**Project Proposal:**

**Fire Risk Prediction**

CS 6140 – Machine Learning - FALL 2023

by Elio Khouri and Ilana-Mahmea Siegel

# 1. Introduction

## 1.1 Problem Statement

California is an epicenter for relentless wildfires that bring devastation year after year. Aspects such as climate change, urban expansion, and mismanagement of forests compound the issue, resulting in an escalation of the frequency and magnitude of these wildfires. These events not only jeopardize the safety of California’s residents but also threaten the state's economic health and the integrity of its ecosystems. Despite the evolution in firefighting methodologies and increased resources, there is a pressing requirement for a proactive strategy to predict and prevent these wildfires. A well-crafted predictive model stands to offer invaluable insights into potential ignition points and the trajectory of active fires, thus affording authorities the advantage of time in resource allocation and strategic planning.

## 1.2 Significance of the Problem

The ramifications of California's wildfires are profound, affecting diverse facets of life in the state. Based on recent data from Cal Fire, the economic burden of these fires is staggering, with damages amounting to over 300 million dollars in 2022 (Tyler, 2022), and a somber tally of 133 lives lost in the past five years alone (*Statistics | CAL FIRE*, n.d.). Yet, these figures only scratch the surface. The broader economic implications span diminished property values, stifled tourism, and long-lasting health repercussions.

A 2016 study suggests that by mid-century, average PM2.5 emissions from wildfires in the Western U.S. could rise by over 60% compared to levels from 2004 to 2009. Additionally, the peak emissions of PM2.5 might see a surge of nearly 400% in certain regions (*Living Under Smoky Skies—Understanding the Challenges Posed by Wildfire Smoke in California*, 2022). Such increases in PM2.5 emissions have direct health implications.

Pratt et al. estimated that due to exposure to ozone generated by wildfire emissions, there's an associated increase of over 2,000 ER visits in children with asthma in the USA annually (Holm et al., 2020). Delfino et al. observed that wildfire PM2.5 exposure led to an 8% rise in asthma visits for children younger than five (Holm et al., 2020). In British Columbia, Canada, a significant association was found between wildfire smoke and adverse health outcomes. For instance, a 30 µg/m3 increase in total PM10 led to increased respiratory physician visits (odds ratio (OR) = 1.05) and specific asthma visits (OR = 1.16) (Youssouf et al., 2014). Furthermore, in Sydney, bushfire smoke events corresponded with a 6% increase in respiratory hospital admissions on the same day, with chronic obstructive pulmonary disease admissions rising by 13% and asthma admissions by 12% (Youssouf et al., 2014). Delfino et al. also studied the 2003 southern California wildfires and found asthma admissions across all ages increased by 4.8% during the wildfire period, with a more pronounced effect in the elderly (10.1% increase) and children aged 0-4 years (8.3%) (Holm et al., 2020). Vulnerable populations, notably children and the elderly, face the brunt of these health adversities. Beyond the immediate health concerns, the environmental ramifications are severe. Wildfires expedite soil erosion, induce water contamination, and disrupt vast habitats. Tackling the wildfire challenge with predictive strategies isn't just a scientific endeavor but an imperative humanitarian and ecological initiative.

Predicting wildfires is crucial for California. With accurate forecasts, we can evacuate areas at risk, reduce casualties and property damage. This foresight also allows us to strategically position firefighters and resources, making response efforts more effective. Furthermore, knowing when air quality will decline helps at-risk groups, like the elderly or children, to take precautions. Overall, anticipating wildfires isn't just about preventing destruction—it's a vital step toward protecting our communities and preserving California's environment.

# 2. Background

## 2.1 Previous Solutions and Approaches

Wildfire prediction has been the subject of myriad studies, spanning from elementary Kaggle challenges to advanced peer-reviewed publications. The studies largely differ in terms of data sets used, complexity of the algorithms, and the depth of the analysis.

Elementary studies, often sourced from platforms like Kaggle, typically leverage limited datasets and pre-built machine learning models. For instance, many Kaggle competitions focus on predicting wildfires based on basic meteorological data such as temperature, wind speed, and humidity (see Appendix A). Some have incorporated additional features like vegetation, topography, and even previous fires in the region. While these elementary studies provide a foundational understanding of wildfire prediction, they often rely on single datasets and might not capture the nuanced interplay of multiple factors influencing wildfire occurrence.

On the other hand, advanced peer-reviewed publications dive deeper, leveraging an array of sophisticated machine learning algorithms, particularly deep learning models. These studies often scrape data from multiple sources, ensuring a comprehensive dataset. For instance, the study on "Wildfire Danger Prediction and Understanding With Deep Learning" leverages daily weather data from ERA-5 Land, including metrics like maximum temperature, wind speed, and dewpoint temperature. They also incorporate satellite variables from MODIS, such as the Normalized Difference Vegetation Index and day/night Land Surface Temperature. Additional features include soil moisture index from the European Drought Observatory, details on roads and waterway distances, population density from WorldPop, as well as elevation, slope, and land cover fractions from Copernicus databases (Kondylatos et al., 2022).

Another notable study, "Wildfire Risk Prediction and Detection using Machine Learning in San Diego, California," utilized a range of machine learning and deep learning algorithms, including SVM, XGBoost, and Random Forest for weather-based fire predictions, and CNN, LTSM, and SVM for remote sensing-based models. These models were combined into a sophisticated ensemble model that leveraged multiple data sources, such as weather data, satellite imagery from Landsat 8, historical fire events, and vegetation indices like NDVI. Their ensemble approach achieved a remarkable 100% accuracy for fire risk prediction, underscoring the potency of integrating diverse algorithms and datasets (Malik et al., 2022). Such refined models, trained on a multitude of datasets, tend to have a more holistic understanding of the factors influencing wildfire danger.

The government's approach to determining fire danger provides another perspective. The U.S. National Park Service and the National Weather Service, for instance, consider an array of factors such as fuel moisture, drought indices, and historical fire data to determine fire danger levels (NPS, 2023). This highlights the multifaceted nature of wildfire prediction, which is not merely reliant on current weather conditions but also on historical and environmental data.

While the field of wildfire prediction research often benefits from the collaborative efforts of larger teams and the mentorship of seasoned experts, Our project is a testament to the dedication and initiative of two individuals taking their initial steps into the vast realm of machine learning. Our goal is clear: to bridge the gap between elementary and advanced studies. By recognizing the limitations of Kaggle-based models and recognizing the strengths of refined academic studies, we aspire to broaden our data scope, inching closer to the insights and performance offered by more seasoned methodologies.

## 2.2 Drawbacks of Existing Solutions

While the existing models provide valuable insights into wildfire predictions, they aren't without limitations. Elementary studies from platforms like Kaggle often face data limitation constraints. Relying on single datasets can lead to overfitting, where models perform exceptionally well on training data but falter when exposed to new, unseen data. Moreover, these models often utilize pre-built classifiers, which might not be specifically tailored for wildfire prediction, leading to potential inaccuracies.

Furthermore, even though advanced studies use sophisticated algorithms and vast datasets, they sometimes overlook crucial factors. For instance, the importance of weather parameters like air temperature, wind, and soil moisture has been underlined in several studies (see Appendix A). While these parameters play pivotal roles in influencing wildfires, not all refined models incorporate them, potentially leading to gaps in predictions.

