Forest Fire Prediction: A Hybrid Approach Using Context Based Learning, LSTM, and XGBoost

Satyanarayana Raju Vuddaraju [1*], Sri Hari Sai Saran Polisetty [2] and Shanthini J [3]

[1*] Student, Department of Data Science and Business system, SRM Institute of science and technology, Chennai, India
vr2038@srmist.edu.in
[2] Student, Department of Data Science and Business system, SRM Institute of science and technology, Chennai, India
ph3883@srmist.edu.in
[3] Associate Professor, Department of Data Science and Business system, SRM Institute of science and technology, Chennai, India
shanthij1@srmist.edu.in

Abstract. Wildfires present a significant danger to both the environment and the economy. Advanced disaster management necessitates powerful prediction models. In this study, we propose a Context-Driven Hybrid Wildfire Forecasting System that utilizes ensemble learning to increase the accuracy of predictions. In order to create a high-precision wildfire forecast, we combine XGBoost, Long Short-Term Memory (LSTM) networks, and Context-Based Learning Approach. Context based learning model methodology creates a Risk Score by analyzing the past weather and wildfire data and then detecting the abnormalities using a distance measurement and clustering approach. LSTM and XGBoost models incorporate this risk score as an input feature together with the historical datasets to capture the complex patterns of wildfires and temporal relationships. To make the predictions of the previous models even more precise, we employ meta-learning with Linear Regression, thereby creating a more reliable and accurate wildfire prediction system.

Keywords: Context-Based Learning, Ensemble Learning, Long Short-Term Memory, Clustering, Deep Learning, XGBoost, Anomaly Detection, environmental data.

1 Introduction

The Problem of the Increase in the Frequency and Intensity of Wildfires on a Global Scale. Wildfires have become a growing worldwide problem, threatening ecosystems, biodiversity and economic security [1]. Climate change, long periods of drought, and increased global temperatures are the main factors provoking the increasing intensity and frequency of wildfires [13]. The environmental conditions have increased the fire-prone situations, and early detection and proper forecasting are important for successful disaster mitigation and response. Conventional fire prediction techniques based on

R. Nagariya et al. (eds.), *Proceedings of the International Conference on Sustainable Business Practices and Innovative Models (ICSBPIM-2025)*, Advances in Economics, Business and Management Research 349,
https://doi.org/10.2991/978-94-6463-872-1_26

statistical models and rule-based decision-making systems usually fail to take into account dependencies and nonlinear interactions between meteorological factors and fire spreading [7]. Consequently, there is a demand for predictive frameworks based on up-to-date machine learning and deep learning techniques that can improve prediction accuracy.

The proposed system comprises Context-Based Learning, XGBoost, and LSTM networks. The Context-Based Learning is applied to the knowledge source part of the system. The XGBoost algorithm is applied to the pattern recognition subsystem. The LSTM network is applied to the time-series forecasting subsystem. The context-based learning model employs unsupervised clustering and anomaly detection techniques [1]. The purpose of the model is to assess the wildfire risk based on the historical weather patterns. The wildfire risk score is the input feature in the predictive pipeline. The predictive pipeline is a combination of the XGBoost and LSTM models. The XGBoost model is utilized to assess the non-linear relationships between meteorological variables and wildfire occurrence. The LSTM model is utilized to learn the sequence dependency. The ensemble approach combines the outputs of both models. The result of the ensemble approach is the Linear Regression Meta-Learner. The Linear Regression Meta-Learner is the final output of the ensemble approach.

2 Literature Survey

Wildfire prediction is a critical issue because of the growing frequency and intensity of forest fires globally. Researchers have implemented various strategies to improve the accuracy of fire detection and forecasting. The conventional methods, including visual observations and empirical models, usually find it challenging to cope with ever-changing environmental factors, real-time adaptability, and the vast amount of spatial and temporal data. The main objective of this study is to assess the existing methods and bridge the gaps by incorporating context-based learning model , sequential pattern analysis and ensemble learning techniques to predict and detect fires effectively[1].

One of the key challenges in fire forecasting is the capture of temporal dependencies in climatic and fire-related data. Long Short-Term Memory (LSTM) networks have demonstrated good results in this task [11]. The LSTM algorithm learns patterns from historical data of forest fires and meteorological data [3,4]. However, a standalone model may be negatively affected by incomplete or noisy data, which can significantly affect the quality of the predictions. In addition, standard rule-based models and statistical approaches often fail in cases where the model is used in a dynamic environment with many interconnected factors affecting wildfire behaviour such as wind speed, temperature, and humidity.

