HIGH SCHOOL EDITION
Journal of Student Research

Volume 14 Issue 1 (2025)

A Rigorous Comparison of Various Machine Learning Models for Next-Day Wildfire Detection

Ayan Dalmia

Millburn High School, USA

ABSTRACT

Wildfires in the western United States pose significant threats to large, barren areas, as well as more populated, urban lands, especially in wildfire-prone areas in the West. This study aims to develop an Artificial Intelligence model that predicts wildfires on a next-day basis, enabling timely mitigation and safety efforts, which is beneficial, especially in high-risk areas like the Western United States. Understand how safety and health-related measures can be implemented by predicting how the wildfire will spread over a day. Environmental factors such as elevation, maximum and minimum temperatures, wind speed, humidity, precipitation, drought indices, vegetation health metrics (Normalized Difference Vegetation Index), population density, energy release, and initial fire masks (spreads) are features used by the model to make predictions about the spread of wildfires over 24-hours. The model's performance was evaluated based on its precision and recall in predicting fire spread within a 64 km x 64 km area. The results demonstrate the neural network's strong capability to identify high-risk areas for wildfires with an 84% accuracy, improved by boosting with a Random Forest Classifier to 99% accuracy. By providing reliable next-day predictions, this model represents a valuable tool for wildfire management, enabling authorities to implement preemptive measures that could significantly reduce the impact of wildfires. This research and model development contributes to the broader field of disaster prevention and management, offering a data-driven approach to mitigating wildfires on a short-term basis.

Introduction

Wildfires pose a significant threat to natural ecosystems as well as human health and property, particularly in regions like the western United States where such events are becoming increasingly frequent and severe. The unpredictability of wildfire spread exacerbates the risks, leading to devastating housing loss and the release of harmful particulate matter into the atmosphere. As wildfires become more common and intense due to climate change and other factors, the need for advanced predictive models to better mitigate their impacts in advance becomes increasingly urgent, especially at specific resolutions to clarify exactly which areas most require protection. A spatial resolution of 1 kilometer by 1 kilometer is effective and specific, having been used in other studies of wildfire-based behaviors (Aguilera et. al 2023). There are immense potential health benefits of smoke mitigation, which can go in hand with predicted locations for wildfires (Yu et. al 2019). The dataset utilized in building the model was sourced from monitoring sites distributed across the United States, capturing a variety of economic conditions, thereby enabling the model to predict wildfire behavior across diverse regions.

Wildfires devastate local environments, particularly through particle pollution of air. Wildfires are a significant source of PM2.5, fine particulate matter consisting of fine particles that can penetrate deep into the lungs, posing serious respiratory, cardiovascular, and cerebrovascular risks, as studied in high-income nations (Chandra et. al 2020). Wildfires have accounted for up to 25% of PM2.5 in recent years across the United States, and up to half in some Western regions as compared to <20% in 2011 (Burke et. al 2021). There have been increasing trends in the 98th quantile of PM2.5 levels at monitoring sites in the Northwest United States of areas burned by wildfires, with a 95% confidence interval, meaning that there is a 95% probability that the observed trend falls within the calculated range, indicating a reliable statistical result (McClure et. al 2018). The Northwest is relevant for wildfire detection because it has seen a rise in wildfires in recent years, and is particularly susceptible to wildfires due to its arid climate and high density of vegetation (Liu et. al 2017). This elevated particulate matter substantially degrades air quality over large areas (PNAS) and creates significant health risks. A study estimated that hazard ratios for cardiovascular mortality, associated with a 10-μg/m3 exposure increment of PM2.5 were 1.34 (Pope et al. 2018). This means that for every 10 μg/m³ increase in PM2.5, the risk of dying from cardiovascular-related causes increases by 34%. Increased mortality rates are coupled with heightened risks for respiratory and cardiovascular conditions in the contiguous United States (Burke et. al 2021) due to PM2.5 correlation with wildfires. Additionally, PM2.5 exposure from wildfires can result in harm to public health, including increased hospital admissions and exacerbated pre-existing health conditions. A 7.2% increase in the risk of respiratory admissions during smoke wave days was found with high wildfire-specific PM2.5 (Liu et. al 2017). In fact, early predictions can avoid premature deaths in the US population aged 65+ based on predictions of high $PM_{2.5}$ concentrations (Burke et. al 2021).

Wildfires also impact housing negatively, leading to significant property damage and economic losses. From 1990 to 2010, the Wildland-Urban Interface (WUI) in the U.S. saw rapid growth, with the number of houses increasing by 41% (from 30.8 to 43.4 million) and land area expanding by 33% (Radeloff et. al 2018), growth that is continual and significant. The intersection of wild spaces and urban environments at the WUI creates a significant risk for property damage by wildfires (Radeloff et. al 2005). From 2000 to 2019, exposure to large wildfires in the contiguous United States of infrastructural roads and powerlines increased 58% and 70%, 82% of which exposure was in the western U.S. (Modaresi Rad et. al 2023). From 2017 to 2021, large wildfires in the United States led to $16.8 billion in damages annually, particularly concerning California (Biswas et. al 2023). From 1985 to 2010, there has been a consistent increase in spending on wildfire suppression by the US government, which has reached nearly $3 billion/year in federal expenditure (Burke et. al 2021). Additionally, increasing frequency and intensity of wildfires, driven by anthropogenic climate change since 2010, exacerbate burned areas by 10 times, increasing costs (Jones et al. 2020). This vulnerability is compounded by the increasing urban sprawl into these high-risk areas, leading to higher potential losses. The combined effects of rising wildfire activity and expanding residential development necessitate improved fire management and mitigation strategies to protect housing and reduce economic loss.

The purpose is to build a neural network that can accurately predict next-day wildfire spread at a spatial level using environmental and geographical data. This study aims to address the critical challenge of forecasting wildfire activity across the western United States, where such predictions are essential for effective disaster mitigation. Wildfire prediction is a classification problem involving supervised learning. The model's objective is to predict whether each square in a 64x64 grid (1 km² per square) will experience wildfire activity the following day. The input data includes quantitative variables such as elevation, maximum and minimum temperatures, wind speed, humidity, precipitation, drought index, vegetation index (NDVI), population density, energy release component (ERC), and previous fire masks for each 1 km² area. These factors are processed into 64x64 pixel grids, normalized for uniformity, and used to train a Neural Network (NN). The importance of this model is its potential to provide accurate, localized wildfire predictions that can enable short-term mitigation. Short-term solutions, such as the application of polyphosphate fire retardants, can be concentrated in areas predicted to be burned (Yu et. al 2019). Accurate predictions would allow for better allocation of resources, potentially reducing damage to life, property, and nature. The NN model, designed to capture patterns in the data, is trained on historical wildfire and environmental data. The output is a probabilistic prediction of fire occurrence for each grid square. Model accuracy will be assessed and validated against historical fire events.

