter-based desktop GUI supporting location-aware, real-time predictions for major Indian cities, augmented with an IoT-style SMTP email alert module for actionable decision support.**

Keywords: **Ensemble Learning, Mutual Information, Rainfall Prediction, Random Forest, Binary Classification, Geospatial Analysis, Decision Support System**

I. INTRODUCTION

Climate variability has intensified the socio-economic importance of precise precipitation forecasting. Irregular rainfall patterns drive agricultural losses, flood events, prolonged droughts, and compromised water security. India, with its diverse agro-climatic zones and monsoon-dependent economy, is particularly vulnerable to forecasting inaccuracies [1].

Conventional approaches primarily Numerical Weather Prediction (NWP) modelsrely on complex atmospheric simulations demanding high-performance computing. Their inability to capture non-linear dependencies among meteorological variables limits suitability for fine-grained, localized real-time forecasting [2].

Machine Learning (ML) presents a compelling data-driven alternative. Ensemble methods demonstrate superior generalization for high-dimensional, non-linear weather datasets. The Random Forest Classifier effectively mitigates overfitting while maintaining computational tractability [3]. This paper proposes a comprehensive ensemble-based rainfall prediction system combining rigorous multi-model evaluation with a user-accessible desktop GUI.

The key contributions are: **(i)** systematic study of four ML classifiers on diverse Indian data; **(ii)** Mutual Information-based feature selection; **(iii)** deployment of the optimal model in a city-aware GUI; **(iv)** integration of an IoT-style SMTP email alert mechanism.

II. LITERATURE REVIEW

Khan et al. [1] benchmarked four classifiers on Aligarh, India data, finding LR and neural networks achieved approximately 82–83% accuracy. Sarasa-Cabezuelo [4] established that city-specific training on Australian weather data substantially improves prediction accuracy. Liyew and Melese [5] demonstrated XGBoost superiority over MLR and RF for daily rainfall intensity prediction using Pearson correlation-based feature selection.

Yen et al. [6] proposed Deep Echo State Networks (DeepESN) for rainfall forecasting in southern Taiwan, achieving lower RMSE than standard ESNs. Latif et al. [7] highlighted LSTM effectiveness for multi-source meteorological data integration. Ojo and Ogunjo [8] uniquely incorporated geoclimatic coordinates (latitude, longitude) as direct model inputsan approach validated and adopted in the present work. Kumar et al. [9] confirmed ensemble algorithms consistently outperform linear models in urban environments. Markuna et al. [10] validated Random Forest for topographically complex Indian regions.

III. METHODOLOGY / PROPOSED SYSTEM

A. Dataset Description

The dataset india_weather_rainfall_data.xlsx consolidates historical records from 2000–2025 across geographically diverse Indian cities: coastal stations (Mumbai, Chennai), high-altitude hill stations (Shimla, Ooty), and plains-based urban centers (Delhi, Punjab). Features include average, minimum, and maximum temperature, wind speed,

International Journal of Engineering Applied Sciences and Technology, 2026
Vol. 10, Issue 11, ISSN No. 2455-2143, Pages 01-05
Published Online March 2026 in IJEAST (http://www.ijeast.com)

atmospheric pressure, elevation, latitude, longitude, and daily rainfall in millimeters.

B. Exploratory Data Analysis

Fig. 1 presents the distribution of daily rainfall across the dataset. The heavily right-skewed histogram confirms inherent class imbalance, with the majority of observations concentrated near zero (rain-free days). This distribution motivates the 75th-percentile threshold for binary label generation.

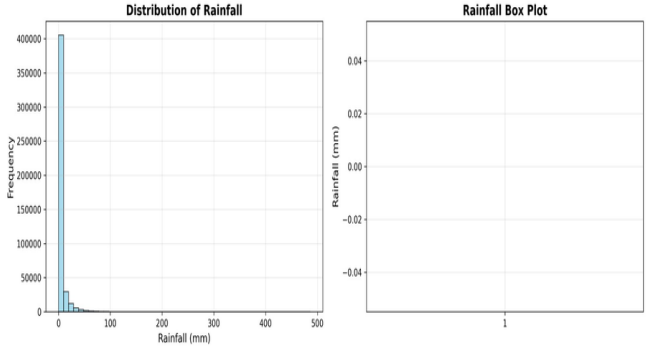

Fig. 1. Distribution of Daily Rainfall (2000–2025) showing right-skewed distribution and class imbalance.