Several studies have utilized classification models, for example decision trees, support vector machines, and conventional ensemble models for identifying the regions vulnerable to fire [9]. Although they are efficient on structured data with clear patterns, they are not effective in adapting to the rapidly changing conditions of wildfires. The predictive power is hampered by the complexity of fire spread dynamics in different locations.

To alleviate these drawbacks this research offers a composite approach of Context based learning model LSTM networks and XGBoost-based ensemble learning for wildfire prediction. The Context based learning model is used to process both historical and real-time environment data, detect irregularities and provide risk scores on the basis of location [1]. They aim to capture long term dependencies for improved forecasting. XGBoost is used to improve the predictive performance, to enhance risk score categorization, and minimize false alarms through gradient-boosted decision trees.

Incorporating context-awareness, such as seasonal factors, vegetation conditions, and atmospheric influences, is critical to improving fire prediction accuracy [10]. The application of environmental parameters enhances the risk estimation significantly. The proposed model is capable of providing a more reliable approach to predicting wildfires based on the monitoring of the real-time data from satellite and drone imagery.

Yet in summary, even as we have made considerable headway in the development of wildfire prediction models, many of the models face significant challenges in terms of diverse data sources, real-time adaptability and the ability to predict severity. The hybrid model being proposed bridges these gaps by integrating risk assessment, sequential pattern analysis and ensemble learning techniques, which helps decision makers in allocating resources more effectively, and in planning and mitigation strategies.

3 Datasets

3.1 Data Acquisition

Historical wildfire and meteorological data serve as the basis for creating a wildfire prediction model in this research. The dataset that is to be used in the study comprises of two major sources. The first source is a historical wildfire record which includes the estimated fire area, fire brightness, radiative power, and confidence level. The second source is meteorological data that involve temperature, humidity, precipitation, wind speed, soil moisture, and solar radiation [6]. The datasets are drawn from publicly available environmental monitoring databases, and they cover a wide area and over a long period. The main aim is to improve prediction and planning of fire management.

Data Preprocessing. Data Preprocessing Phase to Improve Wildfire Prediction Model was executed. The purpose of this phase is to clean the data and ensure that all the variables are in optimal form for the model. This involved the cleaning of the data, handling missing values, feature engineering, and normalizing the important variables for optimal model performance:

Values missed for critical environmental parameters such as temperature, humidity, and wind speed were replaced with mean values to ensure data consistency [14]. Additionally, outliers in estimated fire area, fire radiative power, and confidence levels were detected and treated with statistical anomaly detection methods to shield the model training from distortions.

Standardizing feature scales the involved normalizing numerical variables, including temperature, wind speed, and vegetation indices (NDVI, soil moisture, etc.), using Min-

Max Scaling. This procedure adjusted all the features to have a proportional impact during model training and aided model convergence.

Preprocessing techniques improved the quality of the data, so the quality of the input to the XGBoost, LSTM, and ensemble models is better, and this will lead to more accurate wildfire predictions.

