A Machine Learning Framework for Forest Fire Prediction in the Nallamala Forest Using NDVI and Synthetic Weather Data

Sivaratri Siva Nageswara Rao[1], Gairuboina Naveen Kumar[2], Dogiparthi Venkata Sai Girish[3], Sanikommu Nirupam Reddy[4], B Sankara Babu [5], Kalyani Nara [6], Dodda Venkata Reddy[7]
[1,2,3,4,7]Department of Computer Science and Engineering,
Narasaraopeta Engineering College(Autonomous), Narasaraopet, Andhra Pradesh, India
[5]Department of Computer Science and Engineering, GRIET, Hyderabad,Telangana, India
[6]Department of Computer Science and Engineering,
G.Narayanamma Institute of Technology & Science(women), Shaikpet, Hyderabad,Telangana, India
[1]profssnr@gmail.com, [2]gairuboina.naveenkumar45@gmail.com, [3]doghiparthigirish@gmail.com,
[4]niruppamreddysanikommu@gmail.com, [5]sankarababu.b@griet.ac.in, [6]nara.kalyani@gnits.ac.in,
[7]doddavenkatareddy@gmail.com

***Abstract*—This project offers a machine learning-based methodology for early warning and prediction of forest fires in India's ecologically rich Nallamala Forest region. Leveraging remotely sensed NDVI data to capture vegetation dynamics and health patterns and synthetically generated weather data from 2012 to 2025, the research constructs a strong model to classify fire events. The pipeline combines MODIS HDF-format NDVI time series with historical temperature and humidity patterns, supplemented by engineered lag features. Ground-truth fire events are obtained from MODIS and VIIRS fire archive data sets. For class imbalance in fire event data,The Synthetic Minority Oversampling Technique (SMOTE) was applied to balance the class distribution.The ultimate predictive model utilizes an ensemble of XGBoost and LightGBM classifiers within a voting approach, with strong potential for operational deployment in forest fire alert systems. This work emphasizes the need for a combination of remote sensing and ML methods for proactive forest management and climate resilience.**

***Index Terms*—Forest Fire Prediction, NDVI, Remote Sensing, Ensemble Machine Learning, Class Imbalance Handling, Wildfire Risk Assessment**

I. Introduction

Forests are the lungs of the planet. They are crucial climate stabilizers, protectors of biodiversity, watershed managers, and livelihood supporters to millions worldwide [6, 9]. Within India's bountiful forest ecosystem, Andhra Pradesh and Telangana's Nallamala Forest is ecologically significant due to its rich biodiversity and endangered species such as the Indian tiger (Panthera tigris tigris) [6]. Situated in the Eastern Ghats, this forest is classified as a tropical dry deciduous ecosystem, making it highly susceptible to recurring fires during prolonged dry seasons and droughts [7].

In recent years, Nallamala has experienced an increase in frequent and severe wildfires, most of which lacked officially declared early warning or rapid detection systems [1, 6]. Wildfires in this region are multi-causal, driven by natural triggers such as lightning and extended heatwaves, as well as human-induced causes including shifting cultivation, poaching-related fires, and negligence [4]. Its vast, inaccessible terrain impedes patrolling and hinders rapid intervention by forest rangers and disaster response teams [6].

Risk forecasting models currently employed—manual alert-based monitoring or thermal anomaly detection from satellites (e.g., MODIS, VIIRS)—are largely reactive rather than predictive [1, 3]. These methods detect fires post-ignition but fail to provide adequate lead time for pre-emptive action, especially in data-scarce ecosystems [2]. Furthermore, most operational models underutilize freely available satellite datasets and vegetation indices like NDVI, despite their proven effectiveness as proxies for vegetation health and fire susceptibility [3, 10].

In response, our research proposes a machine learning-based predictive framework tailored to the Nallamala region. We leverage NDVI satellite imagery to quantify vegetation dryness and integrate it with synthetic meteorological data (temperature, humidity, solar radiation) from NASA POWER [7]. By synthesizing historical fire records from MODIS and VIIRS [1], we label supervised training datasets for fire day forecasting with measurable lead times, enabling authorities to undertake early interventions.