We plan to make next-day predictions, by answering the questions:

I. *How can the spread of wildfires be accurately predicted by base models and Neural Networks?*
II. *How can the robustness of these models be tested?*

In the next section, we will consider the methods and models used in data manipulation to effectively undergo wildfire prediction, followed by a discussion and conclusion of the results of the spread predicted by the models.

Methodology and Models

Our study aims to build upon the work of those such as Sayady et al. (2019), who utilized Artificial Intelligence, Big Data, and Remote Sensing to predict wildfires using satellite data such as NDVI, LST, and Thermal Anomalies from MODIS. While their approach focuses on large-scale, broad-area predictions, our research narrows the scope to focus on localized environmental and geographical data, in some cases focusing on 1 km x 1 km areas. Simultaneously, our data is capable of studying large 64km x 64km areas, allowing for large-scale mitigation in targeted areas such as the Western United States. Additionally, our study will use similar variables, including NDVI and temperature data, but will also incorporate other variables such as population, wind speed and direction, energy components, and more.

Our study also aims to build upon the work of Shadrin et al. (2024), who utilized deep neural networks and multimodal data to predict wildfire spread over large areas. Their approach leverages environmental and climatic data, including geospatial data, weather conditions, and land cover parameters, to make predictions about fire spread. While their study focuses on large-scale predictions, our research narrows the scope to target more localized wildfire behavior, incorporating fine-grained environmental data, including wind speed and direction, temperature, and other geographical factors. Additionally, our study focuses on integrating a more diverse set of features such as population density and specific land cover classifications to enhance prediction accuracy, particularly in areas such as the Western United States.

The purpose of this study is therefore to build on methods such as machine learning and model development strategies proposed by these papers and to attempt to build similarly successful models using different data features, based on data from the United States, with the aim of impact in the Western United States.

Data Preprocessing

Data Preprocessing began with writing TFRecord (binary tensor files) files, from the original data source to CSV format. We then parsed through the features of the TFrecord files: elevation, wind direction (th), wind speed (vs), minimum temperature (tmmn), maximum temperature (tmmx), specific humidity (sph), precipitation (pr), drought index (pdsi), vegetation index (NDVI), population density, energy release component (erc), the fire mask at time t = 0 days (PrevFireMask), and the fire mask at time t = 1 days (FireMask), which were stored in the TFRecord Files in the shape [64,64] with type float32.

To incorporate spatial context into the dataset, we created a new column called fireNeighbors, which captures the number of neighboring cells in a 64x64 array that have a Fire value (1.0). For each cell, we calculated the total count of its adjacent cells—both directly and diagonally connected—that contained the Fire value using a convolutional 3x3 grid. This feature provides the model with information about local fire density, enhancing its ability to predict wildfire spread based on proximity to already ignited areas. This is related to the discussion about the importance of the spatial context in the wildfire prediction process, and the importance of the geographical spread in predictions.

The dataFinal data frame was then created, with columns th, vs, tmmn, tmmx, sph, pr, pdsi, NDVI, population, erc, fireNeighbors, and PrevFireMask. Then, the dataLabels dataframe, with only 1 column: FireMask. They were created by converting the data stored in the CSV files as strings to floats using casting from string to float (float(x)).

The FireMasks were then modified for experimental purposes, assuming that all unsure data (stored as -1.0), have no fire and will be stored as 0.0 instead.

The Train-Test Split served as a division of the dataset into training and testing sets for model training. The test_size variable was set to 0.2, meaning that 20% of the data will used to test, rather than train, the model. Setting random_state = 1 results in uniform shuffling of data. X_train, X_test, y_train, and y_test are the 4 dataframes formed after using the Scikit-Learn train_test_split method. X_train and y_train are the training datasets, and this data will be used to train the various models in recognizing patterns in the data. X_test and y_test are the test data and will be used to verify the accuracy of the model post-training. Logically, the training set must be much greater than the testing set, as the testing data must only be used a few times in the predictions based on learning from the training data.

The data was then normalized to have a mean of 0 to ensure uniform input. The dataFinal data frame was normalized using the formula, $z = \frac{x-\mu}{\sigma}$, where the normalized value, z, was calculated with: μ= mean of the given distribution, and σ = standard deviation of the given distribution. Each column (“elevation”, “th”, “vs”, …) was normalized individually in the X_train dataset first, then stored the mean and standard deviation (STD) for each column. The same mean and STD values from the X_train data were used to normalize the testing data, to ensure uniformity.

There was an extreme imbalance in the data, having unequal representations of negative and positive classes. In each 64x64 FireMask sample, there were typically thousands of negative (0.0) samples and few positive (1.0) samples. Most samples had a majority of the 64x64 grid that was negative, so the positives were very sparse in the FireMask. To address this data imbalance, the data reshaping process transformed each index in the original dataset, which previously contained 64 x 64 arrays, into individual float values, significantly increasing the number of data points. The new X_train and y_train datasets were balanced to contain an equal number of positive (1.0) and negative (0.0) values. This was done by identifying all indices in the original y_train where the value was 1.0, then adding each index from the 64 x 64 array individually to the new X_train and y_train. To maintain balance, an equal number of randomly selected indices where y_train was 0.0 were also added.

Table 1. Description of the Data Features (Independent Variables) used in the creation of the models and the prediction of Next-Day Wildfire Spread

Data Feature and Variable Used	Description
Elevation (elevation)	Terrain height affects fire behavior, as fires spread more easily uphill due to wind patterns and fuel availability.
Wind Direction (th)	Guides the potential spread of fire based on the direction the wind is blowing.
Wind Speed (vs)	Influences rate of fire spread, with stronger winds pushing fires to move faster.
Minimum Temperature (tmmn)	Lower temperatures can suppress fire activity, impacting daily fire intensity.
Maximum Temperature (tmmx)	High temperatures can exacerbate the effects of wildfires or even increase the intensity and spread of wildfires.
Specific Humidity (sph)	Affects the moisture content in the air, influencing how quickly vegetation dries out and ignites as well as the fire intensity.
Precipitation (pr)	Provides moisture to the environment, reducing the likelihood of fire ignition and spread.
Drought Index (pdsi)	Indicates long-term soil moisture deficits, with higher values signifying drier conditions favorable for fire spread
Vegetation Index (NDVI)	Measures vegetation health and density, with lower values indicating more dry and dead vegetation that fuels wildfires
Population density (population)	Can affect fire spread based on factors like the wildland-urban interface.
Energy Release Component (erc)	Represents potential energy available to fire, higher values indicate more intense fires.
fireNeighbors	Reflects the local fire density by counting neighboring cells with fire, informing the model of proximity-based fire spread.
Previous Fire Mask at time t = 0 days (PrevFireMask)	The initial fire presence, showing areas already burning or previously affected.
“Current” Fire Mask at time t = 1 days (PrevFireMask)	Captures the fire presence at time = t+1 days, serving as the target for any predictions from the model to be compared to.