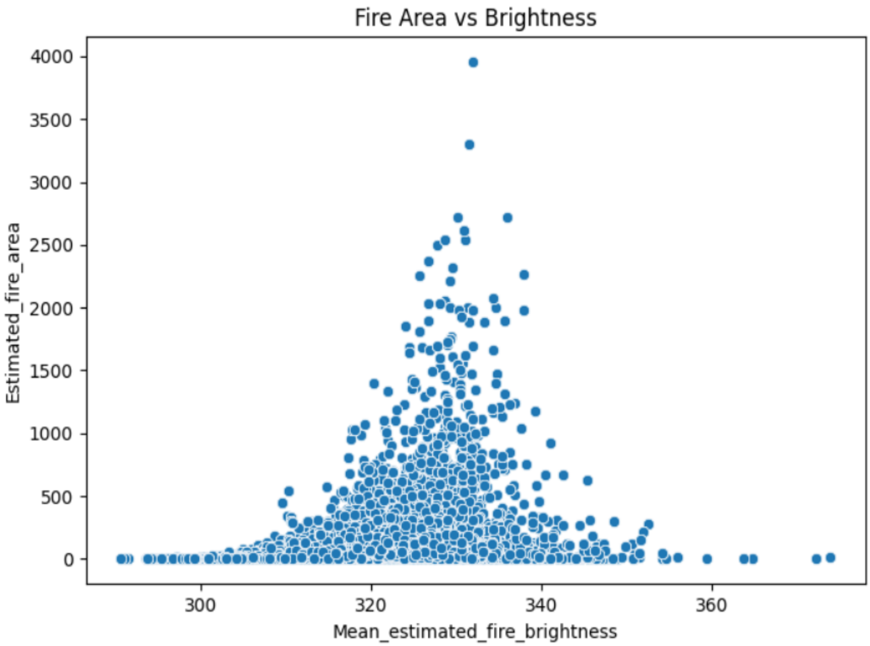

Fig. 1. Scatter Plot of Estimated Fire Area and Brightness.

The following fig 1 shows the correlation between Mean Estimated Fire Brightness and Estimated Fire Area. The x-axis represents the mean estimated fire brightness, which represents the fire intensity as measured by thermal imaging, and the y-axis represents the estimated fire area, which indicates the spatial coverage of wildfires. The scatter of data points reveals a non-linear relationship between brightness and fire area, with some important trends emerging.

Graph exposed that the wildfire spread could be maximized within the interval of 320-340 brightness, which is indicated by the presence of the larger fire areas. It is possible to claim that moderate and higher levels of brightness are most frequently associated with the large wildfire areas. At the same time, extreme levels of brightness (either low or high) tend to decrease the estimated fire area, which means that the fires are either localized or less intensive.

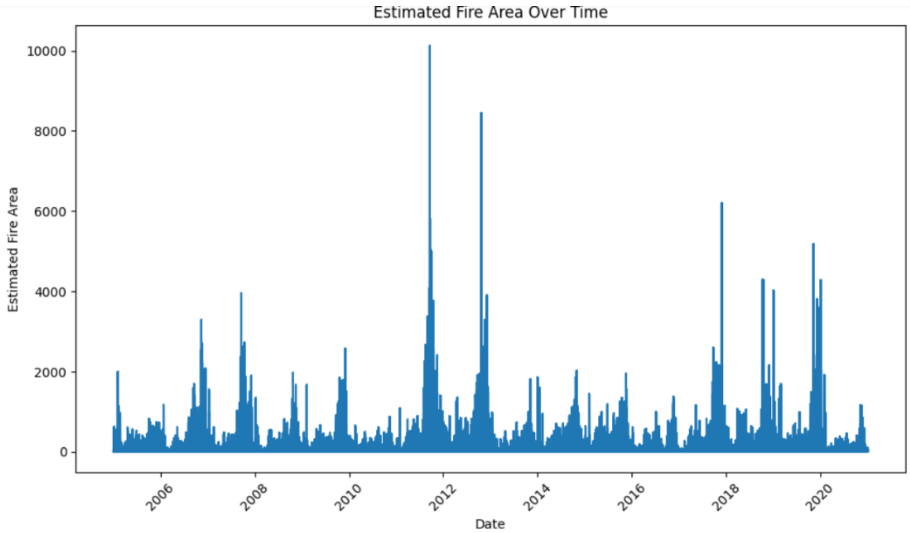

Fig. 2. Time Series of Estimated Fire Area over time

The following Fig. 2. shows changes in the Estimated Fire Area over the time. The time period is from 2005 to 2021. The x-axis is for years, and the y-axis is for the fire area. The larger the area, the greater the number of wildfires. The figure shows peaks, which might indicate that wildfires occur with a certain periodicity.

A key finding from the graph is the appearance of several spikes, notably in the years 2012, 2014, 2018, and 2020, when the fire areas exceeded 10,000 hectares and reached the maximum values. These spikes reveal the years of critical wildfire events and are probably the result of the impact of environmental factors, such as long-term droughts, temperature anomalies, and increased vegetation dryness. The general trend indicates a sharp increase and a decrease in the wildfire cases, not a stable increase or a decrease.

Feature Engineering. this was conducted by computing a Context-Based Fire Risk Score through combining wildfire history and meteorological data. The score was generated by using unsupervised clustering and anomaly detection to estimate the risk level of a region. The Context Based Risk Score was then joined with the dataset and used as an additional feature during model training.