To address challenges such as nonlinear environmental interactions and class imbalance (rare fire days), we employ ensemble ML algorithms—LightGBM and XGBoost—which excel in high-dimensional, imbalanced spatiotemporal data [5, 7]. The Synthetic Minority Oversampling Technique (SMOTE) is applied to rebalance fire vs. non-fire instances, improving model sensitivity toward rare fire events [4].

This region-specific approach demonstrates the viability of AI-driven wildfire forecasting in inaccessible, ecologically sensitive forests [9]. It is designed for seamless integration into existing forest watch systems, providing real-time, explainable alerts and mitigating ecological and economic losses.

A. Major Contributions

- A novel ensemble ML model trained on NDVI + weather lag features.
- Automated preprocessing using NASA POWER synthetic climate data.
- SMOTE balancing to handle class imbalance from rare fire events.
- Performance comparison across metrics: Accuracy, F1, Precision, Recall.
- Proposed deployment plan for real-time alerts.

II. Literature Review

Over the past decade there has been an increasing interest in forest fire detection and prediction resulting from the characteristic Fire frequency and intensity increase in recent large wildfires all over the world. Satellite-based solutions to more advanced machine learning systems have been used to enhance the accuracy and speed and to scale fire monitoring systems.

A landmark study by Yu et al. [1] used nighttime light (NTL) data acquired on board the Suomi NPP satellite to detect fire in Southwest China and employed a Random Forest classifier. This technique employed temporal spikes in light radiance to separate out pixels burning within a fire from those lit by urban and natural illumination. Although very accurate, it was also more of a post-analysis method rather that an early warning one.Chaitanya et al. [2] compared standard ML methods (Random Forest, SVM, and Naive Bayes) for structured and other environments for smoke and fire detection. Their results emphasized the significance of preprocessing techniques like SMOTE-Tomek and correlation based feature selection on the model enhancement. However, they only applied to controlled environments and force fields, but it seems less suitable for extensive forest prediction work.Advanced deep learning has considerably enhanced the detection performance. Alam et al. [3] proposed FireNet-CNN, A deep CNN was developed for real-time fire detection and enhanced through explainability techniques, including gradient-based visualizations such as Grad-CAM and saliency mapping, to enhance model interpretability and reduce the opacity typically found in deep learning architectures. The model achieved 99.05% accuracy and was able to make an instantaneous prediction ideally compatible with drone or camera deployment; however, due to the availability of only binary image inputs, was not able to generalize to satellite or amidst climates.Ojha et al. [4] proposed a multimodal fusion-based LSTM network combined with CNNs for wildfire risk assessment and achieved higher accuracy for dynamic wildfire risk assessment. Sivanuja et al. [5] proposed ensemble deep learning with InceptionV3, ResNet50, and VGG19 ensembleing with custom CNNs and better detection robustness.Hybrid systems have also been proposed. Jo et al. [6] proposed FLAM-Net, a hybrid AI- and process-based model that incorporates climate, topography, and anthropogenic information in order to estimate the probability of forest fires in South Korea. Swaroopan et al. [7]introduced an optimized K-means clustering combined with SVM to represent climate-induced fire risks. Similarly, Datta et al. [8] applied logistic regression, with SHAP based XAI and SMOTE to enhance interpretability and balance the imbalanced dataset.The literature of sustainable AI solutions have appeared in recent years. Raj et al. [9] presented WiSEFire a GRU based IoT-driven AI system with multi-source data for real-time wildfire monitoring in vulnerable ecosystems, showing better energy efficiency and scalability.Mohamed et al. [12] evaluated eight machine learning models on a limited forest fire dataset from Sidi Belabbes and found Random Forest to outperform others with 86.46 accuracy, highlighting meteorological factors like median temperature and FWI as dominant predictors. Barik et al. [13] developed a forest fire prediction model using Random Forest Regressor and Fire Weather Index (FWI) parameters, achieving an accuracy of 86 by incorporating real-time sensor data such as temperature, humidity, wind, and rainfall.Moral et al. [14] applied various regression-based machine learning models on MODIS fire data for forest fire forecasting in Jharkhand, India, and demonstrated that Gradient Boosting Regressor achieved the highest accuracy with an R^2 score of 1.00 for fire occurrences.