C. Correlation Analysis

Fig. 2 presents the Pearson Correlation Matrix across all meteorological features. Strong positive correlations exist among temperature features ($r > 0.77$). A notable negative correlation between temperature and atmospheric pressure ($r = -0.70$ to -0.75) reflects fundamental thermodynamic relationships. The RainTomorrow target exhibits the strongest associations with RainToday ($r = 0.66$) and Rainfall ($r = 0.57$).

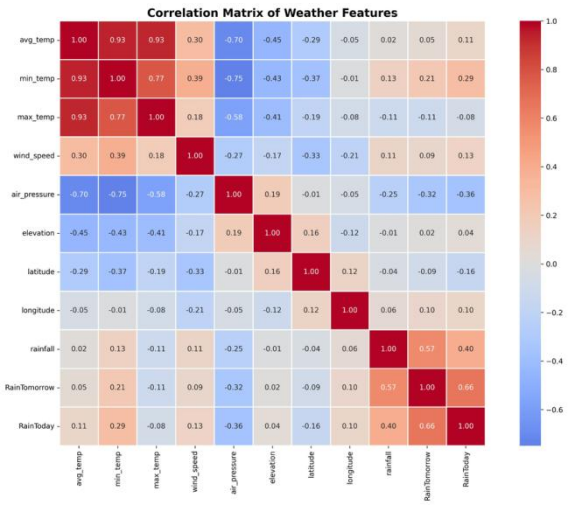

Fig. 2. Pearson Correlation Matrix Heatmap of all meteorological weather features.

D. Data Preprocessing

Raw meteorological records contain missing values attributable to sensor failures. Numerical features are imputed using column-wise median to preserve distributional properties. The binary target variable RainTomorrow is generated dynamically:

$$\text{RainTomorrow} = 1 \text{ if } \text{Rainfall} > Q75, \text{ else } 0 \quad (1)$$

Feature standardization applies the z-score transformation (Eq. 2), ensuring high-magnitude features such as elevation do not disproportionately influence model weights:

$$z = (x - \mu) / \sigma \quad (2)$$

E. Feature Selection via Mutual Information

Mutual Information (MI) quantifies the statistical dependency between each feature Xi and target Y, capturing both linear and non-linear associations:

$$I(X;Y) = \Sigma\Sigma\, p(x,y) \cdot \log[\, p(x,y) / (p(x) \cdot p(y)) \,] \quad (3)$$

The top five selected features are: Maximum Temperature, Wind Speed, Elevation, Latitude, and Longitude. Fig. 3 presents post-training Random Forest Gini Impurity importance scores confirming RainToday (≈0.35) and Air Pressure (≈0.19) as dominant predictors, consistent with MI-based selection.

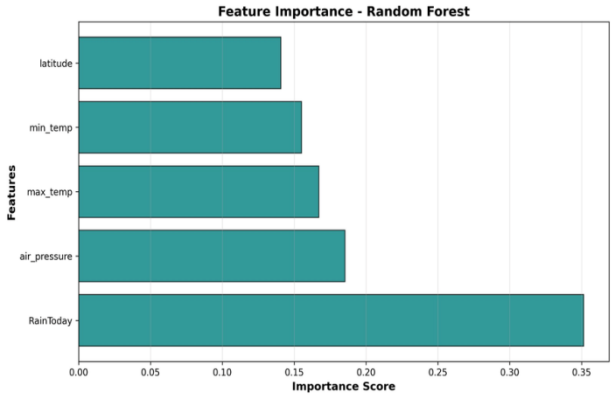

Fig. 3. Feature Importance Plot — Random Forest Gini Impurity scores confirming top predictors.

