Proceedings of the 12th World Congress on Electrical Engineering and Computer Systems and Sciences (EECSS 2026)
August 16 – 18, 2026 – London, United Kingdom
Paper No. ICBES 107
DOI: 10.11159/icbes26.nextgen.107

Machine Learning-Based Prediction of Wildfire Risk Using Environmental and Meteorological Data in California

Avery Li*, Emily Huang*, Olivia Lee*, Kaiwen Zhao*, Linda Shi
Institute of Engineering in Medicine, University of California San Diego
9500 Gilman Drive, La Jolla CA 92093, United States of America
averylifangzhi@gmail.com, ephsd1@gmail.com, olivia.p.a.lee@gmail.com, kaiwenzhao0129@gmail.com, zshi@ucsd.com
*High school students participating in the IEM OPALS program

Abstract - Wildfires are large-scale, uncontrolled events that cause significant ecological, economic, and public health impacts, including habitat destruction, air pollution, and soil degradation. In the United States, an average of over 60,000 wildfires occur annually, burning millions of acres of land, with a substantial proportion occurring in California. Human activities account for the majority of wildfire incidents, and the economic impact of wildfires in California alone is estimated to reach several billion dollars per year. Given the increasing frequency and severity of these events, there is a growing need for predictive tools that can identify high-risk conditions based on environmental and meteorological factors. In this study, a data-driven approach was used to investigate the relationship between environmental variables and wildfire occurrence. A dataset consisting of 128,125 records from California was analyzed, with each record labeled as either fire or no fire. The dataset includes 14 features, such as precipitation, vapor pressure, air temperature (maximum, minimum, and average), humidity (maximum, minimum, and average), dew point, wind speed, wind run, soil temperature, and ethylene oxide levels. A decision tree classifier was implemented in Python to model wildfire risk based on these variables. The model identifies hierarchical decision rules that partition the feature space and classify conditions associated with wildfire occurrence. Model performance was evaluated using standard classification metrics, with an overall accuracy of 96.2% on the test dataset. Analysis of the trained model indicates that variables such as soil temperature, air temperature, and humidity play key roles in determining wildfire risk. The decision tree structure provides interpretable rules that illustrate how combinations of environmental conditions contribute to fire likelihood. These results demonstrate the potential of machine learning methods to support wildfire risk prediction using environmental and meteorological data, providing a foundation for further development of data-driven risk assessment tools.

Keywords: wildfire prediction, machine learning, decision tree classifier, environmental data, meteorological factors, risk modeling

1. Introduction

Wildfires are an ongoing and persistent issue that impacts communities, economies, and ecosystems all around the world. Some fires are caused by factors like temperature, soil moisture, vegetation, and more. However, 80% of fires are caused by human activity, whether it's negligence or intentional. Victims of wildfires can potentially face not only respiratory illnesses but also eye issues, life-threatening diseases, loss of homes, and deaths of family and friends around them. Economies can also suffer due to government costs, property damage like cars, and insurance. Environmentally wise, wildfires are considered a source, meaning they release a large amount of air pollution. Plants absorb greenhouse gases, but because they cannot constantly keep up with sources, for example wildfires and livestock, carbon dioxide can travel to the Earth's atmosphere. The presence of CO2 inside Earth is one of the major causes of climate change, which is hurting animal habitats as well as raising temperatures leading to drastic weather changes [1].

Wildfires are unpredictable and problematic but are becoming more common. At any given time, communities, habitats, and wildlife surrounding them are at risk. In the past decade, the United State's annual fire death rate per million population has increased 27%. For example, in 2023, there were approximately 13 deaths per million people. These fire death rates also consist of children. Around 4.74 deaths per million people occurred in children under the age of 14 in 2023 and the fire death rate for children has barely decreased since 2014 [2]. Wildfires not only have an effect on people,

but also animals and wildlife. A lot of wildfires start in forests, where animals live. Plants and vegetation could be burned down, stripping many different species of their food source. These smaller species can face population depletion, and that causes a domino effect on the food chain because their predators have no prey to consume, therefore risking a population crash too. Also, ashes from the wildfire can contaminate the rivers and streams with ashes where animals drink from. Adding on, debris in the water can negatively affect aquatic animals such as fish or frogs. As wildfires burn the vegetation, they also burn the soil, lowering the chance that plants and trees will regrow. The after effect of wildfires on homes and forests have to be covered by insurance companies [3]. After the LA wildfires, insurance claims billed to 20.4 billion dollars, and this only partially covered 93% of insurance claims. The effects on economies as a result of wildfires include escalating costs of living, loss of jobs, and tighter government budgeting [4]. We built a machine learning model that learns from historical data to predict wildfires accurately, so deaths, environmental impacts, and economic consequences can be minimized.