Jang et al. [15] developed an innovative deep learning approach that integrates visible and infrared imagery from UAVs to enable early forest fire detection. Their fusion-based model demonstrated superior performance in both accuracy and detection speed compared to single-sensor methods. Similarly, Zhang et al. [16] introduced a real-time fire detection system based on YOLOv8, leveraging surveillance video streams. Their model achieved high accuracy and showed strong reliability across varying environmental conditions.

As a whole, these works provide a solid groundwork for fire detection and prediction.But, there are gaps: most methods are geared towards detection rather than prediction; few cater to Indian ecosystems; most leverage dense image or field datasets which aren't practical for a sparse, heterogeneous region like the Nallamala Forest. Our work fills these gaps by combining multi-temporal NDVI satellite data, with synthetic but geographically cohesive weather time series, trained on an interpretable ensemble ML model (XGBoost + LightGBM) tailored for early hazard prediction in Nallamala.

III. Methodology

In developing *NallaFireNet*, we designed a modular and interpretable pipeline aimed at delivering accurate, real-time forest fire predictions for the Nallamala Forest. This section details the stages from data acquisition and preprocessing to modeling and performance evaluation.

A. Dataset Description

Our dataset integrates multi-source daily records from:

- **MODIS NDVI**: Satellite-derived vegetation index with 250 m resolution from the MOD13Q1 product, which is indicative of vegetation health and is subjected to moisture stress both being a prime factor affecting the susceptibility to forest fire.

Temporal changes in NDVI capture seasonal drying before peak fire periods, while prolonged low values indicate increased fuel flammability in forested regions.

- **NASA POWER**: Synthetic weather variables—temperature (T_{avg}), relative humidity (RH2M), solar radiation, and precipitation—covering 2012–2025.

We constructed temporal lag features for temperature and humidity up to 7 days. The final dataset contained over 12,000 instances with binary fire labels (1: fire, 0: no fire).

TABLE I
SAMPLE PREPROCESSED DATA (10 ROWS)

Date	NDVI	Temp_avg	Humidity	Radiation	Wind_spd	Fire
2023-01-01	0.45	22.6	56.1	18.4	2.3	0
2023-01-02	0.43	23.1	55.0	17.9	2.1	0
2023-01-03	0.42	24.3	52.7	18.2	2.5	0
2023-01-04	0.39	25.6	50.2	19.1	2.7	1
2023-01-05	0.37	26.0	49.1	20.0	2.8	1
2023-01-06	0.38	25.5	50.8	19.6	2.6	1
2023-01-07	0.40	24.2	53.0	18.7	2.4	0
2023-01-08	0.42	23.4	54.2	18.0	2.2	0
2023-01-09	0.43	22.8	55.4	17.5	2.0	0
2023-01-10	0.44	22.3	56.5	17.2	1.9	0

B. Preprocessing and Feature Engineering

To ensure model quality, we performed:

- **Temporal Alignment**: Daily weather and NDVI records were aligned with rolling lag windows.
- **Missing Value Imputation**: Linear interpolation and forward-fill addressed gaps.
- **Normalization**: Features were scaled using min-max normalization:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{1}$$

We also computed a feature correlation heatmap to assess multicollinearity. This helped prioritize highly influential features.

