A Comparative Study of Predicting Wildfire Likelihood and Causes Using ML Models

Amishaben Natavarbhai Chaudhari
Department of Computer Science
Montclair State University
Montclair, New Jersey, USA
CWID: 50129394

Muhammad Kamal Subhani
Department of Computer Science
Montclair State University
Montclair, New Jersey, USA
CWID: 50136417

Punithan Thangavel Sathiyamoorthy
Department of Computer Science
Montclair State University
Montclair, New Jersey, USA
CWID: 50135877

Syed Mehrose Javed
Department of Computer Science
Montclair State University
Montclair, New Jersey, USA
CWID: 50128848

Venkata Appala Manoj Muvvala
Department of Computer Science
Montclair State University
Montclair, New Jersey, USA
CWID: 50122981

***Abstract*— This study sets out to predict both the occurrence and the prospective causes of wildfire events using state-of-the-art machine learning algorithms. Utilizing the extensive Fire Program Analysis Fire-Occurrence Database (FPA FOD) that contains in excess of 1.88 million geo-coded wildfire data from 1992 to 2015, we implement and benchmark two supervised learning models: XGBoost and Random Forest. These models incorporate temporal and spatial features such as fire magnitude, location, discovery date, and even the ownership of the land to predict wildfire occurrence and cause its classification into one of the 13 NWCG defined categories. The experimental result of this study indicates that both models perform well in predictive accuracy, although Random Forest outperformed XGBoost in overall classification accuracy as well as in managing class imbalance. The findings of this study demonstrate the crucial role of machine learning techniques in improving wildfire forecasting systems and reinforces the need for proactive wildfire prevention and resource allocation at federal, state, and local levels.**

***Keywords*— Wildfire Prediction; Machine Learning; Random Forest; XGBoost; Prophet**

I. Introduction

The primary factors contributing to the destruction of wildfires in the United States are climate change, droughts, and the increase of human activities in areas prone to wildfires. The level of ecosystems, infrastructure, and human life is at huge risk because of these fires. The traditional approaches used in wildfires predicting are often limited by the complex nature and scope of fire data, thus creating a necessity for better data solutions. This research is geared towards aiding that lack with the application of technology by utilizing machine learning methods for better predictive analysis for causal analysis associated with wildfires.

In specific, this research is focused on the prediction of wildfires XGBoost and Random Forest machine learning models are sought for use, assigning causative classification as a secondary concern. Moreover, the objective of the study is to evaluate the dual model in respect to modeling accuracy, generalization ability, and multi-faceted fire feature adaptability. The outcome thereof is intended to enhance decision making and optimization of resources towards wildfire management.

The dataset we are using in this research is the Fire Program Analysis Fire-Occurrence Database (FPA FOD) which contains records for over 1.88 million wildfires occurring from the year 1992 to 2015; the dataset encompasses relevant attributes including fire discovery date, the size of the fire, geographical coordinates, the reporting agency, the ownership type of land, and fire's statistical cause amongst others. Conducting feature selection to remove uninformative variables was done after dealing with the missing data and encoding the categorical variables.

Both XGBoost and Random Forest models were assessed based on accuracy and F1-score after being trained on data. Results indicated that, in comparison to Random Forest, XGBoost was better at predicting wildfire occurrences while Random Forest was better at classifying the causes of wildfires than XGBoost. Features that were the most important to the models included the time-based features (discovery date), geographical factors (latitude, longitude) and size of the fire. Moreover, Random Forest was more accurate than the other at managing class imbalance issues for the cause classification portion completed later.

The use of machine learning for the prediction and causal understanding of wildfires is promising, as demonstrated in this study. The ability to predict when fires will occur and their likely locations can significantly bolster prevention plans, emergency response planning, and even guide policy formulation. In summary, applying XGBoost and Random Forest as shown in this study could significantly reduce wildfire risks and support long-term environmental sustainability.

II. Related work

Wildfire prediction has been an active area of research in recent years, with a focus on utilizing machine learning algorithms such as XGBoost and Random Forests for more accurate forecasting. Several significant works have influenced the development of predictive models for wildfire analysis [1]. This study investigates wildfire susceptibility in Portugal using Random Forest and logistic regression to analyze both environmental and human-related variables. The authors emphasize the role of features such as vegetation cover, terrain slope, and proximity to urban infrastructure in fire risk modeling. The Random Forest model demonstrates superior accuracy and the ability to handle complex, non-linear relationships, making it a valuable tool in spatial fire prediction. The paper reinforces the effectiveness of ensemble models in building robust wildfire forecasting systems.