Moreover, while the government's fire danger assessment tools, such as the Wildfire Danger Factors and the Wildfire Fire Index (WFI) codes (appendix), provide a comprehensive understanding of fire danger, they sometimes lack real-time adaptability. These systems are based on accumulated knowledge and historical data, and while they are robust, they might not be agile enough to predict sudden, anomalous wildfire outbreaks.

Recognizing these challenges, our approach seeks to strike a balance. We understand that as a team of two, in our first machine learning course, our logistic regression model might not rival the advanced solutions developed by larger teams with extensive resources and seasoned expertise in the field. However, by integrating a diverse set of data sources, we aim to leverage the strengths of both elementary and advanced models. Our goal is to improve upon the limitations of elementary studies while acknowledging and learning from the sophisticated methodologies of advanced research. In doing so, we hope to offer a fresh perspective that bridges the gap between these two ends of the spectrum.

# 3. Proposed Solution

## 3.1 Addressing the Drawbacks

Most of the papers and projects that we examined use the day’s precipitation, humidity, temperature and wind measurements as input variables. Many also consider the amount, type and dryness of fuel either by including fuel buildup estimates and fuel moisture codes (e.g. Zaidi, 2023), or by using vegetation data and longer-term measurements of humidity, precipitation and temperature (e.g. Kondylatos, 2022). We have not found studies that explicitly compare models trained on the inclusion of this longer-term data vs. models that are trained using only the given day’s weather data. Rainfall and water from humid air can be absorbed by living and dead vegetation, wooden fences, etc., while dry air leaches moisture from its surroundings (NPS, *Fire Danger*). High temperatures cause faster evaporation. All of these factors, as represented by medium- and long-term measurement features in our dataset, can indicate the dryness and flammability of fuel in a given area. We intend to demonstrate that the inclusion of longer-term weather data improves model performance.

As critical as weather, vegetation and land use data are to determining fire risk, they do not encompass all variables that impact the likelihood of fires. Fire danger is complex and is likely influenced by many confounding variables, which cannot be directly measured. For example, many forest fires are caused by campfires[[1]](#footnote-0), which occur more frequently in areas with heavy camping use. Camping activity on its own may be difficult to measure, and of course doing so would not begin to address the myriad of other unknown variables that contribute to fires. We hypothesize that adding another set of input features representing a region’s fire history may be used as a proxy to account for the influence of some of these confounding variables on the risk of fire in a region, whether that’s due to camping activity, local government policies, or other factors.

While a region that has frequent fires is likely to have more, however an area that has recently been burned is unlikely to burn again soon since there isn’t as much fuel present. Including fire history could also be an important addition to the use of the fuel moisture codes, or vegetation and drought data. Even if the long- and short-term weather factors indicate that a location is experiencing fire weather, that location is unlikely to burn if it has recently done so and all that’s left is charred. Only one paper we found included regional fire history in the set of input features (Shmuel, 2023). In their experimentation, including fire history data improved model performance. We will conduct an additional experiment to test whether and how much the inclusion of the fire history as a factor improves model performance.

## 3.2 Model/Framework/Approach Description

### 3.2.1 Features and Factors Contributing to Fire Risk

As previously discussed, there are many factors, known and unknown, that contribute to fire risk. These include the amount and dryness of fuel that make an area ripe for burning, the possible ignition causes like lightning or campfires, and more. Ideally, we would like to incorporate many of these factors as input features, however we have not been able to find sources for many of them (see section 4 for a discussion of the data). As it stands, we are able to extract data of varying quality (again, see section 4) representing the location (latitude and longitude) and time (date and time) for which to predict fire risk, as well as nearby measurements of elevation, temperature, humidity, precipitation, wind direction and wind speed. We will use this data to construct other features representing medium- and long-term averages of the nearby temperature, humidity and precipitation. We also hope to construct fire history features for each location that include the number of acres burned in the wider region and the last time the area around the point in question was burned. Section 4 discusses the data sources in more detail.

### 3.2.2 Exploratory Data Analysis

Once we have the proposed features listed above, we will perform exploratory data analysis to observe their relationships to each other and to fire danger, and to determine which features and model type to use.

#### 3.2.2.1 Feature Selection

We will first perform univariate analysis on each independent variable to determine whether there is a correlation between it and the dependent variable. It is possible that some of our proposed features do not correlate with fire danger, either because of the nature of fire danger or because of the quality of our data. Regardless of the reason, we will remove independent variables that do not strongly correlate with the dependent variable.

Next, we will perform multivariate analysis to determine the covariance of independent variables. If two or more features are strongly correlated with each other, we will select only one of them to be an input to the model. We will choose whichever of the covarying features correlates most strongly with the dependent variable.

During the exploratory data analysis, we will also experiment with different ways of representing input features. It is possible that some representations of the features are more useful in determining the dependent variable than others. Outliers or trends that we observe during this exploratory data analysis may also guide our continued search of other data sources.

#### 3.2.2.2 Model Selection

Our observations of the nature of the relationships between the independent variables and the dependent variable will determine which model type we select. If the relationships between the input features and fire danger are linear, which we do not expect to be the case, then we will use a logistic regression model. If, as we expect, many of the relationships are polynomial, we will choose between a Support Vector Machine and a Random Forest approach. Depending upon the strengths and weaknesses of SVMs and Random Forests (neither of which we have yet learned about), we will select whichever model seems best suited to our data.

### 3.2.3 Baseline and Experimental Models

As mentioned previously, we intend to conduct two experiments exploring our hypotheses that the addition of long-term weather data and of fire history data both improve model performance. Since our experiments are based entirely on changes to the input features, we will use the same model type and training methodology for the baseline and both experiments.

The baseline model will be of whichever type we select based on the observations made during the exploratory data analysis step. The exact input features may also change after exploratory data analysis, but at the moment, we expect them to be:

* latitude,
* longitude,
* month,
* day,
* hour,
* the nearest weather station’s elevation,
* and the nearest weather stations’ daily measurements of:
  + precipitation,
  + humidity,
  + dew point temperature,
  + dry bulb temperature,
  + wind direction, and
  + wind speed.

The first experiment will add features to the input features used for the baseline. These additional features are also not finalized, but they are expected to be the one-week, two-week, one-month and/or six-month averages of the nearest temperature and wetness measurements:

* average precipitation,
* average humidity,
* average dew point temperature,
* and average dry bulb temperature.

The second experiment will also add features to the baseline’s input features. These are expected to be:

* the area burned within 0.5-1 degrees of latitude and longitude within the last ten to twenty years,
* the area burned within 0.1-0.25 degrees of latitude and longitude within the last one to three years,
* and the days since the last fire to occur within 0.1-0.25 degrees of latitude and longitude.

This is similar to the process Shmuel and Heifetz used for their history features (Shmuel, 2023). The exact distances and time ranges will be determined during exploratory data analysis.

# 4. Data

## 4.1 Dataset Description and Types

### 4.1.1 Data Sources

We are not using a pre-made dataset. Rather, we are compiling a dataset from several different resources, each of which provides features of interest. Our primary data resources will be:

1. *Incident Data* from Cal Fire,
2. *Global Historical Climatology Network - Daily (GHCN-Daily), Version 3* from the National Oceanic and Atmospheric Administration (NOAA),
3. *U.S. Local Climatological Data (LCD)* from NOAA, and
4. *Spatial wildfire occurrence data for the United States, 1992-2015* by Karen Short.