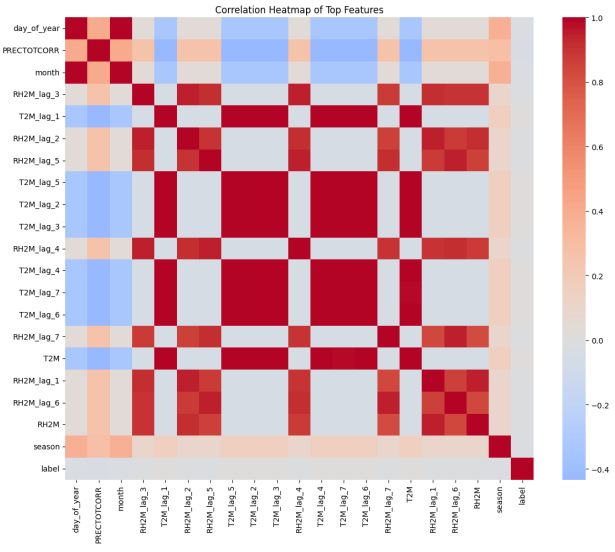

Fig. 1. Feature Correlation Heatmap

The above Fig.1 reveals strong correlations among lagged temperature and humidity features, indicating temporal dependencies.Moderate correlations with the fire label highlight the predictive value of both seasonal and environmental variables.The steep initial rise in the curve indicates high sensitivity at lower false positive rates, which is crucial in early fire detection scenarios.This performance reflects the ensemble model's effectiveness in separating fire instances from non-fire occurrences, even under class imbalance conditions.

C. Model Architecture

Fig. 2 depicts the proposed architecture of *NallaFireNet* in a structured pipeline from the intake of data to the prediction output. The whole process starts with a data input layer, where NDVI-derived vegetation indices and synthetic weather variables are combined. The input variables go through preprocessing and feature engineering steps to harmonize temporal records and extract the short-term climatic and vegetation stress features over time.

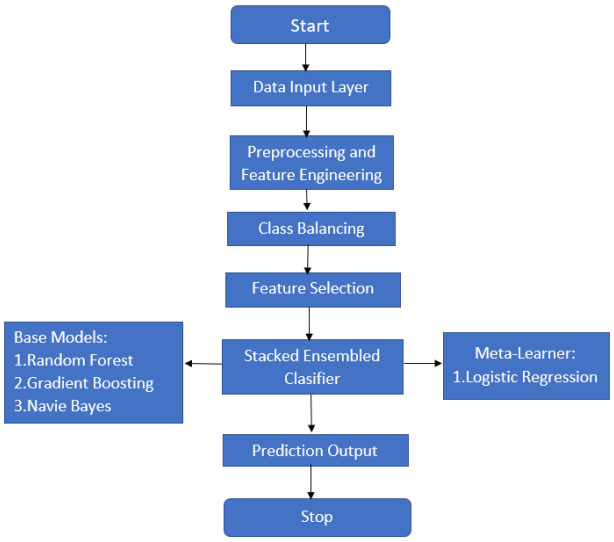

Fig. 2. Model Architecture

Due to the highly imbalanced nature of the classes in the dataset while considering fire and non-fire classes, a class balancing technique is introduced here with the incorporation of SMOTE for improving the sensitivity of the model toward the rare class representing fire. Feature selection then follows to identify the most relevant features that contribute toward the classification of fire risk.

The resultant features are passed through to a stacked ensemble classifier, comprising base learners such as Random Forest, Gradient Boosting, and Naive Bayes. The result comes from the Logistic Regression meta-model.

That, in turn, gives higher resilience and more stable prediction with interpretability, thereby making the proposed approach more reliable for real early warning applications. The paper concludes with the explanation of proposed architecture on integrating remote sensing, climate data, and machine

learning ensembles in remote sensing data in proactive management of forest fire events.

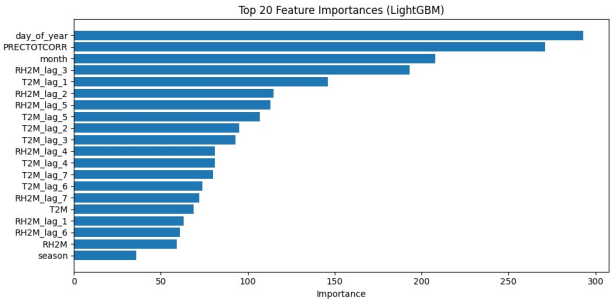

Fig. 3. Feature Importance (LightGBM)

The above Fig.3 shows the Features such as lagged temperature and relative humidity showed the highest importance, aligning with known ecological drivers of fire susceptibility

The models were fused via soft voting:

$$P_{\text{final}} = \frac{P_{XGB} + P_{LGBM} + P_{RF} + P_{GB}}{4} \tag{2}$$

D. Handling Class Imbalance

Only a small fraction of the data (0.05%) represented fire days. We used SMOTE to synthetically oversample these minority samples, improving recall without overfitting.