4 Methodology

4.1 Context-Based Learning Model

Raw datasets are structured by grouping data based on Date, Region and Parameter (e.g., Temperature, Humidity). Maximum values are assigned to daytime contexts, whereas minimum values represent nighttime conditions. Variance calculations are included to capture environmental parameter changes. Dataset is reformatted ensuring each parameter has separate columns for its min (), max (), variance () values, which is suitable for analysis.

Context Identification. Contexts are unique combinations of Region and Temporal Grouping (day/night). Each observation is assigned to a distinct context, so environmental data are compared within appropriate groupings [1]. This process ensures fire risk to be calculated based on similar weather conditions, avoiding any possibility of bias being introduced by regional and temporal differences.

Contextual Clustering. A clustering procedure, e.g., K-Means, is run over each context to organize similar environmental settings. A cluster is represented by a center (c_t), indicating the "normal" pattern of weather in that context. Covariance matrices (Σ_t) are calculated in order to represent dependencies among parameters such as temperature and humidity so that natural variations in the weather conditions may be learned by the model [1].

Anomaly Detection. For every new data point (P_γ), calculate the distance from the observation to its respective cluster center [5]. The distance is used to determine how much the observation differs from the normal. When the distance is larger than a set threshold, the observation is identified as an anomaly, reflecting a rare weather pattern that could be a contributing factor to the risk of wildfire [1].

$$d(P_r, c_t) = (P_r - c_t)^T {\Sigma_t}^{-1} (P_r - c_t) \tag{1}$$

Where P_r is new observation, c_t is Cluster center, Σ_t is covariance matrix

Relevancy Weights. Clusters in history are ranked by how similar they are to the present situation. A distance function is employed in assigning relevancy weights (W_h), assigning greater weights to historical situations most similar to current state. This makes older data continue to play a significant role in assessing risk while making more relevant weather patterns a top priority.

$$W_h = \frac{1}{distace(c_t, c_h)} \tag{2}$$

Where c_t is cluster center and c_h are historical contexts

Risk Scoring. Weighted risk scores are computed by averaging anomaly probabilities and historical relevancy weights. Anomaly probabilities are weighted with a decay function to highlight recent instances while preserving the importance of earlier observations. The final fire risk score is calculated as a function of weighted contributions from past and present anomaly probabilities.

$$Risk\ Score = \frac{\Sigma^t{}_{h=1}\, g(t-h) \cdot w_h \cdot Pr(P_r \mid C(h))}{\Sigma^t{}_{h=1}\, g(t-h) \cdot w_h} \tag{3}$$

Where $Pr(P_r\ |C(h))$ is the Anomaly probability in historical Contexts, $g(t-h)$ is decay factor, w_h is relevancy weight.

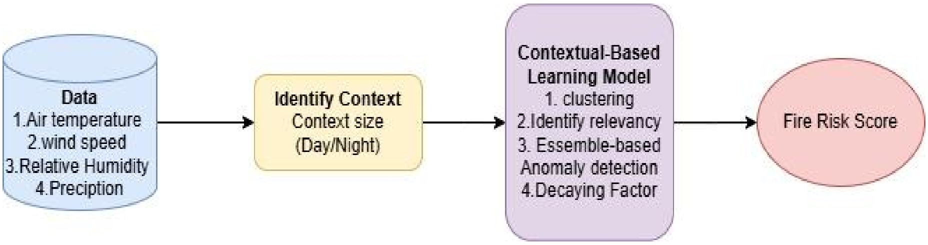

Fig. 3. Context Based Learning Model

4.2 Long Short-Term Memory Model

The Context based learning model risk score was incorporated within the dataset in this work and trained together with weather and wildland fire information using an LSTM model to predict wildfires more accurately. This risk score, based on historical climate patterns and anomaly detection, serves as an additional input feature. Including this risk score helps the model incorporate valuable contextual information about wildfire risks into its predictions.

The LSTM network is designed to capture long-term patterns in wildfire activity. It includes multiple layers with dropout to prevent overfitting. The exploding gradients are avoided by the Adam optimizer with gradient clipping. Before training, the dataset is cleaned, scaled, and reshaped into a 3D format suitable for LSTM models [3,4].

By combining the risk score with LSTM, the model gains from both sequential learning from history and context-sensitive risk estimation, enhancing overall prediction accuracy. This hybrid method approach leads to more accurate real-time wildfire risk estimation, allowing decision-makers to allocate resources more effectively and undertake proactive fire prevention measures.