At the time of writing, the Cal Fire *Incident Data* (dataset 1) includes 2179 entries for fire incidents that occurred from 2013 to the present. It also includes two incidents from 1969, which we will drop. From this dataset, we will extract the date and approximate time (ISO date/time format) that each fire began, as well its latitude (float) and longitude (float) and the acres that were burned (int). We may also use other features from this dataset, such as the ones describing each incident’s administrative unit and county (both strings), since it is possible that variables related to government and administration also contribute to the likelihood of an incident occurring.

The *GHCN-Daily* dataset contains daily (string: “yyyymmdd”) weather data from 1,377 stations within California. We will primarily be considering data from stations within the state, but may also include data from stations in neighboring states that are near the California border. Data for each station includes the station’s latitude, longitude and elevation (all floats). We will use the latitude and longitude to map this data to nearby fires in the Cal Fire *Incident Data* and will use the elevation as another feature for our model. Each station also provides daily precipitation and snow measurements (ints) dating back a varying number of years. If we cannot find a critical number of stations with data dating back to 2013, then we may filter out older fire data so we can train the model with a full set of weather features.

The *LCD* dataset contains hourly (ISO date/time strings) observations from 140 observation posts located in 48 of California’s 58 counties. These observations include measurements of dew point temperature, dry bulb temperature, humidity, wind direction and wind speed (all ints). We have not yet determined how to incorporate this data into our training dataset, given that it is more geographically sparse than the other data we will be using. However, temperature, humidity and wind are important factors in fire prediction (Gutierrez, 2021)(National Park Service, *Understanding Fire Danger*), so we will experiment with how best to include this data and continue to look for other datasets that include more locations.

As with the *GHCN-Daily* dataset, we will use distance to map the *LCD* data to the Cal Fire *Incident Data*. This will result in fire/no-fire examples with weather measurements from varying distances, which upon first look is not ideal. However, such variation in fire location to weather measurement distance reflects the reality of the world in which the model needs to function. Even so, we will set a threshold for determining how far away is too far to provide useful data. If an example event cannot be matched with near enough weather data, then the example will not be included. If disallowing examples in this way biases the examples in the training and test sets away from the normal geographic distribution in the state, then we may have to leave out this dataset all together.

The *Spatial wildfire occurrence data for the United States, 1992-2015* dataset contains, as the title suggests, wildfire data from across the United States between 1992 and 20015. Crucially, it includes features for the state, latitude, longitude, discovery date, discovery time, and acres burned for each fire. We will use this dataset to construct fire history features for our training and test examples, which will all fall between 2013 and 2023.

In addition to the four datasets listed above, we are actively seeking out data on vegetation and groundcover, elevation and topography, population density and land use, and lightning storms. If we are able to find them, these datasets will likely be stored as GeoJSONs or Shapefiles, with regions represented by a series of latitude/longitude coordinates. We will map each fire-risk/no-fire-risk example to whichever region contains its latitude and longitude.

### 4.1.2 Balancing Positive and Negative Class Examples

So far, this section has only discussed the considerations for constructing a set of positive (fire-risk) class examples by combining the Cal Fire *Incident Data* with the other datasets. Of course, the Cal Fire *Incident Data* does not track examples of the negative (no-fire-risk); no dataset would. However, negative examples are crucial for constructing a training dataset, so we will construct one.

We will construct the negative examples by selecting location and time features like those we would extract from the Cal Fire *Incident Data* dataset and matching them with data from the other datasets in the same way as will be done for the positive examples. Locations and times for the negative examples will be selected in two ways: (1) entirely randomly (random locations and random days/times) and (2) from locations that appear in the positive dataset, with random days in the month preceding the positive fire and from random days throughout the entire dataset time window. We will balance the resulting data to ensure locations are spread around the state and times are spread through our training time window (2013-2023). In this way, we hope to provide many negative examples to the model so it can learn to distinguish both locations that are more prone to fire and other contributing factors that make fire more or less likely in those locations at any given time, without being biased towards or away from specific locations by the training data.

While the most accurate representation of the real world in terms of fire/no-fire ratio would be to include examples from every location across the state for every day (possibly even every hour or minute) since 2013, doing so would create too large a dataset for us to work with. Such an approach would also overwhelm the model’s ability to pick out the factors that contribute to fires since the ratio of positive to negative examples would be so imbalanced. A classification model trained on such an imbalanced dataset would score high on both precision and recall by always predicting no-fire. Additionally, fire danger (which we are attempting to predict) is much more prevalent than actual fire occurrences (the positive examples that we have), but we do not know the actual ratio of high fire danger locations/times to low fire danger locations/times. Given all that, we think it would be ideal to construct a dataset with an even ratio of positive to negative examples. However, since the locations and times of the positive examples are not evenly distributed, and since we want to intentionally add negative examples in some similar locations as the positive examples, including only as many negative examples as positive ones would not allow for as widely distributed locations and times as is necessary to reduce bias in the location and time features. Instead, we will create twice as many negative examples as positive ones. One quarter of the negative examples will have approximately the same locations as positive examples and three quarters will have entirely random locations. We will inspect the combined positive/negative set to ensure that there is a reasonably even distribution of locations and times throughout the dataset.

### 4.1.3 Data Types and Type Conversion

Features are represented in the source datasets as follows:

* Date/time: ISO date-time string, or as “yyyymmdd” with the hour listed separately as “000”,
* Latitude, longitude and elevation: floats,
* Precipitation, humidity, acres, etc.: ints, but other datasets may use floats, and
* County: string.

We do not know how other datasets we find may encode their features, but we expect measurements to be in the form of ints or floats, classes to be labeled with ints or strings, latitude/longitude to be floats, and date/time to be in some sort of parsable string. We will convert the combined date/time features into separate features for the hour, day, month and year. Other non-numeric data types will need to be represented numerically before they can be used as inputs to the model.

We will convert any non-numeric features, such as vegetation type, to numeric ones, either by representing the data in question as a single numeric feature or by one-hot encoding a series of new features representing class labels from the original feature. We will take the former approach for non-numeric data that represents a spectrum. For example, if an input feature originally takes on values like “No Population,” “Sparsely Populated” up to “Very Densely Populated,” we could convert it to a numeric scale by representing “No Population” as 0 and “Very Densely Populated” as 10. For non-numeric features that represent distinct classes, like counties or vegetation types, we will take the second approach to conversion. For example, one feature that could originally take on ten different values including “Agriculture,” “Conifer Forest,” “Desert” and “Shrub” would be converted to a series of ten features for one-hot encoding, so the values of “Agriculture” and “Shrub” would be represented as [1,0,0,...,0] and [0,0,0,1,0,...,0], respectively.

If we are able to access datasets in the form of GeoJSON, like the vegetation dataset we are pursuing, the data stored in them will be nested in JSON objects and arrays. The exact formatting of these objects and their attributes will, of course, matter as we are using them to gather further features, but we expect them to follow similar standards to the ones above, with latitude and longitude as floats and class labels as strings or ints.