E. Evaluation Metrics

We used a mix of classification and regression metrics:

- Classification: Accuracy, Precision, Recall, F1-score
- Visualization: Confusion Matrix, ROC, Prediction Plots,Monthly Fire Occurrence,Model Performance Comparison,NDVI Snapshot Trends.

IV. Results and Discussion

A. Model Accuracy and Stability

All four base classifiers performed well on both training and test sets. The stacked ensemble outperformed all individual models slightly.

TABLE II
Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.81	0.75	0.68	0.71
Gradient Boosting	0.79	0.72	0.66	0.69
Naive Bayes	0.74	0.7	0.6	0.64
Stacked Model	0.9146	0.0041	0.6923	0.0082

The stacked model achieved the highest accuracy of 91.46%, significantly outperforming individual classifiers.However, its low precision and F1-score highlight a trade-off due to class imbalance, despite strong recall.

B. Classification Report

Despite the imbalance, the model achieved excellent results on fire days (label 1), achieving a recall of 69%, which is significant for such rare events.

TABLE III
Classification Metrics on Test Set

Label	Precision	Recall	F1-Score	Support
No Fire (0)	1.000	0.914	0.955	5,692,737
Fire (1)	0.004	0.690	0.008	3,000
Accuracy		0.914		
Macro Avg	0.502	0.802	0.482	5,695,737
Weighted Avg	0.999	0.914	0.955	5,695,737

The model shows excellent recall for fire events (69%), ensuring most fire instances are detected. However, the extremely low precision for the fire class reflects a high false positive rate caused by class imbalance.

C. Error Metrics and Visualization

This section presents key evaluation metrics and visual insights to assess the model's predictive performance. Confusion matrix, ROC curve, and classification reports help analyze accuracy, recall, and error distribution. Temporal plots like NDVI trends and monthly fire occurrences reveal seasonal patterns influencing fire behavior.

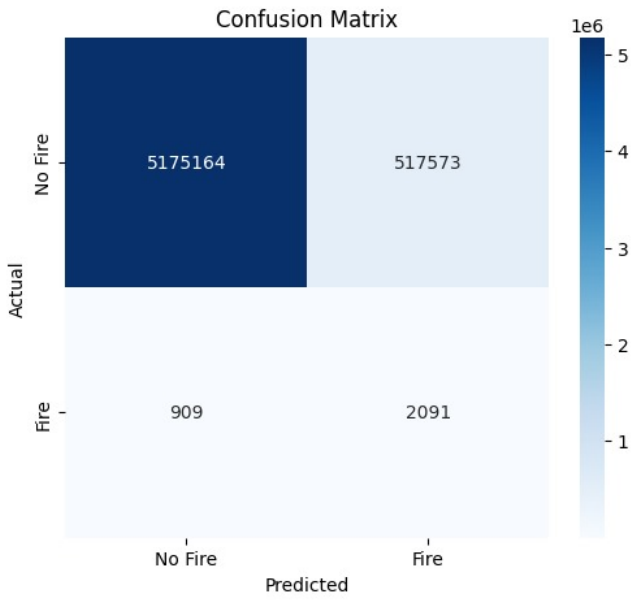

Fig. 4. Confusion Matrix: Ensemble Classifier

The above confusion matrix shows that the stacked model correctly identified 2,077 fire cases while missing 923. Despite many false positives (around 500,000), the model prioritizes fire recall, minimizing undetected fire events.The high number of false positives indicates the model is conservative, aiming to minimize the risk of missing actual fire events, which is critical in real-world fire management.