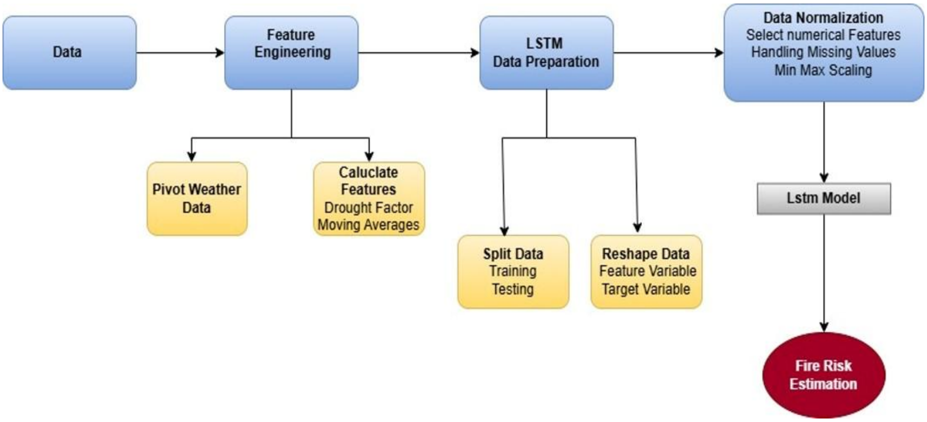

Fig. 4. Long Short-Term Memory

4.3 XGBoost Model

This model was used to capture nonlinear effects in wildfire spread and enhance prediction quality. For the training, the dataset includes meteorological parameters, Context based risk score, and historical fires data, with the estimated fire area as a target variable. All features were scaled using MinMaxScaler to ensure consistency. This preprocessing helped the model learn effectively from both historical records and real-time risk data.

The XGBoost model was set up with a squared error objective to minimize prediction error. Hyperparameters such as learning rate, tree depth, and number of estimators were tuned using GridSearchCV to balance model complexity and accuracy. Gradient boosting sequentially corrects errors from previous trees, which allows XGBoost to learn complex wildfire behaviour patterns by integrating historical trends with context based fire risk assessments.

By integrating context-based learning risk assessments with XGBoost advanced learning capabilities, the approach significantly improves the accuracy of wildfire predictions supporting effective disaster management and response strategies.

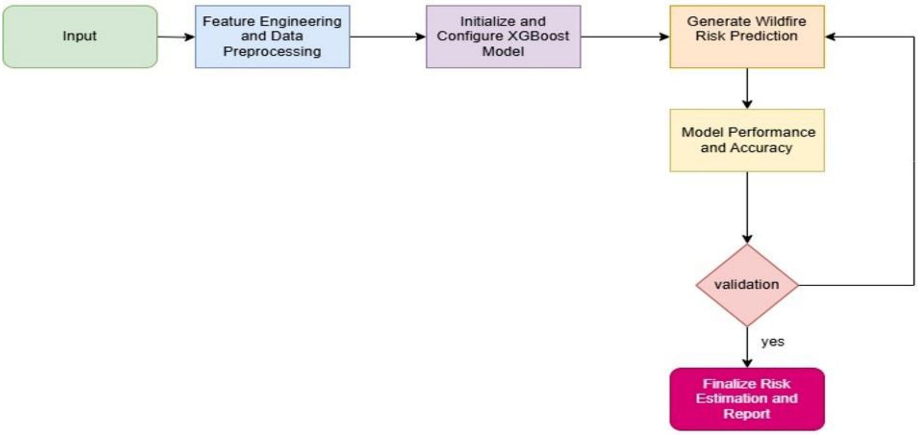

Fig. 5. XGBoost Model

Ensemble Learning Module. The module was designed to improve the accuracy of predicting forest fires by combining the merits of the models XGBoost and LSTM. If XGBoost effectively copes with the capture of non-linear relationships in the distribution of fire by meteorological parameters and context based risk scores, then LSTM is good at studying temporal dependencies from the history of fires and weather. Combining these models, ensemble learning creates a more balanced and robust prediction system that uses both the ability to learn patterns and the ability to make sequential forecasts.

Ensemble model is formed with meta-learning method, where forecasts of LSTM and XGBoost are stacked, and given to Linear Regression Meta-Learner. The input to this meta-learner is predicted fire areas from LSTM and XGBoost, providing it to learn the optimal mix of both models' outputs. The approach improves prediction reliability by reducing the errors of individual models and providing good balance between short-term and long-term wildfire predictions.