Model Architecture, Training, and Evaluation

Base models K-Nearest Neighbors (KNN), Logistic Regression, and RandomForestClassifier were used to establish a foundation before moving on to more complex neural networks. KNN can give insights into the local structure of the data, while Logistic Regression provides a perspective on class separability. RandomForestClassifier, with its ensemble approach, captures non-linear patterns and helps in handling overfitting. Using these models first allowed the possibility of gauging the effectiveness of simpler algorithms and fine-tuning the feature engineering process, and served as a key baseline.

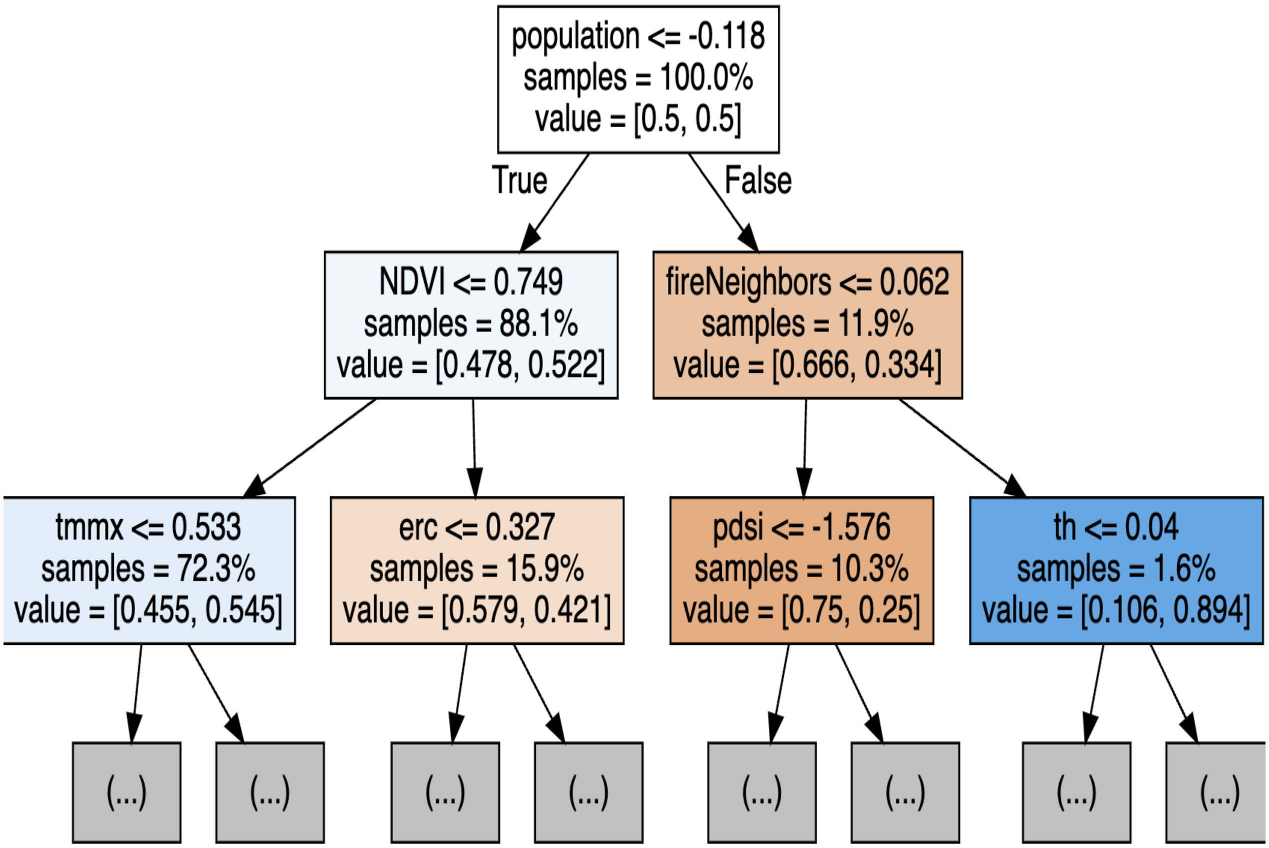

Figure 1. Architecture of the RandomForestClassifier model, a model that uses an ensemble, combining various decision trees to make a final prediction.

The neural network model was created using the Keras Sequential API. The Sequential model type allows layers to be stacked linearly, where each layer passes its output to the next. The first layer is an input layer with 13 neurons (each corresponding to one of the 13 input values), from which each of the data features will be taken into the model. This layer expects 13 input features, indicated by input_shape=(13,), meaning the model takes the 13 features from the dataset as input. A Dropout layer follows, where the dropout argument defines the fraction of neurons that are randomly "dropped" during training, to prevent overfitting by reducing reliance on specific neurons, forcing the model to learn more random representations of the data. The next layers added are Dense (fully connected) layers with 512 neurons and *ReLU* activation functions. The ReLU (Rectified Linear Unit) activation function introduces non-linearity to the model by outputting the input value if it is positive and zero otherwise, helping the network to learn complex patterns by activating only certain neurons. A loop “numTimes” determines how many of these Dense layers are added to the model, each with 512 neurons. After the loop, the final Dense layer outputs a single value, a prediction, which is followed by an activation function. This single output neuron is used for binary classification, where the activation function, sigmoid, produces a probability between 0 and 1 that any 1km x 1km area will have a fire using a logistic function based on the various interconnected neurons.