F. Classification Algorithms

Logistic Regression: Linear baseline, $P(y=1|X) = \sigma(w^T X+b)$. **Decision Tree:** Non-parametric Gini-impurity tree induction. **MLP Classifier:** Feed-forward neural network with ReLU activations. **Random Forest:** Ensemble of N=100 trees with majority voting:

$$\hat{y} = \arg\max_c \left[\Sigma\, I(hi(x) = c) \right], \quad i = 1...N \quad (4)$$

All models are trained on a 75:25 stratified split (random_state=42) and evaluated across six metrics: Accuracy, Precision, Recall, F1-Score, Cohen's Kappa, and ROC-AUC.

International Journal of Engineering Applied Sciences and Technology, 2026
Vol. 10, Issue 11, ISSN No. 2455-2143, Pages 01-05
Published Online March 2026 in IJEAST (http://www.ijeast.com)

IV. SYSTEM ARCHITECTURE

The proposed system operates as a five-stage pipeline: (1) Data Processing median/mode imputation, encoding, null elimination; (2) Feature Selection MI-based Select KBest top-5; (3) Model Training standardization and parallel training of four classifiers; (4) Model Comparison multi-metric evaluation; (5) Deployment serialized .pkl model integrated into Tkinter GUI with SMTP alert.

Table -1 System Requirements

Component	Specification
Language	Python 3.9+
ML Framework	Scikit-learn, Pandas, NumPy
GUI Framework	Tkinter (Desktop Application)
Notification	smtplib — Gmail SMTP Protocol
Serialization	pickle (.pkl model files)
Dataset	india_weather_rainfall_data.xlsx
Training Split	75% Train / 25% Test (Stratified)

V. IMPLEMENTATION

The system is implemented through five coordinated modules, all executing locally with no cloud dependency. The Tkinter GUI allows users to select any major Indian city and adjust weather parameters interactively to obtain live predictions from the serialized Random Forest model.

A. GUI Application India Live Mode

Fig. 4 shows the main interface providing city/region selection, interactive sliders for Max Temperature and Wind Speed, and a real-time results panel displaying Rain Probability, No-Rain Probability, geospatial coordinates, and context-specific agricultural recommendations.

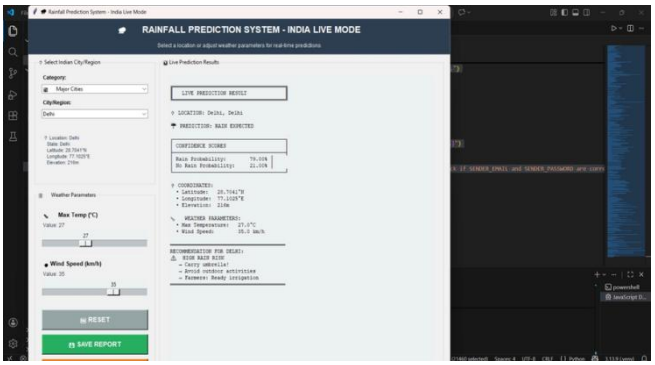

Fig. 4. GUI India Live Mode — Delhi Prediction showing 79% Rain Probability.

B. IoT Email Alert System

Fig. 5 (left) shows the email recipient input dialog and Fig. 5 (right) shows the successful delivery confirmation. Upon high-risk prediction, the SMTP module dispatches a structured alert containing prediction class, confidence scores, GPS coordinates, and actionable recommendations.

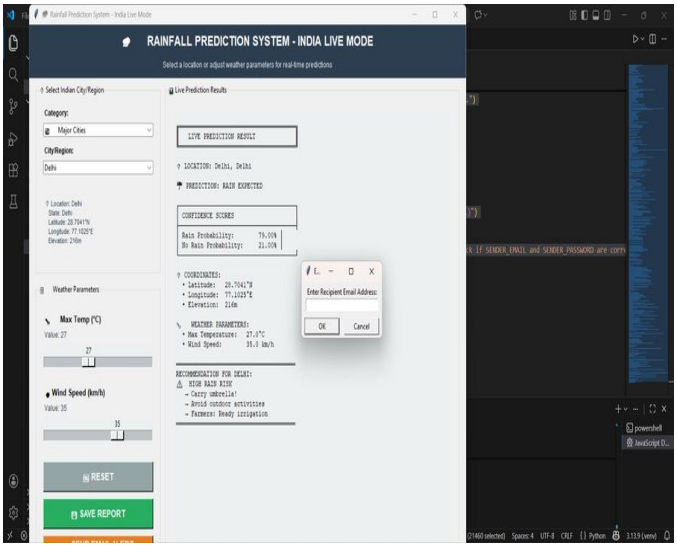

Fig. 5. Email Alert: Recipient Input Dialog (left) and Delivery Confirmation (right).