### 4.1.4 Combining Datasets

As previously discussed, we will construct the examples in our training and testing datasets by merging examples from multiple source datasets based on location and time. We do not expect any of the locations to match exactly (it would be a surprise to find a fire that ignited at the exact site of a weather monitoring station). In reality, the closest weather station to a fire ignition point will likely be miles away. The location and time features for the input to our model will be the exact location and time features from the Cal Fire *Incident Data* dataset or from our constructed, random locations/times for negative examples. Using these location and time features as anchors, we will select weather measurements (and vegetation or other measurements) from the nearest measurement locations and times.

Beginning with location features for a positive example (from the Cal Fire *Incident Data* dataset) or negative example (from our constructed, random locations/times), we will select weather measurements from the nearest weather measurement location in each of the weather datasets. If we are able to obtain shape datasets representing land use, vegetation, etc., we will simply check which shape in that dataset contains the lat/lon point for the example in question. Once we have found the nearest weather station to an example’s latitude and longitude, we will select the most recent measurement time.

### 4.1.5 Additional Feature Construction and Data Cleaning

We add features derived from the original datasets to provide the model with a longer-term view of the area in question. These features may include medium- and long-term precipitation, humidity and temperature, represented by the average of those values over a series of days (stored as a float). Given that the source datasets include the hourly or daily measurements for precipitation, humidity and temperature, calculating the medium- and long-term averages should be straightforward. We will also add features representing the fire history in the area around an example location, including the number of acres burned nearby in the preceding year or two, the number burned in a larger area in the preceding decade or two, and the time since the last nearby fire.

We will filter the datasets to extract only the features we need. We will then join the datasets together and clean them, eliminating invalid values and outliers. Then we will normalize the feature values so a feature like longitude with high values doesn’t overwhelm a feature like precipitation with low values. Depending upon the ultimate size of our final dataset, we may extract a random subset to use to train our model. Afterward, we will begin exploratory data analysis, as outlined in section 3.2.

# 5. Other Considerations: Assumptions and Vulnerabilities

## 5.1 Fire vs Fire Danger

Although the purpose of the model is to predict the danger or risk of a fire occurring, the class labels for each example are determined by whether or not a fire actually occurred. On the one hand, this could be considered a good way to label a location and time as having a high fire risk, since there is no better indicator that there is a risk of fire than for a fire to have occurred. On the other hand, this method of determining positive and negative class examples would mean that, at the location of a fire, the day before a fire, the fire-risk/no-fire-risk class label would be no-fire-risk, which would likely be inaccurate. This means that the model would be less likely to predict positive class labels, even for times when the conditions (location, weather, etc.) are similar to those when fires did occur. We think that is alright because the actual danger of a fire occurring at the exact location and time of an example is so small. The model inputs take exact latitude/longitude coordinates rather than larger geographic areas and times measured by the hour instead of by the day. While the fire danger for a larger geographic region over the course of a whole day may be high, the probability that a fire will break out at or right around the given coordinates and at or within an hour of the given time is very slim. If we see that the model is not sensitive enough to fire risk, we will adjust our labeling methods.

## 5.2 Data Quality

As discussed in section 4, the data we have access to is not particularly robust. For example, we may have to accept a wide variation in distances between fire risk locations and the weather monitoring stations that they are matched with. This will drastically impact the quality of our model’s predictions, but we believe that this will not have an overly negative impact on the outcomes of our experiments. If we were attempting to create a model to predict fire risk more successfully then the models discussed in section 2, which use high quality data, then the quality of our data would prevent us from being able to do so. However, we are attempting to show that considering longer term data, whether that is weather data or fire history, has a positive impact on model predictions. The quality of data will remain consistent across our baseline and experiment models, with the only changes being the additions of the features that we hypothesize will be helpful. Although data quality is a weakness, we believe that it will not prevent us from conducting worthwhile experiments.

# 6. Works Cited

Cal Fire. (2023). *Incident Data* (No version listed)[Data set]. Cal Fire. <https://www.fire.ca.gov/incidents#:~:text=FOR%20SOFTWARE%20DEVELOPERS-,Incident%20Data,-Incident%20data%20is>

Gutierrez, A. A., Hantson, S., Langenbrunner, B., Chen, B., Jin, Y., Goulden, M. L., & Randerson, J. T. (2021). Wildfire response to changing daily temperature extremes in California's Sierra Nevada. *Science advances, 7*(47), eabe6417. <https://doi.org/10.1126/sciadv.abe6417>

Holm, S. M., Miller, M. D., & Balmes, J. R. (2020). Health effects of wildfire smoke in children and public health tools: a narrative review. *Journal of Exposure Science and Environmental Epidemiology*, *31*(1), 1–20. <https://doi.org/10.1038/s41370-020-00267-4>

Kondylatos, S., Prapas, I., Ronco, M., Papoutsis, I., Camps-Valls, G., Piles, M., et al. (2022). Wildfire danger prediction and understanding with Deep Learning. *Geophysical Research Letters* *49*(e2022GL099368). <https://doi.org/10.1029/2022GL099368>

*Living under Smoky Skies—Understanding the challenges posed by wildfire smoke in California*. (2022, November 14). Legislative Analyst’s Office. Retrieved October 22, 2023, from <https://lao.ca.gov/Publications/Report/4644>

*Malik, A., Jalin, N., Rani, S., Singhal, P., Jain, S., & Gao, J.* (2021, October 1). Wildfire Risk Prediction and Detection using Machine Learning in San Diego, California*. IEEE Conference Publication | IEEE Xplore.* Retrieved October 22, 2023, from <https://ieeexplore.ieee.org/document/9604370/authors#authors>Malik

Menne, M. J., Durre, I., Korzeniewski, B., McNeill, S., Thomas, K., Yin, X., Anthony, S., Ray, R., Vose, R. S., Gleason, B. E. & Houston, T. G. (2012). *Global Historical Climatology Network - Daily (GHCN-Daily)*(3.30)[Data set]. NOAA National Climatic Data Center. doi:10.7289/V5D21VHZ [Accessed 10/19/2023] <https://www.ncei.noaa.gov/metadata/geoportal/rest/metadata/item/gov.noaa.ncdc:C00861/html>

National Centers for Environmental Information, NESDIS, NOAA, U.S. Department of Commerce. (2005). *U.S. Local Climatological Data (LCD)*(No version listed)[Data set]. [Accessed 10/19/2023] <https://www.ncei.noaa.gov/metadata/geoportal/rest/metadata/item/gov.noaa.ncdc:C00684/html#>

National Park Service (NPS). *Understanding Fire Danger.* Retrieved October 20, 2023, from <https://www.nps.gov/articles/understanding-fire-danger.htm>

Shmuel, A. & Heifetz, E. (2023). Developing novel machine-learning-based fire weather indices. *Machine Learning: Sci. Technology* *4*(015029) <https://doi.org/10.1088/2632-2153/acc008>

Short, K. C. (2017). *Spatial wildfire occurrence data for the United States, 1992-2015* *(FPA\_FOD\_20170508)*.(4th Edition)[Data set]. Fort Collins, CO: Forest Service Research Data Archive. <https://doi.org/10.2737/RDS-2013-0009.4> Accessed via Kaggle: Tatman, R. *1.88 Million US Wildfires* [Data set]. Kaggle. <https://www.kaggle.com/datasets/rtatman/188-million-us-wildfires/data>

*Statistics | CAL FIRE*. (n.d.). California Department of Forestry and Fire Protection. Retrieved October 22, 2023, from <https://www.fire.ca.gov/our-impact/statistics>