HIGH SCHOOL EDITION
Journal of Student Research

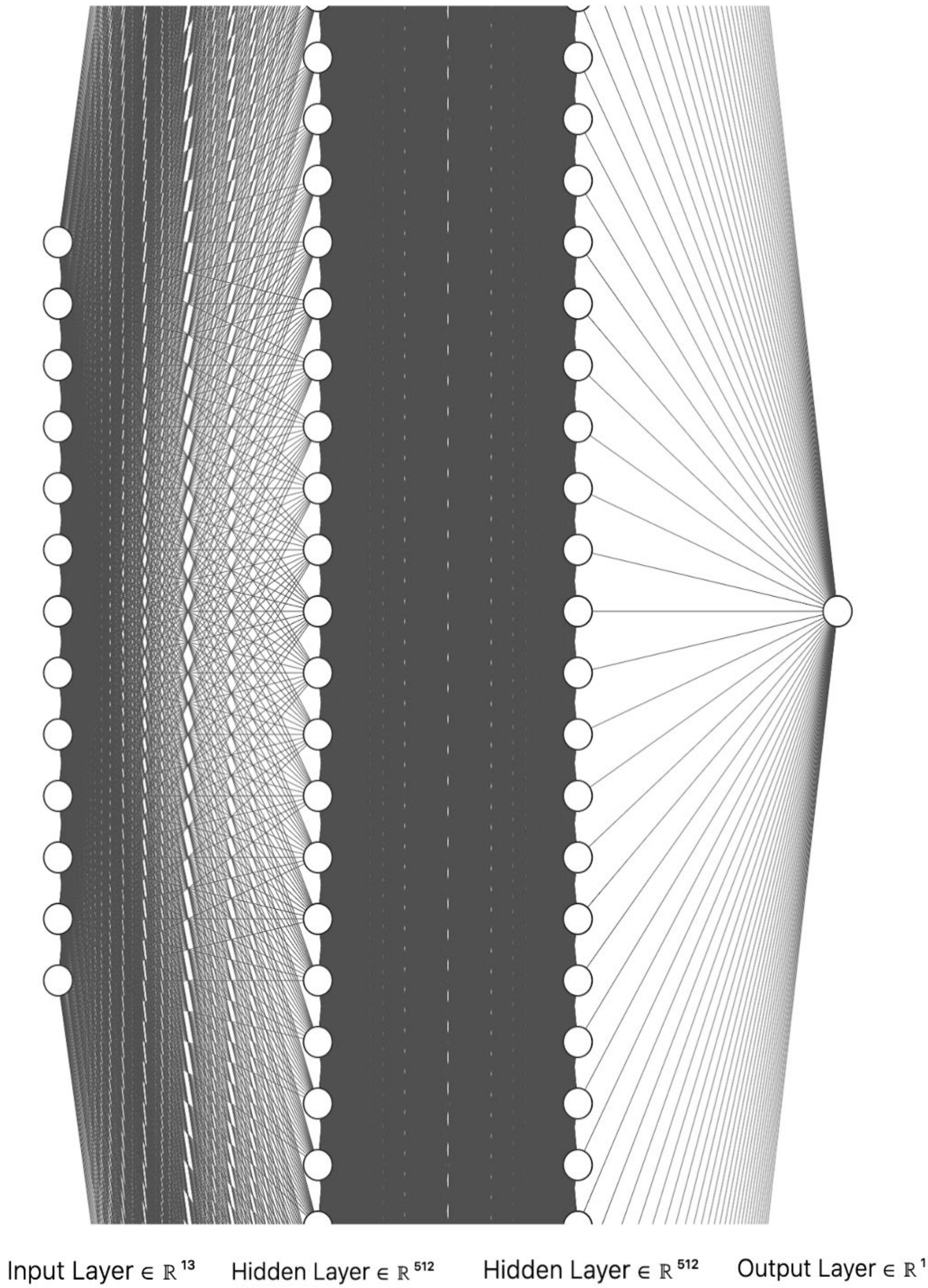

Figure 2. Architecture of the Neural Network model, a model that uses a series of interconnected neurons to make predictions, with dropout to ensure robustness and relieving the model of over-dependence on any specific neuron.

To improve the potential for high model accuracy further, we employed the common technique of "boosting," which involves the combination of the capacities of multiple models in order to maximize the accuracy. We combined the Neural Network and the RandomForestClassifier models, as well as the Neural Network and the Logistic Regression models, in order to complement each other and create an even greater model. For example, in the Neural Network-RandomForestClassifier boosting model, the Neural Network's capacity to capture complex patterns will complement the RandomForestClassifier's capacity to handle overfitting effectively. Boosting iteratively corrected prediction errors by focusing on misclassified instances from previous iterations, creating a weighted ensemble model based on multiple past attempts.

Predictions

The training process for the neural network involved the use of the Adam optimizer, binary cross-entropy as the loss function, and 20 epochs for model optimization. During training and further model development, various data augmentation techniques were applied to enhance the model's ability to generalize across different scenarios. Hyperparameter decision was an essential aspect of model optimization, with several hyperparameters being specified to create the best possible model. These included the number of epochs, the type and rate of dropout, the structure of dense layers, and other key parameters that could affect the model's accuracy and robustness.

Hyperparameters were not only involved in the neural network, but also in the base models. In the KNN model, they were involved in deciding the optimal number of neighbors to try and get the best predictions, which was 10, as decided after empirical testing. In the RandomForestClassifier, the number of decision trees in the ensemble was a hyperparameter, and as the number of trees increased, the prediction accuracy of the RandomForestClassifier increased. Therefore, 100 trees were used in the RandomForestClassifier model.

If the hypothesis is true, then we expect to see that the neural network has a greater capacity to predict wildfires than the base models, as the Neural Network has a more complex architecture that should theoretically allow it to capture intricate, high-dimensional patterns in the data that are critical for accurate wildfire prediction, at a higher level than the more simple base models. Additionally, we expect the boosted ensemble to outperform all individual models by leveraging the complementary strengths of the Neural Network and the RandomForestClassifier.

Results

To validate the neural network (NN) model, we compared its performance with several other models, including the logistic regression, K-nearest neighbors (KNN), and RandomForestClassifier base models.

Among the base models we tested, the RandomForestClassifier demonstrated the best performance. RandomForestClassifier, with its ensemble of decision trees, is particularly effective at handling both linear and non-linear patterns in the data. By averaging the predictions of multiple trees, RandomForestClassifier reduces the risk of overfitting and improves the model's ability to generalize to unseen data. In contrast, logistic regression, while useful for modeling linear relationships, was less effective at capturing the complex, non-linear interactions often present in wildfire prediction. K-nearest neighbors (KNN), which classifies data points based on their proximity to neighboring points, also struggled to handle the patterns needed for accurate wildfire prediction.