C. Email Content Received

Fig. 6 presents the structured alert email received in Gmail, containing the full prediction result, confidence scores, GPS coordinates, weather parameters, and risk-based advisory completing the IoT-style notification loop for agricultural stakeholders and authorities.

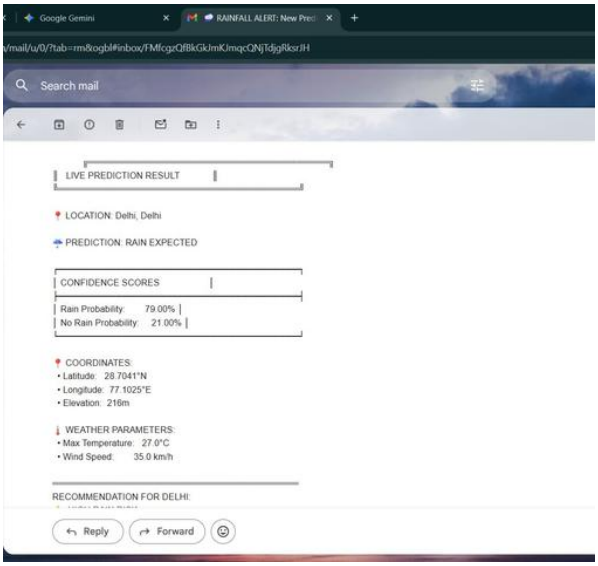

Fig. 6. Alert Email Received in Gmail — Full prediction report with recommendations.

VI. RESULTS AND DISCUSSION

A. Model Performance Comparison

Table 2 summarizes classification performance across all four models on the 25% stratified test partition. Random Forest achieves the highest Accuracy (86.87%) and ROC-

International Journal of Engineering Applied Sciences and Technology, 2026
Vol. 10, Issue 11, ISSN No. 2455-2143, Pages 01-05
Published Online March 2026 in IJEAST (http://www.ijeast.com)

AUC (0.943), confirming ensemble learning superiority. Fig. 7 presents a comprehensive six-metric visual comparison.

Table -2 Performance Comparison of All Classification Models

Model	Acc.	Prec.	Recall	F1	Kappa	AUC
Log. Reg.	84.99%	0.672	0.771	0.718	0.619	0.929
Decision Tree	86.28%	0.710	0.753	0.731	0.638	0.938
Random Forest	86.87%	0.732	0.742	0.737	0.652	0.943
MLP Classifier	86.71%	0.729	0.734	0.731	0.649	0.942

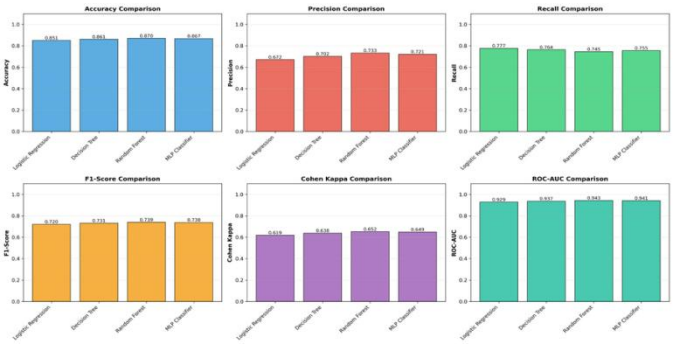

Fig. 7. Six-Metric Performance Comparison for all four classifiers.

B. ROC-AUC Analysis

Fig. 8 presents the ROC curves for all four models. All classifiers substantially outperform the random classifier baseline. Random Forest (AUC=0.943) achieves the highest area, followed closely by MLP (0.941), Decision Tree (0.937), and Logistic Regression (0.929). The steep initial rise confirms strong discrimination at high-confidence thresholds.