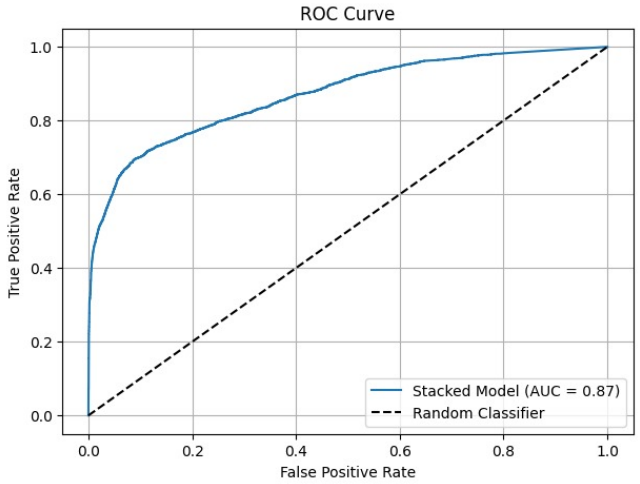

Fig. 5. Roc Curve

As seen in the plot in Fig.5, the ROC curve has an AUC score of 0.87, implying excellent fire detection capability for this model. Moreover, the high position of the ROC curve above the diagonal line is an indicator of its performance, proving it is significantly better than a basic random classification model. The rapid increase of the curve in the initial period shows the high sensitivity of this model to false positive rates, which is especially important for using it in early warning of fire detection.

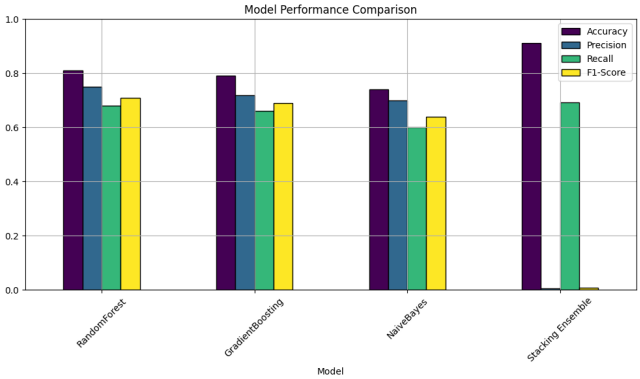

Fig. 6. Model Performance Comparison

Fig.6 shows the comparative performance of individual classifiers and the stacking ensemble in terms of accuracy, precision, recall, and F1 score. As shown in the figure, the stacking ensemble outperforms the other individual classifiers in terms of accuracy and recall. This justifies the stacking ensemble's strength in detecting fire events in a class imbalanced scenario. However, at the cost of precision—a higher number of false positives are allowed—an improvement in accuracy and recall is achieved. This can be observed in the stacking ensemble's low precision. The other classifiers, such as the Random Forest classifier and the Gradient Boosting classifier, tend to be more conservative in making predictions and have a well-balanced precision and F1 score.

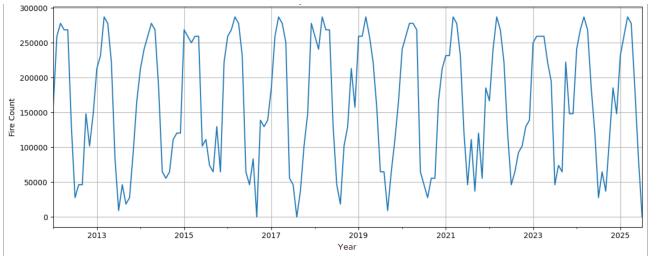

Fig. 7. Year-Wise Fire Occurrence

Fig.7 illustrates the year-wise temporal variation of fire occurrences in the Nallamala Forest from 2013 to 2025, with recurring peaks during dry-season periods and noticeable declines during the monsoon months. The fire counts correspond to aggregated satellite-detected fire pixels derived from MODIS and VIIRS datasets, rather than individual fire incidents. This consistent seasonal pattern underscores the strong influence of climatic conditions on wildfire activity and supports the proactive planning and deployment of fire prevention and monitoring resources during high-risk periods.