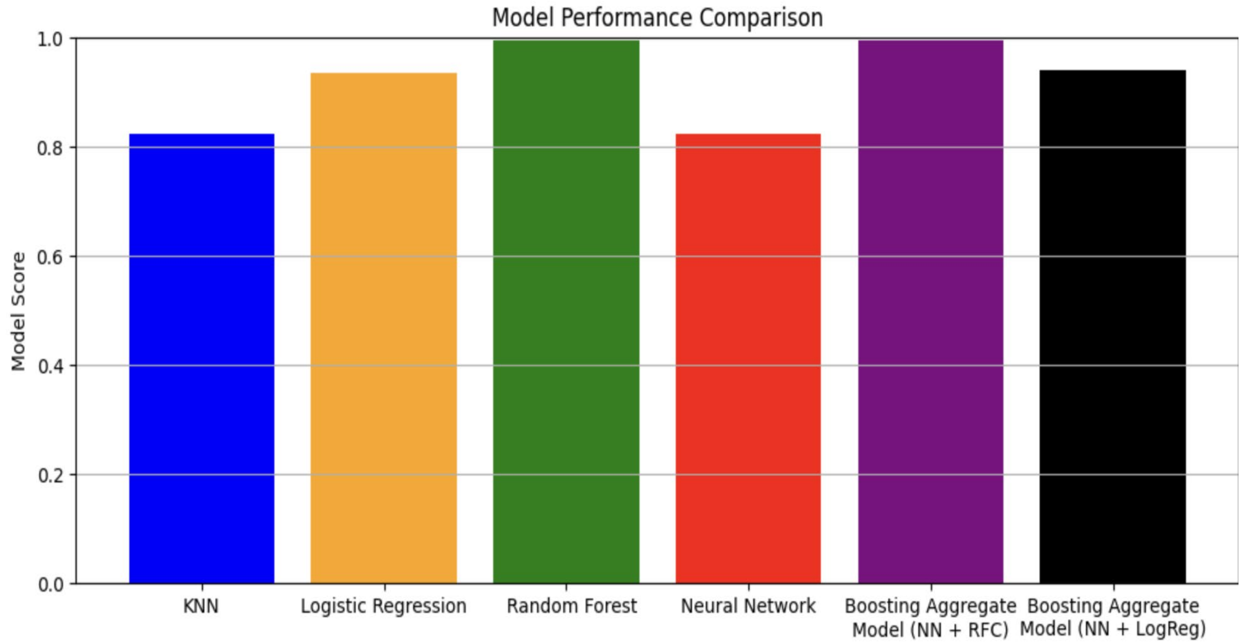

Figure 3. A comparison of the performance of each model, with a "score" between 0 and 1, indicating the accuracy of each model in predicting wildfires based on testing on the testing dataset.

To further improve model performance, we employed boosting, a technique that iteratively improves weak learners by focusing on the errors made in previous iterations. Boosting was applied to create an aggregate model combining the strengths of both the Neural Network and the RandomForestClassifier. The Neural Network, designed to learn complex patterns through multiple dense layers, complements the RandomForestClassifier's ability to handle non-linear patterns. By boosting these models together, we achieved significant improvements in prediction accuracy. This boosted ensemble model proved to be highly effective, outperforming the base models and delivering robust results, especially in capturing the nuances of wildfire spread prediction.

A key part of the wildfire-prediction process was the visualization, to see the way that the model predicts the spread relative to the actual value. This visualization compares wildfire prediction outputs against actual observations and prior fire activity, providing insight into the model's accuracy in predicting at-risk areas. Using an AxesGrid layout, it presents three panels: the Previous Fire Mask (fire-affected areas, at time t = 0 days), the Actual Fire Mask (the true observation at time t = 1 day), and the Predicted Fire Mask (the model's prediction). The model generates raw predictions, which are refined using a threshold—a critical parameter that suppresses low-risk areas by setting predictions below the threshold to zero, thereby only highlighting regions with significant fire risk. This approach ensures that only the most at-risk areas are emphasized, aiding in the evaluation of the model's utility for wildfire spread forecasting. Though the model does not discount predictions below 70% probability, many of those values with a probability around 50% indicate that the model is unsure about that specific region, and those predictions therefore do not hold any value in the ultimate visualization.

The visualization therefore is a 64 x 64 grid (4096 squares), each of which is assigned a certain color based on the probability of fire in that region, with darker shades of red corresponding to a higher probability of fire in that region. This creates a necessary visual of the actual shape of the fire, allowing for a deeper understanding of the different ways in which fire tends to spread, even on a next-day basis. The shades of red represent probabilities over the threshold, which we set at 0.7, while the lighter shades of tan correspond to any areas in which a lower probability of fire was predicted. As used throughout this process, each square in the visualization represents a 1 km x 1 km area, allowing visualization of a relatively large portion of land.