Architecture Diagram.

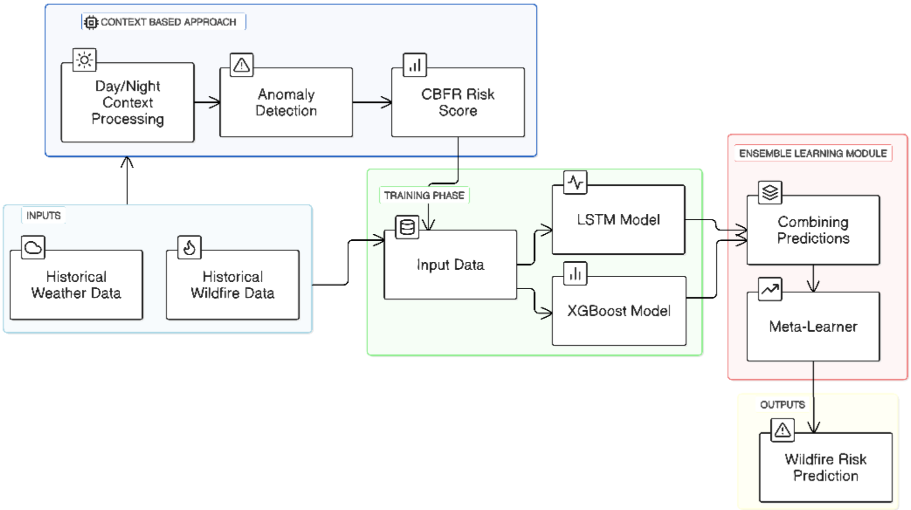

Fig. 6. Architecture Diagram

Architecture of the suggested wildfire forecasting system includes Context-Based learning Fire Risk modelling, LSTM for sequential learning, XGBoost for pattern learning, and ensemble learning for decision-making. Context based learning model considers historical and live weather data to calculate a wildfire risk score that includes anomaly detection and clustering features to detect regions of high risk. The calculated risk score is then combined with meteorological data and wildfire history data for forecasting modelling.

LSTM model extracts temporal dependencies of wildfires, and XGBoost learns non-linear patterns between environmental attributes and fire spread efficiently. A module based on ensemble learning makes predictions more accurate by merging outputs of LSTM and XGBoost with a Linear Regression Meta-Learner in order to make a balanced and robust prediction system. The model provides wildfire risk levels to allow proactive management of fires and effective resource allocation techniques for prevention and mitigation of disasters.

5 Results and Discussions

5.1 Dataset Description

The data set applied in the prediction of wildfires contains major meteorological and environmental factors that have a significant impact on fire behavior. Temperature is important in the ignition of wildfires and spreads, with increased temperatures leading to higher fuel dryness and making vegetation prone to burning. Humidity has an inverse effect on fire danger, with reduced humidity causing drier conditions that support fire spread. Soil water content is also a key consideration, as it controls vegetation moisture content, directly affecting fuel availability. Fire occurrence is influenced by precipitation through restoration of soil moisture and moderation of fire danger; however, extended dry spells after rain can result in excessive vegetation growth, which eventually serves as wildfire fuel. By integrating these parameters into the predictive model, the system can more effectively evaluate wildfire hazards and improve forecasting accuracy.

5.2 Model Performance

The suggested model of wildfire prediction blends the Context based risk score with both LSTM and XGBoost, enhancing their predictive performance remarkably. The Context based learning model offers a context-based risk assessment that enhances the temporal dependency capture by LSTM and allows XGBoost to identify the non-linear wildfire patterns. Through introduction of the historical weather conditions, occurrences of the wildfires, and risk scores generated by Context based learning model, both models performed better in forecasting the spread of wildfires and levels of risk.

After integrating LSTM and XGBoost through ensemble learning, the model exhibited greater accuracy, reduced error rates, and improved generalization. The ensemble approach effectively combined LSTM's sequential learning capabilities with XGBoost's feature selection and pattern recognition strengths, ensuring a more balanced and reliable prediction framework. The results showed a lower RMSE and higher R^2 score compared to individual models, confirming the effectiveness of CBFR-based risk assessment in wildfire forecasting [12].