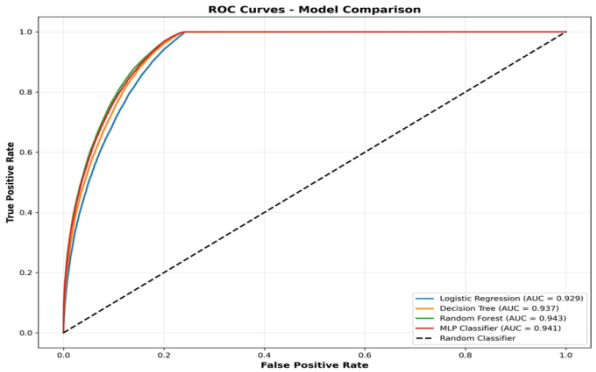

Fig. 8. ROC Curves — Model Comparison showing AUC scores for all four classifiers.

C. Confusion Matrix Analysis

Fig. 9 presents the confusion matrices for all four models on the 25% test partition. Random Forest achieves the highest True Negative count (78,889) and lowest False Positive count (7,747), while maintaining the best True Positive rate (21,259). The low False Negative rate (7,262) minimizes missed rainfall events — operationally critical for agricultural advisory and flood management applications.

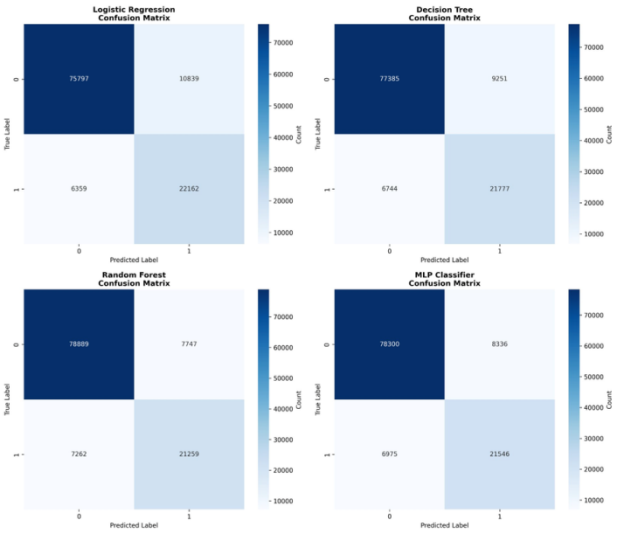

Fig. 9. Confusion Matrices for all four classifiers — Random Forest achieves best balance.

D. Discussion

Random Forest outperforms Logistic Regression by approximately 1.88 percentage points in accuracy. The inclusion of geospatial features (latitude, longitude, elevation), validated by Ojo and Ogunjo [8], proves empirically beneficial with latitude ranking among the top-5 MI predictors. Fig. 10 shows the VS Code terminal confirming successful project execution with full performance metrics for all four models.

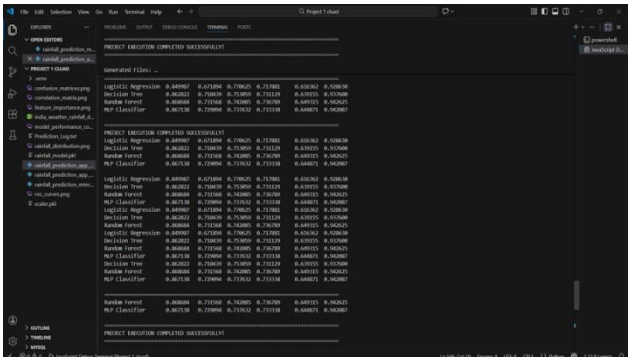

Fig. 10. VS Code Terminal — Successful project execution with performance metrics.

Table -3 System Performance Summary

Metric	Value	Model / Context
Best Accuracy	86.87%	Random Forest
Best ROC-AUC	0.943	Random Forest
Best Precision	0.733	Random Forest
Best Recall	0.777	Logistic Reg.
Best F1-Score	0.739	Random Forest
Best Kappa	0.652	Random Forest
Prediction Time	<100 ms	Local inference
Email Delivery	<5 sec	Gmail SMTP

VII. CONCLUSION

This paper presented a systematic ensemble machine learning framework for binary rainfall prediction on diverse Indian meteorological data. Through rigorous comparative evaluation of four classifiers and Mutual Information-based feature selection, the Random Forest Classifier was identified as the optimal model achieving 86.87% accuracy and ROC-AUC of 0.943. The system was successfully deployed as a location-aware desktop application with integrated IoT-style SMTP email alerts, transforming a research prototype into an actionable decision-support tool for agriculture, water management, and disaster prevention.