We have researched wildfires, creating a machine-learning model that accurately predicts them. We searched what causes them and identified environmental and meteorological factors that can cause wildfires' occurrences. Our machine-learning model uses a decision tree to analyze these factors and find patterns in it to accurately predict whether or not a fire will occur. Because fires in California are so popular, and because of the recent fires in Los Angeles, we decided to focus on California as our main location to predict wildfires. Our model has been able to give us results with an accuracy of 96.2%. Our work and research will contribute towards addressing this problem.

2. Methodology

2.1. Study Area

The study area is the state of California. The data spans from January 1st of 2018 to September 18 of 2020, and data is collected from 153 different weather stations. We chose California because of its abundance of wildfire occurrences and abundance of data that we can use for predicting wildfires.

We have also studied many other models that predict wildfires and what methods they use. Over the course of the past decade, researchers have increasingly examined the application of artificial intelligence to predict wildfires. One study uses Big Data, remote sensing, satellite data, and machine learning algorithms such as Artificial Neural Networks (ANN) and Support Vector Machines (SVM) to analyze satellite imagery [5]. ANN is a computer system designed to recognize patterns in data like a human brain and make predictions based on those patterns by using artificial neurons [6]. SVM is a classification algorithm which uses binary classification which is backed by mathematics to ensure accuracy and prevent overfilling [7]. First, this study put a variety of data into one dataset that they got from the MODIS satellite which uses remote sensing. MODIS stands for Moderate Resolution Imaging Spectroradiometer, and it moderates under NASA and is an instrument used to monitor Earth's atmosphere and climate [8]. Then they input the dataset into machine learning models. The ones they used were ANN and SVM. They split the dataset of 804 total samples (386 fire and 418 non-fire events) into 70% training and 30% testing and validated their models by using classification metrics, cross-validation, and regularization. Classification metrics evaluate the performance using quantitative measures by analyzing accuracy (overall correctness), precision (positive predictive value), and recall (the percentage of actual fires correctly detected). Cross-validation includes splitting the dataset into multiple training and testing groups to see if the model still predicts well each time. Recall measures how many real fires the model successfully detected. They also compared their models to other wildfire prediction systems to see if theirs performed better. The results from this study demonstrated that the ANN model had an average accuracy of 89.91% and the SVM model had an accuracy 84.03% [5]. Even though wildfires can be unpredictable, recent tactics such as early detection systems and machine learning models are improving our luck with fatality rates and preventing our damage. And now, with our own machine learning model and research, taking inspiration from previous works,, we can assist in contributing to bettering those chances.

2.2. Data sources

The dataset we used is the *California Environmental Conditions Dataset* for *Fire Risk Analysis* from Kaggle. The environmental data collected for this dataset was from the California Irrigation Management Information System (CIMIS), which contains weather station observations across the state. There are a total of 153 weather stations, each consisting of approximately 14 numerical variables used for data analysis. Which are:, ETo (reference evapotranspiration, the rate at which water evaporates), precipitation, solar radiation, average vapor pressure, maximum

air temperature, minimum air temperature, average air temperature, maximum relative humidity, minimum relative humidity, average relative humidity, dew point, average wind speed, wind run, and average soil temperature. In addition, the dataset includes a target variable indicating whether a fire has occurred on a given date and location, allowing it to be used for supervised machine learning. Our dataset has a total of 128,125 samples, 5,354 resulting in a fire and 122,771 resulting in no fire. With real-world environmental data specific to California, this dataset is suitable for analyzing wildfire patterns and training predictive models [9].

2.3. Model Architecture

We used a machine learning model that we coded using Python. This machine learning model uses a decision tree classifier. Our model splits the data given into training and testing A decision tree classifier analyzes the data given and finds the patterns that determine the outcome. Then it creates a tree-like structure. When given a new set of data to predict whether or not there is going to be a wildfire, it starts at the Root Node and asks a main question about the data. Then based on the answer to the question it goes to the Internal Nodes and asks one of two questions based on the question. This process continues until reaching the terminal (leaf) node, the model then outputs the probability of whether or not a fire will occur or not. There are many advantages to using a decision tree classifier such as its understandable; however one big disadvantage is overfitting. Overfitting occurs when the tree was built such that it memorizes historical data using various idiosyncratic characteristics, instead of finding broader applicable patterns that generalizes; as a result this has a negative effect on the performance of the model when deploying the model for real use[10]. In the next section, we show how we use special train-test split procedure (by station rather than random sample) to detect and overcome this overfitting issue

3. Results and Discussions

3.1. Target Definition and Train-Test Split

When we first created the model with the decision tree having its unbounded depth, using random sample-based train-test split, the test accuracy was unusually high, reaching 98.9%. To build a robust and generalized model, we performed additional analysis and identified a column named “Stn ID”, which means the weather station’s ID. As mentioned before there are 153 unique weather stations recorded in the data. It goes up to weather station 262 because some weather stations are skipped which is why there are only 153 despite going up to 262. We split the training set as everything before the Stn ID 212 and the testing set as everything after or equal to Stn ID 212. This resulted in the model using 89,419 samples to train itself, and 38,706 samples for testing. The reason we chose this way to split the data is because the training set is 69.8% of the whole dataset and the testing set is 30.2%. This ratio is the closest we could get to the optimal 70% training, 30% testing split that another model used (we mentioned before). This was the optimal way to split the data and also a way to prevent the model from memorizing. This made our accuracy more plausible, being 93.2%.