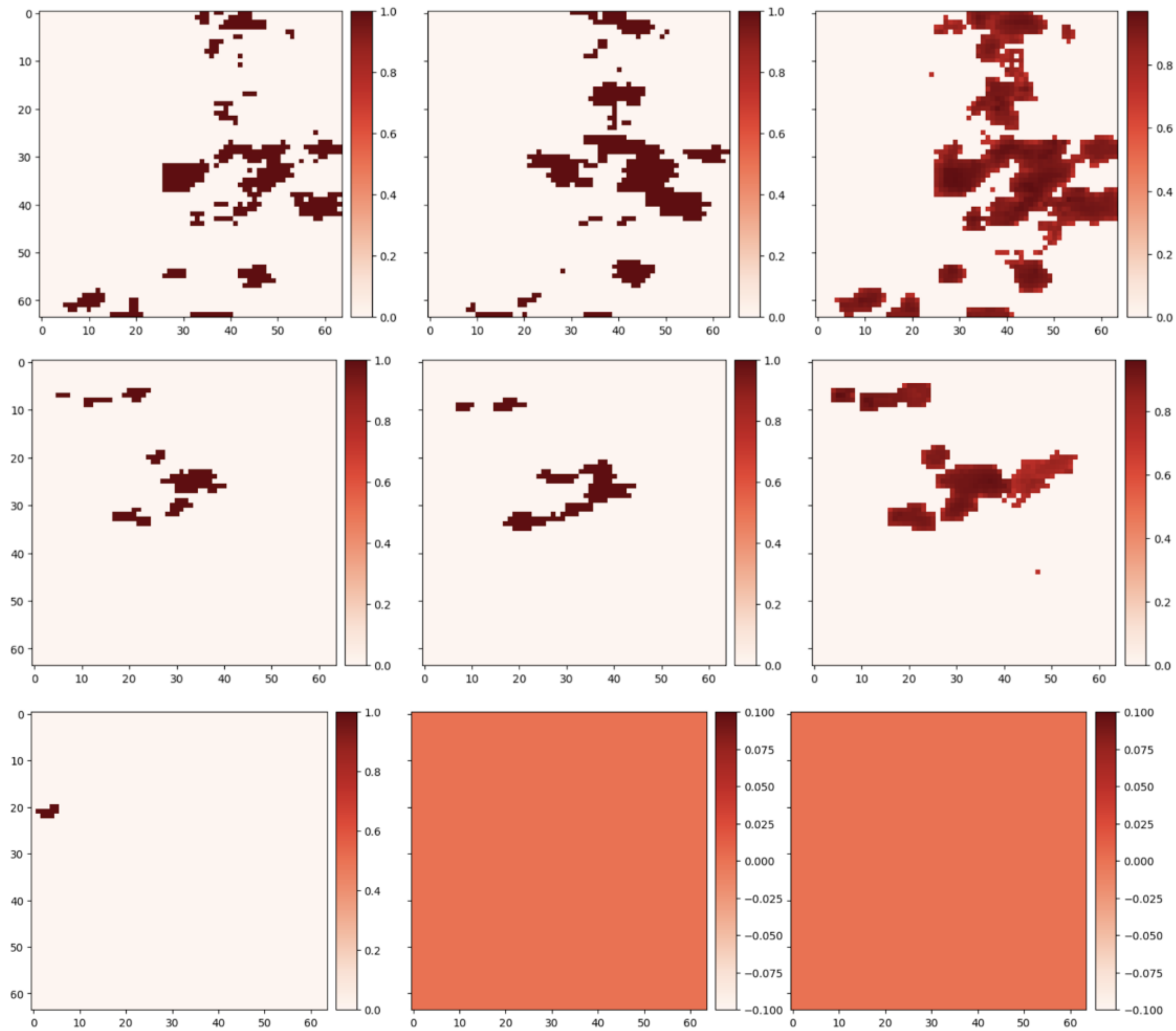

Figure 4. A visualization of fire spread with three different examples of the predicted spread in comparison to the actual values. In each row, from left to right, the first graph represents the *previous fire mask* (fire locations from time t=0 days), the second graph shows the *actual fire mask* (the true fire the next day), and the third graph illustrates the *predicted fire mask* (model's prediction for the current time). Darker red shades indicate higher probabilities or intensities of fire activity, and tan colors represent areas with probabilities lower than the threshold of 70%. Each graph represents a 64 km x 64 km area, containing 4096 1 km x 1 km areas, the minimum areas on which wildfires are predicted.

Discussion

The results indicate that the RandomForestClassifier performed best among the base models tested: RandomForestClassifier, K-Nearest Neighbors, and Logistic Regression. The RandomForestClassifier likely performed best because it used an ensemble of 100 decision trees, allowing it to capture complex relationships between the environmental features and wildfire spread. Each tree in the forest provides a unique perspective, reducing overfitting and increasing generalization. RandomForestClassifiers also handle nonlinear relationships and interactions between all 13 features, which are all critical in modeling wildfire behavior. This diversity in decision-making helps the model make more accurate predictions compared to simpler models like K-Nearest Neighbors and Logistic Regression. Ultimately, the most successful model overall was the model combining the RandomForestClassifier and the Neural Network, enabling generalization with decision trees along with complex analysis through deep learning with the Neural Network.

The success of the ensemble-based RandomForestClassifier, as well as the ensemble of the Neural Network and the RandomForestClassifier, shows the value of ensemble-based machine learning methodology. By combining two different models with vastly different capabilities, a far more powerful model can be generated. This is exemplified by these models, which had much higher accuracies than any other models, which can at least in part be accredited to their ensemble-based architectures. Although the hypothesis was proven wrong, as the RandomForestClassifier and Logistic Regression models were ultimately more accurate than the Neural Network alone, this study reveals that in terms of spatial wildfire predictions, aggregate models, either developed through bagging or those such as RandomForestClassifier than inherently use an ensemble-based approach in their creation, are the most effective in creating accurate predictions.

Another key observation from the predictions shown in *Figure 4* is that when comparing the predicted FireMasks to the actual FireMasks from time t = 1 day, the predicted FireMasks tended to be centered in the same areas as the actual FireMasks and hold approximately the same shape, however, it would also predict some extraneous areas around the extremely confident areas (with a high probability of a wildfire) with a lower probability, shown by lighter hues of red. The model's prediction of extraneous areas with lower probability leads to an approach that errs to the side of caution, allowing the mitigation of the risk of underpredicting wildfire spread, ensuring that resources are allocated to potentially affected areas, and although not predicted areas are fully accurate, the areas with the darkest hues are generally the most accurate when compared to the actual FireMasks. Although the extra predictions of fire may create some confusion, this is simultaneously beneficial when applied practically as this enables extra caution.

Additionally, based on observation and discussion, it is clear that the most important feature in the model was the PrevFireMask, which provided critical insight into the spatial distribution of fire activity at time t=0, serving as the foundation for predicting the fire mask at time t+1. This feature allowed the model to capture the shape and boundaries of active fires, making it significantly easier to forecast how these fires would evolve in the next time step. When the PrevFireMask was included, the model effectively leveraged this initial fire distribution to generate accurate 64x64 predictions. However, when this feature was eliminated, the model's performance declined sharply, as it lacked the direct spatial context needed to predict the future fire spread accurately, and the images predicted were completely inaccurate and dissimilar to the initial FireMasks. The absence of the PrevFireMask caused notable distortions in the predicted fire shape, underscoring its crucial role in influencing the final wildfire prediction. In addition to the PrevFireMask, the FireNeighbors feature played a crucial role in enhancing the model's predictions by providing context on the density of nearby fires. By counting the number of adjacent cells containing fire, fireNeighbors helped the model assess local fire behavior and its potential spread, complementing the spatial information from the PrevFireMask and further refining the shape of the final predictions. The model gave significant weight to PrevFireMask and fireNeighbors as they provided insight into what the shape of the 64km x 64km prediction should look like, while all other features only provided context for each 1 km x 1km region independently. Spatial data was the most important because it directly captured the dynamic nature of fire spread, allowing the model to predict how the fire's boundaries and shape would evolve. However, other variables must also be accredited, as they helped refine the model's understanding of each 1 km x 1 km region's local conditions and enabled the model to predict each specific area, as the spatial data only allowed the model to create a general image of what the prediction would look like.