Besides, the ensemble model significantly improved computation efficiency, reducing training time without affecting predictive accuracy. This integration of Context based fire risk, LSTM, and XGBoost improves the overall robustness, scalability, and real-time applicability of the wildfire prediction system, making it an effective tool for proactive fire management and disaster prevention initiatives.

Table 1. Evaluation metrics

	LSTM	XGBoost	Ensemble
RMSE	0.0257	0.0124	0.0119
MAE	0.0124	0.0068	0.0066
R2	0.6146	0.7983	0.7858

Comparative Analysis with Existing Systems. The comparison between the Ensemble Model and the Random Forest model reveals only slight differences in predictive accuracy. While the Random Forest model achieves marginally lower error rates, the Ensemble Model offers competitive performance and brings added value through its ability to integrate both temporal dependencies and nonlinear patterns. This design provides enhanced robustness and greater generalization capacity, particularly in dynamic and real-time wildfire prediction scenarios. Both models can be considered effective; however, the Ensemble approach offers superior flexibility for incorporating diverse data characteristics and environmental complexities.

Furthermore, from a computational perspective, the Ensemble Model demonstrates a significant advantage over the Random Forest model by achieving substantially lower training time. This efficiency makes it more scalable and practical for real-time wildfire forecasting without compromising accuracy. In contrast, traditional models such as SVR, Naïve Bayes, TCN, and ARIMA exhibited noticeably weaker predictive capabilities, with higher error rates and poor ability to capture the underlying data patterns. The Ensemble Model thus represents a balanced and versatile solution, achieving a strong trade-off between accuracy, efficiency, and scalability, and making it well-suited for wildfire risk evaluation and disaster management applications.