3.2. Decision Tree Depth Adjustment

We then realized that because of overfitting, our accuracy was low when testing and could have been higher. Our original decision tree had a depth of 31 and had overfitting issues, memorizing station-specific ideosynchracy that fails to generalize to new stations. We performed experiments with different tree depth, ranging from 1 to 31, and identified the optimal depth was between 1-4. In other words, the simpler model (low depth) is more robust and generalizes well to unseen stations.

Table 1: Accuracy of model for decision tree depth of 1-31

Depth	Accuracy
1	0.9617480486
2	0.9617480486
3	0.9617480486
4	0.9617480486
5	0.960623385

6	0.9567406178
7	0.9535272932
8	0.9529247948
9	0.9538084591
10	0.9505148014
11	0.9481181968
12	0.9461634243
13	0.9455207594
14	0.9421735463
15	0.9408614388
16	0.9387460001
17	0.9376481142
18	0.9353318427
19	0.9351711765
20	0.9345419004
21	0.9335109588
22	0.933310126
23	0.9338055135
24	0.9315026309
25	0.9326272945
26	0.933310126
27	0.9331762375
28	0.932948627
29	0.9317971857
30	0.9311812985
31	0.9323327398

As you can see from the above table, the best accuracy is achieved when the depth of the decision tree is less than or equal to 4. At depth of 4, compared to depth of 31, it improved the accuracy, measured on the test station samples, from 93.2% to 96.2%.

3.3. Our Decision Tree and How it Works

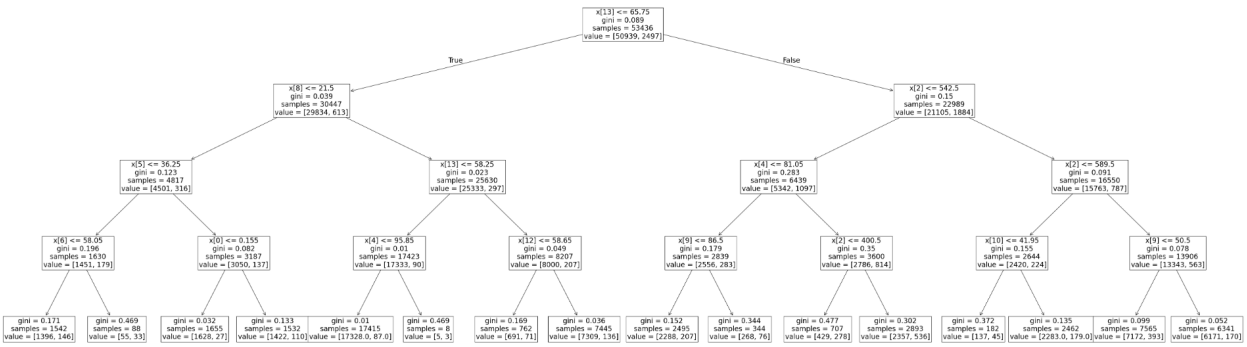

Fig. 1: Decision Tree with depth of 4

Figure 1 shows the decision tree the machine learned from the data. Our decision tree works by analyzing our data, progressively identifying from most predictive factors to least ones. We use Gini impurity criteria to search and identify splitting factor and splitting point (thresholds) to optimize the model performance. For example, the model identifies the most predictive factor is soil temperature: when splitting all samples into two groups based on if a station's observed soil temperature is higher or lower than 65.75 Fahrenheit. This split, among all possible factors and splitting values, maximizes the reduction of the gini impurity. It split the 53,436 training samples into the left group (29,834 samples with soil temperature equal to lower than 65.75) and the left group (22,989 samples with soil temperature higher than 65.75). The Gini impurity is reduced from 0.089 (2,497 positive, i.e., fire samples and 50,939 non-fire samples) into weighted total Gini 0.086, where left group Gini is reduced to 0.039 while right group increases to 0.15. This splitting procedure continues until depth 4, at which point, the model performance starts to degrade when evaluating on the testing stations, which indicates additional patterns identified from training station samples does not generalize to unseen testing stations.