Other work [2] provides a comprehensive overview of machine learning models applied to forest fire prediction and detection. The authors examine various algorithms including Decision Trees, Random Forest, Support Vector Machines, and Deep Learning methods such as CNNs and LSTMs. The study highlights challenges such as data scarcity, imbalanced datasets, and real-time deployment limitations. This work supports the integration of both classical and deep learning models for effective early detection and management of wildfires.

Another research [3] highlighted the strong performance of XGBoost in terms of both classification accuracy and area prediction, showcasing its suitability for real-world wildfire forecasting tasks. In the study published in MDPI Forests (2024), researchers developed ensemble machine learning models, including XGBoost, Random Forest, and Multilayer Perceptron, to predict wildfire occurrences and burned areas. By integrating meteorological, topographical, fuel, and population data from global satellite sources, the XGBoost model achieved an AUC of 0.97 for fire occurrence prediction.

In their 2024 review [4], Xu et al. provide a comprehensive analysis of wildfire risk prediction methodologies, emphasizing the integration of various independent variables with machine learning techniques. They categorize these variables into four key aspects: climatic and meteorological conditions, socio-economic factors, terrain and hydrological features, and historical wildfire records. The authors discuss preprocessing strategies tailored to diverse data magnitudes, spatial-temporal resolutions, and formats, alongside methods for assessing variable collinearity and importance. The review highlights the application of statistical models, traditional machine learning algorithms, and deep learning approaches in wildfire risk prediction, with a particular focus on recent advancements in deep learning methods. Furthermore, the paper addresses current research limitations, advocating for the development of more effective deep learning time series forecasting algorithms, the utilization of three-dimensional data encompassing ground and trunk fuel, the extraction of more accurate historical fire point data, and the enhancement of model evaluation metrics.

Sobha and Latifi (2023) [5] present a comprehensive survey on machine learning models used for forest fire prediction and detection, evaluating various algorithms such as decision trees, support vector machines, and neural networks. The paper also addresses significant challenges faced in the domain, particularly the scarcity of data and issues with imbalanced datasets. Their work highlights the need for more robust models capable of handling these challenges effectively, offering valuable insights into the application of machine learning techniques in wildfire risk assessment.

III. Approach and implementation

As we are working on predicting the likelihood and the causes of wildfires, we have used predictive data mining techniques to tackle the problem. We have tried the following models/methods:

- Facebook Prophet
- XGBoost
- Random Forest

A. *Data Preprocessing*

We have standardized the date and location information, where the date conversion is done from Julian day format to the standard day format. Text values have been normalized for counties and states. Missing FIPS codes have been handled. Extra spaces have been removed for consistent data values.

Feature engineering is done to prepare the data for better performance accuracy. New feature like "PREV_MONTH_FIRE_COUNT" is created to represent the fire count from the previous month for the same location (state, county). This is helpful as fire occurrences might be influenced by recent fire activity. Rows with missing data have been removed as part of data cleaning. Rows with missing values in the PREV_MONTH_FIRE_COUNT column are removed using dropna to avoid issues during model training. Categorical features like (FIPS_CODE, STATE) have been converted into a numerical format suitable for ML models. A new column duration is created to store the fire's duration in minutes. It calculates the difference between containment and discovery datetimes and converts it to minutes.

One-hot encoding creates new binary columns for each category within a feature. The data is divided into training and testing sets using train_test_split. 80% of the data is used for training the ML models, and 20% is reserved for evaluating their performance. The random_state ensures consistent splitting for reproducibility. StandardScaler is applied to standardize the features by removing the mean and scaling to unit variance. The scaler is fit on the training data and then applied to both the training and testing sets.

B. *Model Selection/Training*

We implemented XGBoost and Random Forest algorithms for our research. We trained models using both algorithms after the required feature engineering and data cleaning. For the forecast of wildfires chances, XGBoost showed slightly better accuracy over Random Forest. Hence, we chose XGBoost to train the final model, hyperparameter tuning for better accuracy and predict the chances of wildfires.

For predicting the cause of wildfires, we tested multiple algorithms and recorded their accuracies, where XGBoost and Random Forest resulted in highest accuracy. The following classification models are implemented and the resultant accuracies are recorded:

- Random Forest : 0.62
- XGBoost Classifier : 0.60
- DecisionTreeClassifier : 0.50
- KNeighborsClassifier : 0.47
- Logistic Regression : 0.35
- Linear SVC : 0.25
- SGD Classifier : 0.17

The results show that the decision tree classifiers and simplistic models (like Logistic Regression, Linear SVC) are not as good as the ensemble methods such as Random Forest and XGBoost classifiers. Therefore, we used Random Forest for wildfire cause prediction, where we are able to achieve 83% accuracy after implementing SMOTE (Synthetic Minority Over-sampling Technique).