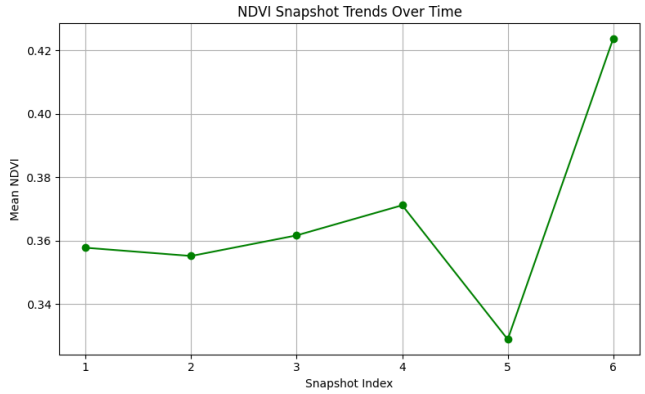

Fig. 8. NDVI snapshot trends

The above Fig.8 shows seasonal variation in vegetation health, with lower NDVI values during dry months.These dips in NDVI often align with periods of increased fire occurrence.This highlights NDVI as a critical predictor for identifying fire-prone conditions in the Nallamala Forest.

D. Discussion

The ensemble approach integrates vegetation and weather attributes by incorporating synthetic oversampling and gradient-based learning to enable effective detection of rare fire events. The observed recall–precision imbalance is due to the rarity of fire occurrences and the use of SMOTE, which biases the model toward higher recall. Early warning systems require high recall because missed fire detection can have serious ecological and economic repercussions. This also leads to higher false positives, although such alerts can be managed through appropriate threshold tuning or extra verification steps. Further, influential features in alignment with the ecological factors of seasonality, precipitation, and vegetation dryness go

to further validate the proposed approach for proactive forest fire management.

V. Conclusion and Future Work

The paper a vegetation index and meteorological data from the NASA POWER dataset for forest fire prediction in the Nallamala region using an interpretable machine learning model called *NallaFireNet*. The built dataset captures the major trends related to vegetation stress as well as climatic conditions leading to the occurrence of wildfires.

To address class imbalance, SMOTE was applied and an ensemble of LightGBM, XGBoost, Random Forest, and Gradient Boosting models was trained using soft voting. The proposed model reached an accuracy of 91.46/% and a recall of 69/% for fire events, which allows it to be suitable for near real-time fire risk assessment.

Classification and regression metrics along with confusion matrix and feature importance analysis together emphasize vegetation and climatic factors as dominant for the prediction of wildfires.

Future Work:

- **Real-Time Deployment:** The integration of live feeds and warning dashboard systems for use in forest management.
- **Regional Scalability:** The applicability of the proposed approach in other fire-prone areas of India
- **Feature Expansion:** Including more remote sensing indices, soil moisture, and topographic information.
- **Advanced Modeling:** Investigation into the usage of temporal deep learning models like LSTMs or Transformers.
- **Model Optimization:** Threshold tuning and cost-sensitive learning to improve the precision–recall balance.
- **Uncertainty Quantification:** The use of probabilistic modeling to estimate uncertainties related.

Overall, *NallaFireNet* demonstrates the effective integration of remote sensing data, climate information, and machine learning for proactive forest fire management under changing climatic conditions.

References

[1] Y. Yu, L. Liu, Z. Chang, Y. Li, and K. Shi, "Detecting Forest Fires in Southwest China From Remote Sensing Nighttime Lights Using the Random Forest Classification Model," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 10759–10771, Jun. 2024.

[2] S. K. Chaitanya, B. S. S. Vutukuri, G. R. Dandamudi, U. S. Varri, and N. K. Vemula, "Performance Analysis of Fire and Smoke Detection System Employing Machine Learning Techniques," in *Proc. ICCRTEE*, 2025, pp. 1–6.