Tyler, J. (Director). (2022). *2022 Wildfire Activity Statistics*. California Department of Forestry and Fire Protection. Retrieved October 22, 2023, from <https://34c031f8-c9fd-4018-8c5a-4159cdff6b0d-cdn-endpoint.azureedge.net/-/media/calfire-website/misc-doc/2022redbookfinalada.pdf?rev=19e8d1a007884c34a181e8138fc7017e&hash=7BA59394712297355511D58394E7B8E2>

USFSEnterpriseContent. (2022). *National USFS Fire Occurrence Point (Feature Layer)* (Modified 2023-09-28)[Data set]. U.S. Forest Service. <https://catalog.data.gov/dataset/national-usfs-fire-occurrence-point-feature-layer-d3233>

Youssouf, H., Liousse, C., Roblou, L., Assamoi, E., Salonen, R. O., Maesano, C. N., Banerjee, S., & Annesi‐Maesano, I. (2014). Non-Accidental health impacts of wildfire smoke. *International Journal of Environmental Research and Public Health*, *11*(11), 11772–11804. <https://doi.org/10.3390/ijerph111111772>

Zaidi A. (2023). Predicting wildfires in Algerian forests using machine learning models. *Heliyon, 9*(7), e18064. <https://doi.org/10.1016/j.heliyon.2023.e18064>