C. *Evaluation Metrics*

The We used different evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), R-squared (R^2) and F1 score which is calculated from the precision and recall values. MAE, MSE and R^2 help in identifying the accuracy for regression models which in our case we used for the XGBoost accuracy in case of predicting the likelihood of wildfire while F1 score helps in classification models which we used for evaluating the accuracy for wildfire cause prediction.

IV. EXPERIMENTS AND OBSERVATIONS

The models performed moderately accurate for the likelihood of the wildfire and fairly accurate for classification of the cause of the wildfire after performance tuning.

A. *Model Performance*

Initially, because of its robustness in time-series extrapolation, Facebook Prophet was implemented to forecast the likelihood of a wildfire as part of the exploratory analysis. However, it performed poorly with an R^2 of about -25, indicating a failure to capture meaningful trends. Afterwards, XGBoost and Random Forest models were implemented, both performing significantly better. XGBoost outperformed Random Forest by a small margin and was selected as the final model. After hyperparameter tuning using RandomizedSearchCV, the best XGBoost model achieved with MAE $\approx$ 3.37 and R^2 $\approx$ 0.48. The model is moderately effective but struggles due to the inherently random and non-seasonal nature of wildfire occurrences.

TABLE I. LIKELIHOOD MODEL PERFORMANCE

Model	MAE	R^2	Interpretability
Random Forest	~3.79	~0.35	Moderately Low
XGBoost(tuned)	~3.37	~0.48	Moderate

XGBoost was preferred for its slight edge in performance, though both models faced interpretability limitations due to their ensemble nature.

Likewise, multiple classification algorithms were evaluated to predict wildfire causes. The resultant accuracy of each of them is presented before:

TABLE II. CAUSE MODEL PERFORMANCE

Model	Accuracy
Random Forest	62%
XGBoost (clean data)	60%
XGBoost (with NaNs)	56%
Decision Tree	50%
KNN	47%
Logistic Regression	35%
Linear SVC	25%
SGD Classifier	17%

Random Forest emerged as the best-performing model. Labels like "Miscellaneous" and "Missing/Undefined" were dropped to reduce noise, which improved accuracy to 66%. Subsequently, SMOTE was used to balance class distribution, raising accuracy to 83%.

B. *User Interaction*

A command-line interface (CLI) tool was developed to enhance accessibility of the predictive system allowing users to input a county, year, and month to receive a wildfire risk prediction with a bootstrapped confidence interval. Key capabilities include:

- User input: County, Year, and Month
- Output: Wildfire risk score with 95% confidence interval
- Real-time query processing using pre-trained XGBoost model

Additionally, county-wise risk was also visualized on a U.S. map using choropleth layers via Folium.

C. *Model Strengths and Weaknesses*

- XGBoost demonstrated strong performance in regression tasks, particularly after tuning, but lacked transparency.
- Random Forest proved superior for classification, particularly after SMOTE application, but is prone to overfitting.
- Simpler models (e.g., Logistic Regression, SVC) underperformed due to the complex and non-linear nature of the wildfire dataset.

D. *Data Analysis*

The dataset used in this study has 1.88 million geo-coded wildfire data from 1992 to 2015. From this data we can visualize multiple relations among the data values.

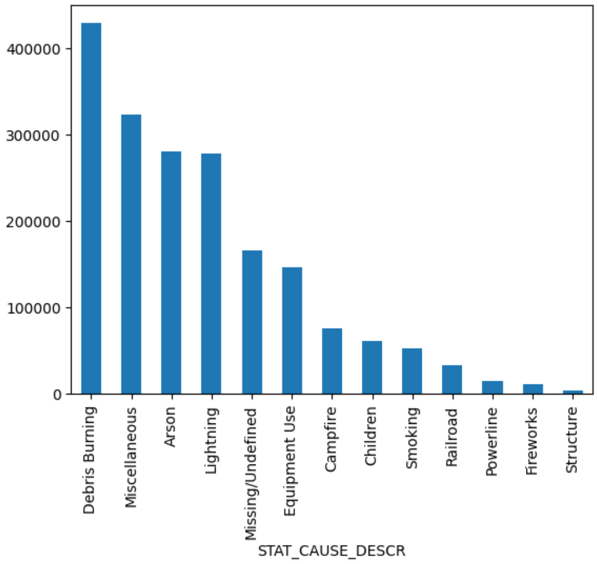

Fig 1. Bar Chart for Number of Fires by Cause

The chart shows the number of fires for each cause of fire reported in the dataset. This shows that top causes for the wildfires were debris burning, arson, lightning and many fire causes were reported as miscellaneous.