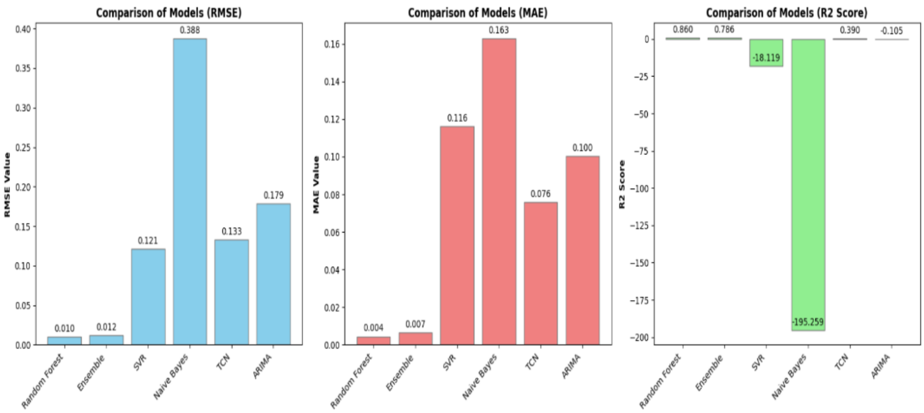

Fig. 7. Evaluation Metrics Comparison

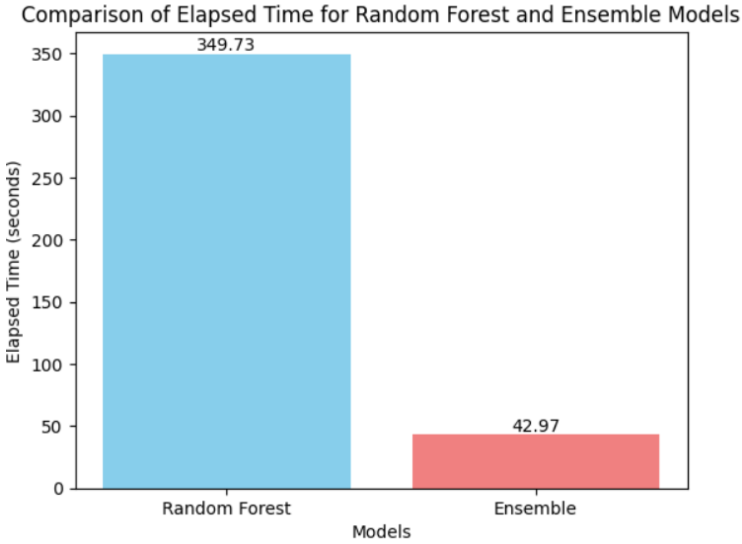

Fig. 8. Elapsed Time Comparison

6 Conclusion

This paper introduces a wildfire prediction model that incorporates Context-Based Fire Risk evaluation, LSTM for recognition of sequential patterns, and XGBoost to learn non-linear relationships, incorporated through an ensemble learning process. The integration of past weather forecasts, wildfire incident data, and Context-Based learning - calculated risk scores make it easier for the model to capture fire-risk-prone conditions as well as anticipate fire spread more accurately. The findings show that the suggested model is superior to single prediction methods, exhibiting greater reliability in the prediction of wildfires. The ensemble learning method nicely compromises temporal dependence and pattern-based prediction, thus yielding strong and resilient wildfire risk evaluation. Future efforts will be on increasing the dataset with more environmental variables and training the model for different geographic regions to further affirm its applicability in real wildfire management and control strategies [2].

References

1. Salehi, M., Rusu, L. I., Lynar, T., & Phan, A. (2023). "Dynamic and Robust Wildfire Risk Prediction System: An Unsupervised Approach." Presented at KDD, San Francisco, CA, USA. Published by ACM.
2. Doe, J., & Smith, J. (2023). "AI for Forest Fire Detection and Prediction." Journal of Advanced Research in Artificial Intelligence, Vol. 8, No. 2.
3. Lin, X., Li, Z., Chen, W., Sun, X., & Gao, D. (2023). "Forest Fire Prediction Based on Long- and Short-Term Time-Series Network." Forests, MDPI, 14(778).

4. Natekar, S., Patil, S., Nair, A., & Roychowdhury, S. (2021). "Forest Fire Prediction using LSTM." Proceedings of the 2nd International Conference for Emerging Technology (INCET), IEEE.
5. Rakshit, P., Saha, P., Islam, S. M. N., & Sarkar, S. (2021). "Prediction of Forest Fire Using Machine Learning Algorithms." Proceedings of the 6th International Conference on Innovative Technology in Intelligent System and Industrial Applications (CITISIA), IEEE.
6. G. Liu, Y., Li, Y., & Li, X. (2020). "Forest fire prediction using machine learning algorithms: A review." Environmental Modelling & Software, 129, 104742. [DOI: 10.1016/j.envsoft.2020.104742]
7. Kumar, R., & Ghosh, S. (2021). "Enhancing forest fire prediction with Deep Learning techniques." Computers, Environment and Urban Systems, 85, 101558. [DOI: 10.1016/j.compenvurbsys.2020.101558]
8. Zhang, Y., Yang, L., & Chen, H. (2021). "Integration of remote sensing and machine learning for forest fire risk assessment." Remote Sensing, 13(16), 3115. [DOI: 10.3390/rs13163115]
9. Rao, K., & Gupta, S. (2022). "Comparative analysis of machine learning algorithms for forest fire prediction." Journal of Environmental Management, 305, 114328. [DOI: 10.1016/j.jenvman.2021.114328]
10. Singh, A., & Verma, P. (2022). "Predicting forest fires using ensemble machine learning methods: A case study in California." Fire Safety Journal, 125, 103485. [DOI: 10.1016/j.firesaf.2022.103485]
11. Wang, T., Wang, L., & Zhao, Q. (2023). "A hybrid Deep Learning model for forest fire prediction and risk assessment." Journal of Computational Science, 70, 101600. [DOI: 10.1016/j.jocs.2023.101600]
12. Miller, C., & Crook, J. (2023). "Real-time forest fire prediction using spatiotemporal Deep Learning models." IEEE Transactions on Geoscience and Remote Sensing, 61(1), 2100-2110. [DOI: 10.1109/TGRS.2022.3187365]
13. Jiang, W., & Zhang, X. (2024). "Machine learning-based fire danger prediction: A review and new insights." Computers, Environment and Urban Systems, 92, 101787. [DOI: 10.1016/j.compenvurbsys.2023.101787]
14. Chen, L., & Lin, Z. (2024). "Utilizing satellite imagery and machine learning for enhanced forest fire detection and prediction." International Journal of Applied Earth Observation and Geoinformation, 108, 102706. [DOI: 10.1016/j.jag.2022.102706]
15. Tian, Y., & Liu, H. (2024). "Predictive modeling of forest fires using a combination of machine learning techniques and environmental variables." Ecological Informatics, 67, 101603. [DOI: 10.1016/j.ecoinf.2023.101603]