# 7. Appendices

## 7.1 Appendix A: Literature Review Notes

### 7.1.1 Peer-Reviewed (Advanced) Approaches

* 1. [Wildfire Danger Prediction and Understanding With Deep Learning](https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2022GL099368)
  2. Type of model
     1. Random forest [RF], XGBoost, Long-Short Term Memory [LSTM], and Convolutional Long-Short Term Memory [convLSTM]
  3. Features Used
     1. Daily weather data from ERA-5 Land (Muñoz-Sabater et al., 2021) of maximum 2 m temperature, maximum wind speed, minimum relative humidity, total precipitation, maximum 2 m dewpoint temperature, and maximum surface pressure. The type of daily aggregations chosen represents the most fire-aggravating conditions.
     2. Satellite variables from MODIS including Normalized Difference Vegetation Index (NDVI; Didan, 2015), day and night Land Surface Temperature (LST; Wan et al., 2015).
     3. Soil moisture index from the European Drought Observatory (Cammalleri et al., 2017).
     4. Roads distance, waterway distance, and yearly population density from WorldPop (Tatem, 2017).
     5. Elevation and Slope from Copernicus EU-DEM (Bashfield & Keim, 2011).
     6. Ten variables with the fraction of classes from Copernicus Corine Land Cover (Büttner, 2014).
  4. Keypoints
     1. The study aims to predict wildfire danger using deep learning models. The models were trained on large datasets of meteorological data and satellite images. The study also explored the interpretability of the models to understand which features are most influential in predicting wildfire danger. The models are used over 2 test sets (2020 and 2021) and metrics (F1, recall, etc) provide a gauge on the performance of each.
     2. Drawbacks
        1. "The results are better for 2021 than for 2020 for all the models. Precision is higher, probably because negatives are more distinguishable. We sample negatives on days when no fire occurred, but few days did not have fires in the summer of 2021. Thus, more negatives are sampled outside the summer, which might make their classification easier."
     3. Strengths
        1. The deep learning models, especially LSTM and ConvLSTM, demonstrated superior performance with F1-scores greater than 0.8, outperforming traditional models like RF and XGBoost.
        2. The models showed strong generalization capabilities, especially for the extreme fire season of 2021, indicating their ability to adapt to varying conditions and predict significant fire events.
  5. [Wildfire Risk Prediction and Detection using Machine Learning in San Diego, California](https://ieeexplore.ieee.org/document/9604370/citations#citations)
  6. Type of model
     1. We used Support Vector Machine (SVM), eXtreme Gradient Boosting (XGBoost), and Random Forest (RF) algorithms for weather based fire prediction.
     2. CNN, LTSM and SVM for remote sensing based fire prediction models.
     3. Combine them into an ensemble model.
  7. Features Used
     1. Weather data (temperature, humidity, wind speed, etc.) (SECTION IV.A)
     2. Satellite data (Landsat 8 images) (SECTION IV.B)
     3. Historical fire data (SECTION IV.C)
     4. Vegetation data (Normalized Difference Vegetation Index - NDVI) (SECTION IV.D)
  8. Keypoints
     1. Aimed to develop a machine learning and big data-based fire risk prediction model. Achieved an accuracy of 100% using the ensemble model for fire risk prediction. Achieved 93% accuracy using Faster R-CNN for fire detection. Utilized various machine learning and deep learning techniques to achieve these results.
     2. Drawbacks
     3. Strengths
        1. Multiple data sources are used and they combine 2 ML models into one ensemble model. Obtained accuracy of 100% with our ensemble model for fire risk prediction and 93% for fire risk detection.
        2. Unlike other research that examines either fire risk detection or fire risk prediction with limited data and parameters, our work focuses on understanding these concepts using past fire events, weather, remote sensing, and satellite data.
  9. [Predictive modeling of wildfires: A new dataset and machine learning approach](https://www.sciencedirect.com/science/article/abs/pii/S0379711218303941?fr=RR-2&ref=pdf_download&rr=819eb5e4e95ef953)
  10. Neural nets MLP Classifier, SVM
  11. Features Used
      1. This is based on satellite imagery. They used fire zones to classify the images as fire or no fire.
      2. Go to the appendix at the bottom and you can download the data they used.
  12. Keypoints
      1. General
      2. Drawbacks
         1. Future works will mainly consist of strengthening the model by including weather data. Weather plays a major role in the occurrence, growth, spread and the Extinction of wildfires. It can impact on the strength and movement of fire, and thus burn more land, which makes its extinction even more difficult. There are three weather parameters that can affect wildfires: Air Temperature, Wind and Soil Moisture.
         2. Air Temperature influences the occurrence of wildfires, by heating trees and crops on the ground, which makes them sensitive toward catching fire.
         3. The Wind has the most prominent and strongest impact on wildfires behaviors. Wind speed and direction are unpredictable. Besides, winds supply the fire with additional oxygen, which pushes the fire to move faster across the land.
         4. Soil Moisture is directly affected by precipitation and air humidity. When the soil moisture is low, the risk of wildfires is high. Conversely, high soil moisture lowers the chances of a wildfire igniting,
      3. Strengths
  13. [Developing novel machine-learning-based fire weather indices](https://iopscience.iop.org/article/10.1088/2632-2153/acc008)
  14. Type of Model
      1. We develop four different classification models: (i) RF (Biau and Scornet 2016). (ii) Extreme Gradient Boosting (XGBoost) (Chen and Guestrin 2016). (iii) MLP, a form of Neural Network (Ramchoun et al 2016), and (iv) logistic regression (Lever et al 2016). We perform a simple train-test split where 25% of the observations are used for testing. We perform these analyses using Python's Scikit-learn package (Pedregosa et al 2011 ), apart from the XGBoost model which is based on the XGBoost package (Chen and Guestrin 2016).
  15. Features Used
      1. Daily ignition, 2 m temperature, Relative humidity, 10-m wind speed, Precipitation, Mean relative humidity in previous month, Mean precipitation in previous month, Mean relative humidity in previous year, Mean precipitation in previous year, Mean slope, Population density, Normalized difference vegetation index, Incoming short-wave solar radiation, Daily fire weather index, Daily build up index, Daily danger index, Daily drought code, Daily duff moisture code, Daily initial fire spread index, Daily fine fuel moisture code, Daily fire daily severity rating, Daily Keetch-Byram drought index, Daily fire danger index, Daily spread component, Daily energy release component, Daily burning index, Daily ignition component, Regional wildfire history.
  16. Keypoints
      1. General
         1. To enable accurate wildfire danger predictions by applying advanced ML models using various data. To analyze the most significant factors affecting wildfire risk, both individually and in interaction with additional factors. To demonstrate the potential of applying ML-based FWIs in actual fire warning systems. To examine the contribution of adding traditional FWIs (in addition to the raw data) as input data when training ML-based models. To examine the potential of predicting extreme wildfires using ML-based models.
         2. There is a figure showing the most significant features and their effects on the predictions and another one showing how logistic regression performs on this data.
         3. ML models significantly outperformed traditional Fire Weather Indices (FWIs) and Logistic Regression models.
         4. XGBoost achieved the highest accuracy with an ROC-AUC of 0.98.
         5. Temperature was the most influential factor in predicting wildfire risk.
         6. Complex interactions between features like wind and humidity were captured better by ML models.
      2. Drawbacks
      3. Strengths
         1. This seems like it has some applicable information to our work, they are using some similar data and they are even training a logistic regression model on it. Of course they use some more sophisticated models that perform better than the simpler models and the traditional fire weather indices.
  17. [A review of machine learning applications in wildfire science and management](https://cdnsciencepub.com/doi/full/10.1139/er-2020-0019)
  18. Keypoint
      1. This is actually a scoping review that provides a good background on the topic and I will actually use it for the background section.
  19. [Deep Learning Methods for Daily Wildfire Danger Forecasting](https://arxiv.org/abs/2111.02736)
  20. Type of Model
      1. Four different types of dataset samples (pixel, temporal, spatial, and spatio-temporal) are extracted from the datacube and then fed to the corresponding models (RF, LSTM, CNN, and convLSTM). For a dataset sample, nf is the number of input features, days is the number of days in the time series, h is the height and w is the width, wherever applicable. The pixel dataset has a different number of input features (n′f ), because it also includes aggregations of the dynamic features.
  21. Features Used
      1. Deep Learning Methods for Daily Wildfire Danger Forecasting" study encompass daily weather data from ERA-5 Land, which includes 2 m temperature, 10 m wind u-component, 10 m wind v-component, and total precipitation. Additionally, satellite variables from MODIS are incorporated, such as Leaf Area Index (LAI), Fraction of Photosynthetically Active Radiation (Fpar), Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI), and Day/Night Land Surface Temperature (LST). The dataset also integrates roads density derived from OpenStreetMap, population density from WorldPop, land cover information from Copernicus Corine Land Cover (CLC), topography variables like elevation, aspect, and slope from Copernicus EU-DEM, and historical burned areas data from the European Forest Fire Information System (EFFIS) combined with the MODIS active fires product to determine the fire's start date
  22. Keypoints
      1. General
         1. We approach daily fire danger prediction as a machine learning task, using historical Earth observation data from the last decade to predict next-day’s fire danger. To that end, we collect, pre-process and harmonize an open- access datacube, featuring a set of covariates that jointly affect the fire occurrence and spread, such as weather conditions, satellite-derived products, topography features and variables related to human activity.
         2. We implement a variety of Deep Learning (DL) models to capture the spatial, temporal or spatio-temporal context and compare them against a Random Forest (RF) baseline. We find that either spatial or temporal context is enough to surpass the RF, while a ConvLSTM that exploits the spatio-temporal context performs best with a test Area Under the Receiver Operating Characteristic of 0.926.
      2. Drawbacks
      3. Strengths
  23. [Predicting wildfires in Algerian forests using machine learning models](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10372657/)
  24. Type of Model
      1. We developed an artificial neural network (ANN) with two hidden layers to predict wildfires in these cities (diagram in section 5.1). Next, we trained and compared the performance of our classifier with those provided by the Logistic Regression, K Nearest Neighbors, Support Vector Machine, and Random Forest classifiers, using a 10-fold stratified cross-validation. The experiment shows a slight superiority of the ANN classifier compared to the others, in terms of accuracy, precision, and recall.
  25. Features Used