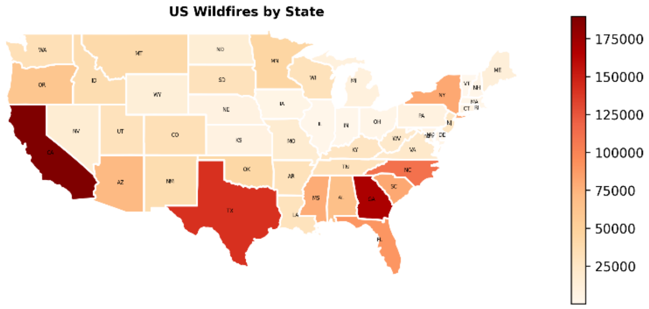

Fig 2. US Map with Wildfires by State

The above figure shows the US map plotted like a heatmap shows the states affected by the number of fires. This helps in visualizing top US states which had the most wildfires. For example California, Texas and Georgia are the top states with most wildfires. There could be multiple factors behind the large number of fires like climate, size of state, dense forests and vegetation, human activity and agricultural burns.

E. Challenges and Limitations

- The lack of strong seasonality or temporal regularity made time series approaches like Prophet unsuitable.
- Data imbalance significantly affected classification accuracy.
- Lack of many contributing environmental variables (e.g., humidity, wind speed) hindered the model accuracy.
- Fire incidents were sometimes aggregated at inconsistent geospatial resolutions (county vs. latitude/longitude), occasionally diluting model sensitivity.

V. Conclusion

This study demonstrated the feasibility and effectiveness of using machine learning for both wildfire likelihood forecasting and cause classification. XGBoost and Random Forest models were successfully applied to these tasks, achieving moderate (R2 $\approx$ 0.56) to high (accuracy $\approx$ 83%) performance with appropriate preprocessing and tuning.

The results highlight that machine learning can be a crucial and powerful tool for supporting wildfire mitigation and effective response planning. Accurate forecasting and cause classification can enable more timely responses, optimized resource allocation, improved risk communication, and comprehensive pre- and post-event analysis.

To carry this analysis further, more environmental variables such as wind speed, temperature and humidity can be added to the dataset for a more robust and accurate predictive system. Moreover, exploring advanced deep learning architectures such as temporal-spatial neural networks may lead to even more robust and scalable solutions.

In addition, there can be many potential applications of this study that can make a positive impact. Some of these are as follows:

- Warning systems for early discovery of fires in the zones which are at most high risk.
- Timely allocation of resources such as firefighting teams and equipment to fight with the fire.
- Policies can be developed on county/state level to influence regulations in the forest areas.
- Satellite imagery can also be incorporated in addition to the fire history data to monitor the areas.
- Reforestation can also be considered for the areas where massive damage has been done.

References

[1] A. Pereira, M. Santos, and L. Rodrigues, "Wildfire susceptibility mapping using machine learning techniques in Portugal," Environ. Rev., vol. 28, no. 4, pp. 402–417, 2020. [Online]. Available: https://doi.org/10.1139/er-2020-0019

[2] P. Sobha and S. Latifi, "A survey of the machine learning models for forest fire prediction and detection," Int. J. Commun. Netw. Syst. Sci., vol. 16, no. 7, pp. 131–150, 2023. [Online]. Available: https://doi.org/10.4236/ijcns.2023.167010

[3] S. Choi, J. Lee, Y. Kim, and H. Jang, "Predicting Forest Fire Area Growth Rate Using an Ensemble Algorithm," Forests, vol. 15, no. 2, pp. 321–335, 2024. [Online]. Available: https://doi.org/10.3390/f15020321

[4] Z. Xu, J. Li, S. Cheng, X. Rui, Y. Zhao, H. He, and L. Xu, "Wildfire Risk Prediction: A Review," arXiv, May 2024. [Online]. Available: https://arxiv.org/abs/2405.01607

[5] O. Bolukbasi, "Wildfire Data Analysis," GitHub, May 2024. [Online]. Available: https://github.com/omerbolukbasi/wildfire_data_analysis/blob/main/wildfire_data_analysis_omer_bolukbasi.ipynb

[6] E. Rodrigues, B. Zadrozny, and C. Watson, "Wildfire Risk Forecast: An Optimizable Fire Danger Index," arXiv preprint arXiv:2203.15558, Mar. 2022. [Online]. Available: https://arxiv.org/abs/2203.15558

[7] I. Zhang, "Project: Wildfire Risk and Information System," GitHub, 2021. [Online]. Available: https://github.com/irenezh1016/Project-Wildfire

[8] P. Ngwu, "Australian Wildfire Data Analysis and Visualization," GitHub, 2021. [Online]. Available: https://github.com/ngwuprince/Australian-wild-fire-data-analysis-and-visualization