HIGH SCHOOL EDITION
Journal of Student Research
Volume 14 Issue 1 (2025)

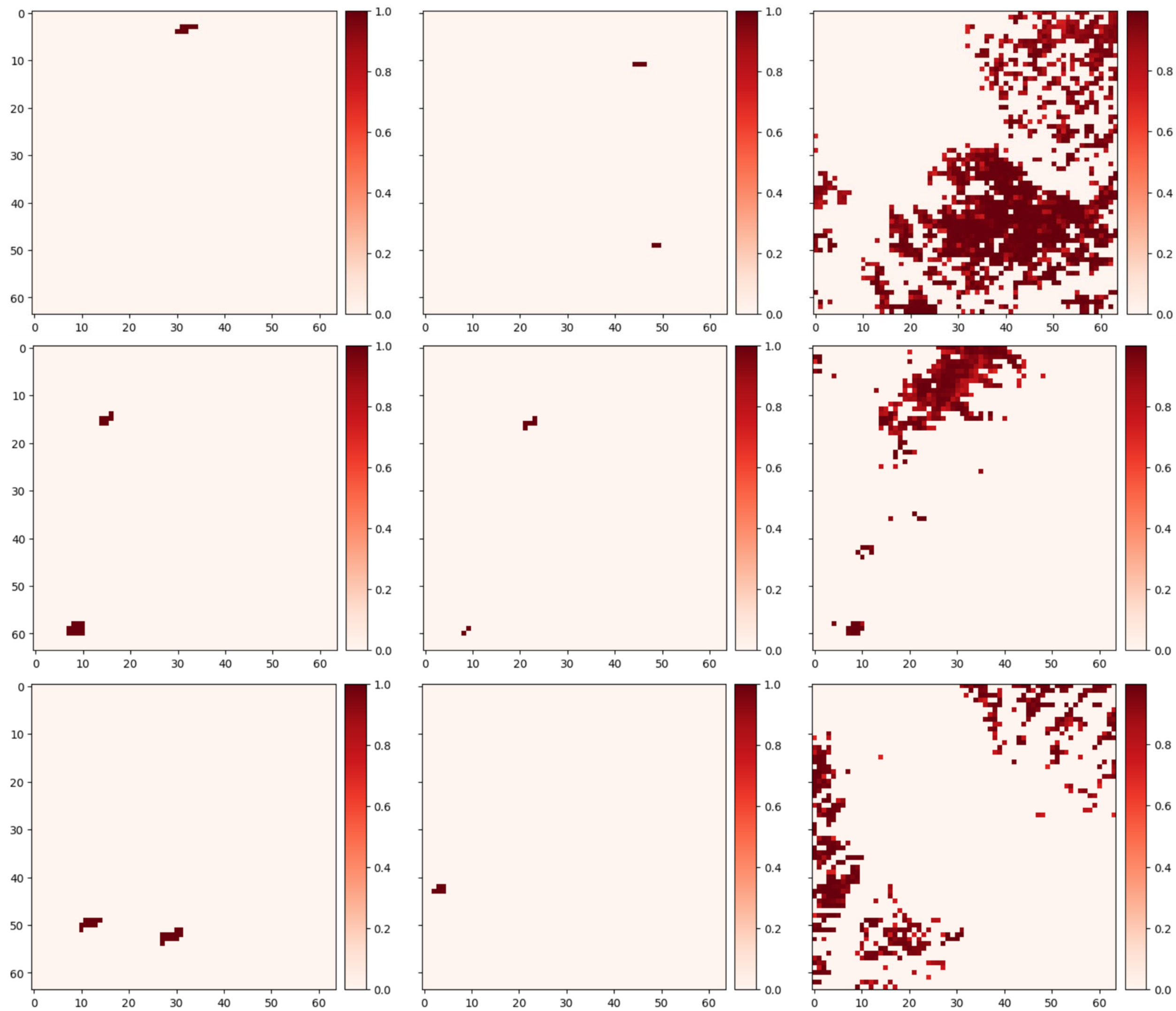

Figure 5. A visualization of fire spread without spatial data given from PrevFireMask and fireNeighbors. Shown are three different examples of the predicted spread in comparison to the actual values when the model is trained without the spatial data mostly used in making predictions. In each row, from left to right, the first graph represents the *previous fire mask* (fire locations from time t=0 days), the second graph shows the *actual fire mask* (the true fire the next day), and the third graph illustrates the *predicted fire mask* (model's prediction for the current time). Darker red shades indicate higher probabilities or intensities of fire activity, and tan colors represent areas with probabilities lower than the threshold of 70%. Each graph represents a 64 km x 64 km area, containing 4096 1 km x 1 km areas, the minimum areas on which wildfires are predicted.

We recognize that our methods have limitations. While our methods successfully demonstrated how models like the Random Forest Classifier and Neural Network can predict wildfire spread, there are both strengths and limitations. The Random Forest Classifier is particularly effective at managing both linear and non-linear patterns in the data, providing robust generalization capabilities. However, it may struggle with very complex interactions that a deeper Neural Network could capture more effectively. On the other hand, while the Neural Network can model intricate relationships, it requires careful tuning and a sufficient amount of data to prevent overfitting. Another limitation of our methods is the inability to effectively utilize precision and recall metrics, which specifically focus on true positive and false positive rates in the context of wildfire predictions. Instead, we relied on accuracy, which may not provide a comprehensive assessment of model performance, particularly in scenarios with imbalanced classes where the costs of false negatives can be significant.

There is potential to improve the accuracy of the Neural Network through the use of more advanced boosting techniques. By combining multiple Neural Networks into an ensemble using boosting methods, we could enable them to learn complex, high-dimensional patterns, enhancing overall model performance. This approach would allow the Neural Networks to complement each other, just as we had done with the RandomForestClassifier and the Logistic Regression models. Additionally, further hyperparameter tuning could be conducted to optimize factors such as the learning rate, batch size, number of hidden layers, and dropout rate. Including more data and tweaking other parts of each model could further increase their capacity to predict wildfires.