      1. The attributes in the dataset include Temperature at 12 noon in degrees Celsius (Temp), Relative Humidity in % (RH), Wind speed in km/h (Ws), 24 hours of accumulated precipitation in mm from noon to noon (Rain), Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC), Drought Code (DC), Initial Spread Index (ISI), Buildup Index (BUI), Fire Weather Index (FWI), and the target variable representing fire occurrence (Target), which takes binary values {fire: 1 | no fire: 0}.
  26. Keypoints
      1. General
         1. There are many images and diagrams summarizing the process, explaining the models and comparing the scores.
      2. Drawbacks
      3. Strengths
  27. [interpretable machine learning fire model for burned-area predictions over tropics](https://gmd.copernicus.org/articles/16/869/2023/)
  28. Type of Model
      1. Note code is available in an online repo.
      2. Burned areas? Is this predicting the spread of wildfires or whether they will start? I dont think it's relevant but I will keep it in.
      3. The AttentionFire model is an interpretable machine learning model based on an attention-augmented Long Short-Term Memory (LSTM) framework, which uses attention mechanisms to dynamically select and assign weights to important drivers and time steps for wildfire predictions, addressing limitations of traditional LSTM models in terms of interpretability and feature importance. (Diagram in figure 1)
      4. Five other widely used machine learning (ML) models are used as baseline models to compare with AttentionFire model: ANN (Joshi and Sukumar, 2021; Zhu et al., 2022), decision tree (DT) (Amatulli et al., 2006; Coffield et al., 2019), random forest (RF) (Yu et al., 2020; Li et al., 2018; Gray et al., 2018), gradient-boosting decision tree (GBDT) (Coffield et al., 2019; Jain et al., 2020), and naive LSTM (Liang et al., 2019; Natekar et al., 2021; Gui et al., 2021; Mei and Li, 2019)
  29. Features Used
      1. Go to table 2 in section 2.3 to see the table of variables.
  30. Keypoints
      1. General
      2. Drawbacks
      3. Strengths
  31. [Fire danger forecasting using machine learning-based models and meteorological observation: a case study in Northeastern China](https://link.springer.com/article/10.1007/s11042-023-15881-1)
  32. Type of Model
      1. LTSM based time series prediction and Random Forest based fire danger occurrence probability.
  33. Features Used
      1. From 2004 to 2015, 3963 fire occurrences were recorded in Northeastern China, comprising 1088 natural wildfires (including forest, grassland, and shrubland fires) and 2875 human-prescribed burnings; data visualizations indicated peak occurrences in 2007 and more frequent fires during spring and summer.
      2. The study employed the Canadian Forest Fire Weather Index (FWI) system, which uses four key meteorological data types—temperature, relative humidity, wind speed, and 24-hour precipitation—sourced from daily observations between 1989 to 2019 at approximately 133 meteorological stations in the region, with fire weather data matched to each fire event based on its date and nearest meteorological station.
  34. Keypoints
      1. General
         1. Proposed a two-stage fire danger rating and forecasting model.
         2. First stage predicts future FWI system indexes using LSTM.
         3. Second stage uses Random Forest to determine fire danger occurrence probability and presents a classification scheme.
         4. Verified the scheme using forest fire danger data from Qipan Mountain.
         5. Fire danger forecasting is complex, and while traditional binary classification methods fall short, our two-stage prediction method uses the FWI system indexes from weather data to create a more nuanced classification, dividing fire danger into four categories.
      2. Drawbacks
         1. The study's model predicts fire danger ratings up to four days in advance, which is an improvement over typical short-term predictions but could benefit from further extension.
         2. The model's applicability is currently limited to Qipan Mountain and Northeastern China due to data collection constraints, but future work aims to expand and verify its effectiveness in other regions.
      3. Strengths
  35. [Predicting California Wildfire Risk with Deep Neural Networks](http://cs230.stanford.edu/projects_fall_2021/reports/103174984.pdf)
  36. Type of Model
      1. Conv Neural Networks (CNNs)
      2. Specifically, pretrained ImageNet models: MobileNetV2 and ResNetV2
  37. Features Used
      1. High-resolution satellite imagery of the California countryside.
      2. Historical fire map data from CalFire.
      3. NSGS Landsat 7 and 8 imagery for California over the last 10 years.
  38. Keypoints
      1. General
         1. The goal is to identify seasonal fire risk in California using satellite imagery.
         2. The models classify terrain based on historical images of areas that experienced wildfires.
         3. The study faced challenges due to high problem bias and applied techniques to reduce this bias.
         4. The input is a satellite image, and the output predicts if the area is at risk of wildfire for the season.
      2. Drawbacks
         1. Initial struggles with effective identification due to high problem bias.
         2. Data imbalance with the not-at-risk class being heavily overrepresented.
         3. Some satellite images were obscured by cloud cover.
         4. The model had difficulty identifying positive classifications.
         5. Fires caused by human incidents might not have identifiable features in satellite images.
      3. Strengths
         1. Utilized advanced CNNs like MobileNetV2 and ResNetV2.
         2. Incorporated learning rate optimization, channel feature addition, and cost weighting to improve model performance.
         3. Demonstrated the potential of AI-based techniques for estimating fire risk in California.
  39. [Identifification and Characterization of Forest Fire Risk Zones Leveraging Machine Learning Methods](https://scholar.smu.edu/cgi/viewcontent.cgi?article=1188&context=datasciencereview)
  40. Type of Model
      1. Binary Classification
  41. Features Used
      1. Historical weather data (temperature, precipitation, wind, etc.)
      2. Past fire occurrences
      3. Population density
      4. MGRS (Military Grid Reference System) coordinates
      5. Historical drought data
      6. Hours of sunlight per day
  42. Keypoints
      1. General
         1. They also have a feature importance figure.
         2. Literature review also included.
         3. Also they implemented a coordinate/grid system in this paper.
         4. The study focuses on the Southern California region.
         5. The model predicts the probability of a fire occurring in a specific area within the next one to five days.
         6. The model uses the MGRS grid system to define geographic areas and correlate multiple fires to a specific area.
         7. The study used a Kaggle dataset with 1.88 million georeferenced fires collected over 24 years.
         8. The model's predictions were visually inspected for accuracy by comparing them to known fire-prone areas.
      2. Drawbacks
         1. The model's predictions remain relatively constant throughout the year, not adjusting well for seasonal variations.
         2. The study is limited to the Southern California region.
         3. Ethical concerns include potential misuse by arsonists or for insurance fraud.
      3. Strengths
         1. The model can be a valuable tool for government and fire officials for proactive fire prevention measures.
         2. The study highlights the importance of feature engineering and the potential for enhancing current fire detection methods.
         3. The model's results align with historically known fire-prone areas, indicating its reliability.
  43. [Combining precipitation forecasts and vegetation health to predict fire risk at subseasonal timescale in the Amazon](https://iopscience.iop.org/article/10.1088/1748-9326/ac76d8/meta)
  44. Type of Model
      1. The model is a hybrid combining both dynamic and statistical methods, based on the NextGen prediction system.
      2. Its main goal is to predict fire risk in the Amazon.
      3. Fire risk is determined by the likelihood of active fires occurring due to precipitation and vegetation health.
      4. The model uses VIIRS active fires data and predicts based on SubX's week-2 precipitation forecasts from 2017-2021.
      5. A secondary experiment adds the NOAA Vegetation Health Index (VHI) as an extra predictor, which detects how ecosystems respond to weather conditions.
      6. The study then examines the model's accuracy in two specific Amazon regions with different fire seasons.
      7. It also investigates how different land covers in these regions affect the model's predictions.
  45. Features Used
      1. VIIRS Active Fires: Active fire detections.
      2. SubX MME Week-2 Precipitation: Rainfall predictions two weeks ahead (2017-2021).
      3. NOAA Vegetation Health Index (VHI): Satellite-based vegetation health measure.
      4. Mapbiomas Land Cover (2019): Brazil's Amazon land cover data.
      5. Fire Activity (FA): Most active fire trimester data from VIIRS.
      6. SubXPr and VHI Weekly Data: 13-week data within the peak fire trimester.
      7. Fire Probability Analysis: Logistic regression models using SubXPr and VHI.
      8. Land Cover Data (2019): Predominant land types in the Amazon from Mapbiomas.
  46. Keypoints
      1. General
         1. The study focuses on bridging the gap between atmospheric scientists and fire ecologists by combining observed states of vegetation with precipitation forecasts.
         2. The research identifies areas in the Amazon where fire risk can be skillfully predicted using SubXPr and VHI.
         3. In areas dominated by forests, fire risk is mainly driven by SubXPr. In contrast, in areas with prevalent savannas and grasslands, the status of VHI plays a significant role in determining fire risk.
         4. The study emphasizes the potential of using subseasonal forecasts in combination with vegetation health to forecast fire risk, providing crucial lead time for fire management resources allocation.
      2. Drawbacks
         1. The role of vegetation health as a predictor of fire risk is modest in areas dominated by forests.
         2. The study suggests that continuous improvement of statistical models' accuracy might be achieved using higher spatial resolution satellite vegetation conditions.
      3. Strengths
  47. [Improved Prediction of Forest Fire Risk in Central and Northern China by a Time-Decaying Precipitation Model](https://www.mdpi.com/1999-4907/13/3/480)
  48. Type of Model
      1. Support Vector Machine (SVM) regression model
      2. Time-decaying precipitation algorithm
  49. Features Used
      1. Relative humidity
      2. Daily maximum temperature
      3. Daily maximum wind speed
      4. Comprehensive precipitation index
      5. NDVI (Normalized Difference Vegetation Index)
      6. VSWI (Vegetation Supply Water Index)
      7. Altitude
      8. Fire density
      9. Slope
  50. Keypoints
      1. General
         1. Four forest areas in central and northern China were selected for training and testing.
         2. The model uses a time-decaying precipitation algorithm instead of traditional equal-weighting.
         3. Gaussian convolution was used to convert discrete fire spots into continuous forest fire density data.
         4. The model achieved accuracies of 0.98, 0.99, 0.95, and 0.93 in the four test areas.
      2. Drawbacks
         1. The study doesn't mention the use of deep learning, which might provide better accuracy with sufficient samples.
         2. The model might require re-optimization for different geographical and climatic conditions.
      3. Strengths
         1. The model improved fire risk prediction accuracy by approximately 10%.
         2. SVM regression model outperformed logistic regression and ANN models.
         3. The model is applicable to diverse climatic and geographical conditions, demonstrating its universality.