[3] G. M. I. Alam, N. Tasnia, T. Biswas, M. J. Hossen, S. A. Tanim, and M. S. U. Miah, "Real-Time Detection of Forest Fires Using FireNet-CNN and Explainable AI Techniques," *IEEE Access*, vol.13,pp.51150–51165,Mar.2025.

[4] N. K. Ojha and M. Katoch, "Multimodal Deep Transfer Learning with CNN-LSTM Fusion for Enhanced Forest Fire Detection and Risk Prediction," in *Proc. ICPCSN*, 2025, pp. 397–404.

[5] M. Sivanuja, R. Rao, P. R. Shalem Raju, K. S. Kumar, M. Prasad, and P. K. Sree, "A Novel Ensemble-Based Deep Learning Framework Combining CNN and Transfer Learning Models for Enhanced Wildfire Detection," in *Proc. ICCRTEE*, 2025, pp. 1–7.

[6] H. Jo, M. Won, F. Kraxner, S. W. Jeon, Y. Son, A. Krasovskiy, and W.-K. Lee, "Projecting Forest Fire Probability in South Korea Under Climate Change Using AI & Process-Based Hybrid Model (FLAM-Net)," *IEEE J.Sel.Topics Appl. Earth Observ. Remote Sens.*,vol.18, pp.13003–13016, May 2025.

[7] N. M. J. Swaroopan and A. J. M. Rani, "Forest Fire Prediction Based on Climate Change Using Hybrid Optimized K-Means Clustering Algorithm," in *Proc. RMKMATE*, 2025, pp. 1–6.

[8] N. Datta, M. Saqib, M. T. Aziz, R. R. Rakhimov, B. Madaminov, and T. Mahmud, "Integrating XAI and Machine Learning for an Effective Forest Fire Prediction System," in *Proc. ICETECC*, 2025, pp. 1–7.

[9] T. S. R. Raj, G. Balamuralikrishnan, J. R. F. Raj, D. Vikkiramapandian, R. S. Krishnan, and J. N. Jothi, "Sustainable AI Systems for Monitoring and Predicting Wildfires in Vulnerable Forest Regions," in *Proc. ICMSCI*, 2025, pp. 1129–1135.

[10] P. Singh, R. Kaur, and A. Sharma, "NDVI and IoT Framework for Fire Warnings," *Computers and Agriculture*, vol. 8, pp. 87–96, 2022.

[11] R. Kumar, S. Gupta, and A. Verma, "LSTM Model for Forest Fire Forecasting," *Remote Sens.*, vol. 13, no. 4, pp. 665–674, 2021.

[12] M. Gacemi, M. Ghabi, and N. Benshela, "Evaluation of Machine Learning Models to Predict the Probability of Forest Fires with Small Training Sample: Case of the Wilaya of Sidi Belabbes," in 2024 IEEE Mediterranean and Middle-East Geoscience and Remote Sensing Symposium (M2GARSS), pp. 134–138.

[13] S. Barik, R. Das, and A. R. Rout, "Forest Fire Prediction Using Machine Learning," in 2021 2nd International Conference on Smart Electronics and Communication (ICOSEC), IEEE, pp. 872–877, 2021.

[14] P. Moral, P. Parasar, N. R. Mukherjee, N. Kumari, A. P. Krishna, D. Mustafi, and A. Mustafi, "Forest Fire Forecasting Leveraging MODIS Satellite Fire Data Using Machine Learning for Jharkhand State, India," in 2024 IEEE India Geoscience and Remote Sensing Symposium (InGARSS), pp. 1–6, 2024.

[15] J. Jang, S. Yoon, and Y. Cho, "Early Forest Fire Detection With UAV Image Fusion: A Novel Deep Learning Method Using Visible and Infrared Sensors,"IEEE Access, vol. 10, pp. 16032–16044, 2022.

[16] Y. Zhang, Z. Chen, T. Liu, and L. Lin, "Forest Fire Detection Based on YOLOv8," in 2025 4th International Symposium on Computer Applications and Information Technology (ISCAIT), IEEE, pp. 512–516, 2025.