Future work will explore: (i) LSTM and Transformer-based architectures for multi-step temporal forecasting; (ii) real-time satellite imagery and radar data integration; (iii) adaptive incremental retraining pipelines; (iv) web-based REST API deployment; and (v) ensemble stacking with XGBoost/CatBoost for further performance gains.

VIII. REFERENCE

[1]. Khan M.U.S., Saifullah K.M., Hussain A., Azamathulla H.M. (2024), Comparative analysis of different rainfall prediction models: A case study of Aligarh City, India, Results Engineering.

[2]. Rahman A.U. et al. (2022), Rainfall prediction system using machine learning fusion for smart cities, Sensors.

[3]. Ghosh S., Gourisaria M.K., Sahoo B., Das H. (2023), A pragmatic ensemble learning approach for rainfall prediction, Discover Internet of Things.

[4]. Sarasa-Cabezuelo A. (2022), Prediction of rainfall in Australia using machine learning, Information.

[5]. Liyew C.M., Melese H.A. (2021), Machine learning techniques to predict daily rainfall amount, Journal of Big Data.

[6]. Latif S.D. et al. (2023), Assessing rainfall prediction models: Exploring advantages of machine learning and remote sensing, Alexandria Engineering Journal.

[7]. Ojo O.S., Ogunjo S.T. (2022), Machine learning models for prediction of rainfall over Nigeria, Scientific African.

[8]. Kumar V. et al. (2023), A comparison of machine learning models for predicting rainfall in urban metropolitan cities, Sustainability.

[9]. Markuna S. et al. (2023), Application of innovative machine learning techniques for long-term rainfall prediction, Pure and Applied Geophysics.

[10]. Mishra P., Al Khatib A.M.G., Yadav S. et al. (2024), Modeling and forecasting rainfall patterns in India using XGBoost algorithm, Environmental Earth Sciences.

[11]. Babu A., Srivastava P., Patel S. (2024), Predicting rainfall using machine learning and deep learning models across altitudinal gradients, Scientific Reports.

[12]. Poornima K., Pushpalatha S., Jana R.B. (2023), Rainfall forecast and drought analysis using LSTM models in India, Water.

[13]. Mathpal T., Rajput R.K.S., Kunwar B., Pandey S. (2024), Development of ensemble probabilistic machine learning models for rainfall prediction, Springer Conference Proceedings.

[14]. Darji M., Dave J.A. (2024), Rainfall prediction in diverse Indian regions using machine learning approaches, Springer Lecture Notes.

[15]. Srinu N., Bindu B.H. (2022), Review on machine learning and deep learning based rainfall prediction methods, IEEE Conference.

IJEAST

INTERNATIONAL JOURNAL
OF ENGINEERING APPLIED SCIENCE
AND TECHNOLOGY

ABOUT IJEAST

International Journal of Engineering Applied Science and Technology (IJEAST) is a peer-reviewed, open access journal that publishes high-quality research papers in the field of Engineering, Applied Science and Technology.

IJEAST aims to provide a platform for researchers, academicians, and professionals to share their innovative ideas, research findings, and practical experiences with the global scientific community.

FOCUS AREAS

- Engineering
- Applied Science
- Technology
- Innovation & Development
- Interdisciplinary Studies

PEER REVIEWED
All submissions are rigorously peer reviewed to ensure quality.

OPEN ACCESS
Free and unrestricted access to research for all.

GLOBAL REACH
Connecting researchers and professionals worldwide.

TIMELY PUBLICATION
We ensure a swift and efficient publication process.

For more information, visit our website
www.ijeast.com

INTERNATIONAL JOURNAL
OF ENGINEERING APPLIED SCIENCE
AND TECHNOLOGY

editor@ijeast.com
www.ijeast.com
India

2455-2143