### 7.1.2 Individual (Fundamental) Approaches

Kaggle Competitions

1. <https://www.kaggle.com/competitions/forest-fire-prediction/data>
   1. Data: day, month, year, temp, relative humidity, wind speed, rain, fine fuel moisture code, duff moisture code, drought code, initial spread index, buildup index, fire weather index
   2. approaches: I wasn’t able to see the competition code or discussion, but that may be because I wasn’t logged in. leaderboard scores are at 1.00000, but idk what that means
2. <https://www.kaggle.com/competitions/ausdm-student-competition-2015/data>
   1. Data: temp min/max, relative humidity min/max, feels like, rain since 9 am, wind speed min/max/avg, wind gust
   2. approaches: again, not able to see. best score on leaderboard was .58050, but idk how the score was calculated
   3. [Wildfires in California prediction (Kaggle)](https://www.kaggle.com/code/sandstorm0123/wildfires-in-california-prediction/notebook)
   4. Type of Model
      1. Machine Learning Model to predict wildfires in California based on weather events.
   5. Features Used
      1. US Weather Events (2016-2019)
         1. EventId, Type, Severity, StartTime(UTC), EndTime(UTC), TimeZone, AirportCode, LocationLat, LocationLng, City, County, State, ZipCode.
      2. California Wildfire Incidents (2013-2020)
         1. Same data we are using
   6. Keypoints
      1. General
         1. The goal is to combine weather events at different stations with wildfire incidents to uncover less obvious relations.
         2. The current version visualizes rain events and acres burned by wildfire.
         3. The model tries to predict if 2019 was a year with a lot of wildfires based on weather data.
         4. The model is trained on data from 2016 to 2018 and tested on 2019 data.
         5. Uses TensorFlow for creating and training the model.
      2. Drawbacks
         1. The approach is still in a rough state and not finished.
         2. The model is specialized in only three years and might have overfitting issues when applied to other years.
      3. Strengths
   7. [AI Geospatial Wildfire Risk Prediction (Kaggle)](https://www.kaggle.com/code/thjaquenoud/ai-geospatial-wildfire-risk-prediction)
   8. Type of Model
      1. Deep Learning Model: U-Net for multi-class semantic segmentation.
   9. Features Used
      1. MODIS: Daily U.S. coverage at 500m resolution.
      2. GRIDMET: Daily meteorological data at 4000m resolution, including temperature, humidity, and drought indices.
      3. LANDFIRE MFRI: 30m pixel categorization based on historical wildfire intervals. Useful for regional trends but static over time.
      4. USDA Cropland Data Layer: 30m resolution land cover map. Used to differentiate cultivated from non-cultivated areas.
      5. Output - Wildfire Hazard Potential (WHP) map: USDA's evaluation of wildfire hazard, available for 2014, 2018, and 2020.
      6. Processing: Data aligned using GDAL and segmented into smaller images for model processing.
      7. Final dataset: 13-layer geospatial images over the CONUS for three years.
   10. Keypoints
       1. General
          1. Objective: Generate wildfire risk maps weekly over the CONUS.
          2. U-Net architecture adapted for the specific input dimensions and multi-class output.
          3. Model trained with a weighted loss function to account for class imbalances.
          4. Achieved 65% accuracy on training and validation data.
          5. Model generalizes to months outside of its training domain, e.g., predicting wildfire hazard in winter months.
       2. Drawbacks
          1. Difficulty in obtaining labeled data for "at-risk areas."
          2. Domain gap due to the use of WHP map which is generated once every few years.
          3. Model sometimes underestimates the highest level of risk and overestimates in certain areas.
          4. Noisy outputs and grid-like artifacts in winter predictions.
       3. Strengths
          1. Demonstrates the potential of using geospatial data in deep learning models for large-scale tasks.
          2. Model can recreate the WHP map for years when it wasn't available.
          3. Evidence of model learning to generalize temporal differences from geographic ones.
   11. [Brazilian Wildfire Prediction (Kaggle)](https://www.kaggle.com/code/data13/brazilian-wildfire-prediction/input)
   12. Type of Model
       1. Linear Regression
   13. Features Used
       1. Year: When Forest Fires happened (Range: 1998-2017)
       2. State: Brazilian State (e.g., Rio, Mato Grosso)
       3. Month: Month when Forest Fires happened (e.g., Janeiro, Fevereiro)
       4. Number: Number of Forest Fires reported (Range: 0-998)
       5. Date: Date when Forest Fires were reported (Range: 01/01/1998 - 01/01/2017)
   14. Keypoints
       1. General
          1. The dataset reports the number of forest fires in Brazil divided by states over approximately 10 years (1998 to 2017).
          2. Data Exploration revealed that the dataset is structured and almost cleaned.
          3. Data Cleaning involved removing duplicates, changing month names to English, and handling special characters in state names.
          4. Exploratory Data Analysis (EDA) showed the distribution of fires per year, month, and state.
          5. Data Preprocessing included encoding categorical data, extracting new features like day of the week, and determining weekends.
          6. The Linear Regression model was used to predict the number of wildfires in the Amazon rainforest.
       2. Drawbacks
          1. The accuracy of the model is not satisfactory. The dataset's small size leads to overfitting, resulting in a low training error and a high test error.
       3. Strengths
          1. The dataset provides a comprehensive overview of forest fires in Brazil over a significant period.
          2. The notebook includes a thorough exploratory data analysis, visualizations, and preprocessing steps.
   15. [Australian Wildfires Prediction (Kaggle)](https://www.kaggle.com/code/ahmednobi/australian-wildfires-prediction/input)
   16. Type of Model
       1. RandomForestRegressor
   17. Features Used
       1. Latitude
       2. Longitude
       3. Brightness
       4. Scan and Track (pixel sizes; later dropped 'track')
       5. Acquisition Date and Time
       6. Satellite (Terra and Aqua)
       7. Confidence (target variable)
       8. Brightness temperature 31 (dropped during feature engineering)
       9. Fire Radiative Power
       10. Day/Night
       11. Type (encoded into type\_0, type\_2, and type\_3 during preprocessing)
   18. Keypoints
       1. General
          1. The dataset has over 36,000 entries and 15 initial features.
          2. Features such as 'track', 'instrument', and 'version' were dropped based on correlation.
          3. Features 'daynight' and 'satellite' were converted to binary values; 'type' was one-hot encoded.
          4. Derived year, month, and day from 'acq\_date' for better feature representation.
          5. The initial RandomForestRegressor had a testing accuracy of 64.66%.
          6. After hyperparameter tuning, the testing accuracy improved to 66.61%.
       2. Drawbacks
          1. Model might be overfitting given the high training accuracy compared to testing accuracy.
          2. Some features like 'bright\_t31' and 'type\_0' were dropped without clear justification.
       3. Strengths
          1. Extensive data preprocessing to enhance model performance.
          2. Hyperparameter tuning helped in improving the testing accuracy.
          3. The model captures a significant amount of variance in the training set.

1. Nine percent of the fires in the *National USFS Fire Occurrence Point (Feature Layer)* dataset were list camping as the cause. [↑](#footnote-ref-0)