Our approach is significant as existing open-source models using the same dataset either do not leverage neural networks at all or fail to incorporate ensemble techniques like bagging, which significantly enhance predictive accuracy. Our approach is also critical because it compares a variety of models, including simpler ones like K-Nearest Neighbors and Logistic Regression, against more complex neural networks, allowing for a comprehensive evaluation of model effectiveness, and a deeper understanding of the capacity of various standard base models in wildfire prediction, to understand the necessary models and data to make accurate prediction. This creates a fair assessment of the accessibility of these predictive software, as it develops an understanding of the minimum complexity of models needed to make effective predictions. The inclusion of bagging within the RandomForestClassifier further reduces overfitting and increases robustness by combining multiple decision trees to make a final prediction, unlike other models that may rely on a single decision path. Additionally, my method addresses data imbalance through reshaping and balancing techniques, ensuring a more reliable prediction of fire spread across the dataset, which other open-source data analysis processes fail to do.

Conclusions

Looking ahead, these models hold promising potential for practical applications in wildfire prediction, especially in regions like the West Coast, where changing climate conditions may increase fire risks. The integrated approach of combining these models can facilitate real-time predictions by analyzing various environmental factors and historical fire data. By accurately forecasting wildfire spread, stakeholders can enhance preparedness and short-term response strategies, ultimately contributing to improved safety and resource management.

Wildfire prediction is a field with great potential, especially with further technological advancement and model development that can predict fires to an even higher degree of accuracy, using further parameters such as microclimates and the historical fire history of a region. Additionally, this specific project had data aimed towards the United States and therefore has the greatest implications on the continental United States, however, wildfires are a global issue, the solutions to which have the potential for international success and mediation.

The greater issue at hand, as discussed in the Introduction, are the threats to human life and global ecosystems caused by destructive wildfires. Therefore, this paper aims to have presented a variety of models and visualizations to further understand wildfire spread, and the models that can be used to predict such spread. Ultimately, this research and further wildfire modeling aims to create large-scale policy change to more effectively mitigate wildfires and save human lives and resources as much as possible.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

Burke, M., Driscoll, A., Heft-Neal, S., Xue, J., Burney, J., & Wara, M. (2021). The changing risk and burden of wildfire in the United States. Proceedings of the National Academy of Sciences of the United States of America, 118(2), e2011048118. https://doi.org/10.1073/pnas.2011048118

Aguilera, R., Luo, N., Basu, R., Wu, J., Clemesha, R., Gershunov, A., & Benmarhnia, T. (2023). A novel ensemble-based statistical approach to estimate daily wildfire-specific PM2.5 in California (2006-2020). Environment international, 171, 107719. https://doi.org/10.1016/j.envint.2022.107719

Sharma, S., Chandra, M. & Kota, S.H. Health Effects Associated with PM2.5: a Systematic Review. Curr Pollution Rep 6, 345–367 (2020). https://doi.org/10.1007/s40726-020-00155-3

Yu, A. C., Lopez Hernandez, H., Kim, A. H., Stapleton, L. M., Brand, R. J., Mellor, E. T., Bauer, C. P., McCurdy, G. D., Wolff, A. J., 3rd, Chan, D., Criddle, C. S., Acosta, J. D., & Appel, E. A. (2019). Wildfire prevention through prophylactic treatment of high-risk landscapes using viscoelastic retardant fluids. Proceedings of the National Academy of Sciences of the United States of America, 116(42), 20820–20827. https://doi.org/10.1073/pnas.1907855116

Pope, C.A., Ezzati, M., Cannon, J.B. et al. Mortality risk and PM2.5 air pollution in the USA: an analysis of a national prospective cohort. Air Qual Atmos Health 11, 245–252 (2018). https://doi.org/10.1007/s11869-017-0535-3

Liu, J. C., Wilson, A., Mickley, L. J., Dominici, F., Ebisu, K., Wang, Y., Sulprizio, M. P., Peng, R. D., Yue, X., Son, J. Y., Anderson, G. B., & Bell, M. L. (2017). Wildfire-specific Fine Particulate Matter and Risk of Hospital Admissions in Urban and Rural Counties. Epidemiology (Cambridge, Mass.), 28(1), 77–85. https://doi.org/10.1097/EDE.0000000000000556

McClure, C. D., & Jaffe, D. A. (2018). US particulate matter air quality improves except in wildfire-prone areas. Proceedings of the National Academy of Sciences of the United States of America, 115(31), 7901–7906. https://doi.org/10.1073/pnas.1804353115

Radeloff, V.C., Hammer, R.B., Stewart, S.I., Fried, J.S., Holcomb, S.S. and McKeefry, J.F. (2005), THE WILDLAND–URBAN INTERFACE IN THE UNITED STATES. Ecological Applications, 15: 799-805. https://doi.org/10.1890/04-1413

Modaresi Rad, A., Abatzoglou, J.T., Kreitler, J. et al. Human and infrastructure exposure to large wildfires in the United States. Nat Sustain 6, 1343–1351 (2023). https://doi.org/10.1038/s41893-023-01163-z

Radeloff, V. C., Helmers, D. P., Kramer, H. A., Mockrin, M. H., Alexandre, P. M., Bar-Massada, A., Butsic, V., Hawbaker, T. J., Martinuzzi, S., Syphard, A. D., & Stewart, S. I. (2018). Rapid growth of the US wildland-urban interface raises wildfire risk. Proceedings of the National Academy of Sciences of the United States of America, 115(13), 3314–3319. https://doi.org/10.1073/pnas.1718850115

Biswas, S., Hossain, M., & Zink, D. (2023). California wildfires, property damage, and mortgage repayment (Working Paper No. 2023-05). Federal Reserve Bank of Philadelphia. https://doi.org/10.21799/frbp.wp.2023.05

Jones, M. W., Smith, A. J. P., Betts, R., Canadell, J. G., Prentice, I. C., & Le Quéré, C. (2020). Climate change increases the risk of wildfires: January 2020. ScienceBrief. University of East Anglia. https://ueaeprints.uea.ac.uk/id/eprint/77982/

Sayad, Y., Mousannif, H., & Moatassime, H. (2019). Predictive modeling of wildfires: A new dataset and machine learning approach. Fire Safety Journal, 104, 130-146. https://doi.org/10.1016/j.firesaf.2019.01.006

Shadrin, D., Illarionova, S., Gubanov, F. et al. Wildfire spreading prediction using multimodal data and deep neural network approach. Sci Rep 14, 2606 (2024). https://doi.org/10.1038/s41598-024-52821-x