We note that in the final depth-4 decision tree, totally 10 features participated: feature 13, 8, 2, 5, 4, 6, 0, 12, 9, 10, in canonical topological order. This shows the decision tree algorithm learned sophisticated 10 factor patterns. Furthermore, the left node size ranges from 8 to 17,415, indicating the algorithm adapted well to heterogeneous sized patterns, which is one of the hallmark and salient characteristics of decision tree approach.

4. Conclusion and Future Work

Wildfires are large, uncontrolled, and unplanned fires that burn in natural areas such as forests and grasslands. They result in habitat loss, air pollution from smoke, soil erosion, and destruction of habitats and infrastructure and many unfortunate deaths. This research helps prevent wildfires by accurately predicting their occurrence.

This study utilizes a Decision Tree Classifier and a data driven approach to predict wildfire risk in California using machine learning. What sets our python-based model apart from others is that we have 128,125 samples and 14 different considered factors. This shows that our model is diverse in many ways. With thousands of unique scenarios, our model can have specific examples and identify various rare situations that other machines may miss, making it more well rounded. Our machine is also taking into consideration several variables which reduces the amount of errors. We tested our model several times and consistently received accurate readings. Our highest output was 96.2%, demonstrating that our python-based model is trained to be precise. With so much to back up our accuracy with, our model is not only diverse and accurate, it is also reliable and robust as it is shown to generalize to different locations as tested on unseen stations.

Future work includes: expanding the model to cover other states and cities across the U.S. beyond California, experiment with different modeling architecture as we incorporate more samples, features, and richer data sources, to further improve the model performance, and lastly to perform more detailed analysis of the model and further improve model robustness by split and test our mode in out of time and out of area situations.

While the model demonstrates strong predictive performance, several limitations should be noted. The study relies on a single regional dataset and one classification method, and additional validation using independent datasets and alternative modeling approaches is needed. Future work will focus on incorporating real-time weather data, investigating temporal patterns, evaluating additional machine learning models, and improving model generalizability.

Overall, our achievements with a python based machine-learning model that has successfully predicted wildfires with an accuracy of 96.2% is a significant step in the right direction for wildfire prediction. We hope our research can guide others to improve and achieve even better results.

Acknowledgements

This study was conducted using material generously donated by Hyperway Pharmaceuticals. This work was supported by Beckman Laser Institute Inc. (LS and VGG). High school students participated in the UCSD IEM OPALS internship program. We gratefully acknowledge the support of Leighton Foundation, Ming- Wei Wu, Shu Chien, Lizhu Chen (CorDx Inc.), Xinhua Zheng, David and Leslie Lee, Mingwei He, and Wen Shi.

References

1. Center for Climate and Energy Solutions, "Wildfires and climate change," c2es.org. Accessed: Jun. 30, 2026. Available: https://www.c2es.org/content/wildfires-and-climate-change/
2. U.S. Fire Administration, "Fire death and injury risk," usfa.fema.gov. Accessed: Jun. 30, 2026. Available: https://www.usfa.fema.gov/statistics/deaths-injuries/
3. M. Snow, "How does wildfire impact wildlife and forests?," fws.gov. Accessed: Jun. 30 2026. Available: https://www.fws.gov/story/2022-10/how-does-wildfire-impact-wildlife-and-forests
4. Department of Economic Opportunity, "Analysis from DEO and LAEDC spotlights the economic toll of the January 2025 wildfires and outlines early recovery trends," opportunity.lacounty.gov. Accessed: Jun. 30, 2026. Available: https://opportunity.lacounty.gov/economictrendfire/
5. O. S. Younes, H. Mousannif, and H. A. Moatassime, "Predictive modeling of wildfires: A new dataset and machine learning approach," Fire Safty Journal, vol 104, pp. 130-146, Mar. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0379711218303941
6. GeeksforGeeks, "Artificial Neural Networks and its applications," geeksforgeeks.org. Accessed: Jun. 30, 2026. Available: https://www.geeksforgeeks.org/deep-learning/artificial-neural-networks-and-its-applications/
7. U. R. Archaya, Y. He, O. Kisi, and D. A. Wood, Natural Remedies for Pest, Disease and Weed Control, Academic Press, 2020. [Online]. Available: https://www.sciencedirect.com/topics/agricultural-and-biological-sciences/support-vector-machine.
8. NASA, "Fire Information for Resource Management System (FIRMS)," firms.modaps.eosdis.nasa.gov. Accessed: Jun. 30, 2026. Available: https://firms.modaps.eosdis.nasa.gov/map/
9. C. Zaloumis, "California Environmental Conditions Dataset," Kaggle. Accessed: Jun. 30, 2026. Available: https://www.kaggle.com/datasets/chelseazaloumis/cimis-dataset-with-fire-target
10. GeeksforGeeks, "Decision tree in machine learning," geeksforgeeks.org. Accessed: Jun. 30, 2026. Available: https://www.geeksforgeeks.org/machine-learning/decision-tree-introduction-example/