Prediction and Inference of Large Wildfire Burn Area in Contiguous United States.

2023 TAMIDS Data Science Competition

Team – Sam's Strikers

Shubham Jain, Sambandh Dhal, Krishna Chaitanya Gadepally, Prathik Vijaykumar

1. Executive Summary

Wildfires are a serious concern for the United States, causing significant damage to communities and natural resources. In this report, we examine the severity of large wildfires in the United States over the past decade (2011-2020). In order to identify potential wildfire hotspots, we examined the spatial variations in climate, topographic attributes, and land cover in various ecosystems. Our analysis shows that the severity of wildfire occurrences, as defined by the burn area, is highly correlated with the climatological forcings and geological characteristics of the ecosystem of its origin and can therefore be modeled for the prediction and inference of wildfire severity using data science tools.

To account for temporal and spatial variations in the dataset's climatological features, we used the location and burn area of large wildfires defined by a threshold of 1000 acres in the western US and greater than 500 acres in the eastern US in the 10-year period (2011 and 2020). Based on the preliminary data analysis, the states of California, Texas, and Idaho have experienced the highest number of large wildfires over the past decade, with lightning accounting for 43 percent of reported large wildfire incidents. Most of the burned area in these wildfires consists of grassland, forest, and shrub land covers, which could be used to determine the size of a fire. Therefore, we used the land cover classes surrounding a 4-kilometer buffer from the location of the wildfire as model features.

For further analysis, a predictive model using a 5-layered Deep Neural Network was trained on the climatological attributes, drought severity index, land cover and the vegetative indices from various satellite data sources at 4,536 large wildfire locations in the US over the last 10 years and the spread of wildfire was predicted. A SHAP analysis was conducted to visualize feature importance and partial dependence. The model revealed the highest importance of land cover around the vicinity of wildfire occurrence for prediction of total wildfire burn area.

Finally, we developed a mobile friendly interactive web-tool to visualize the wildfire burn area and related datasets for the past decade as an aid for data communication for end-users.

Supplementary mobile friendly visualization tool - https://shubhamjain.shinyapps.io/Wildfires/

2. Problem Statement

The 2023 TAMIDS data science competition is focused on two main challenges: predicting wildfire behavior and effective communication of research findings to decision-makers. Wildfire behavior is complex and difficult to predict accurately, making it challenging to develop models and forecasting tools. Effective communication of research findings to end-users such as land managers, policymakers, and the public is also a critical aspect of addressing the wildfire problem.

In 2021 alone, there were 58,733 wildfires in the US, burning a total of 7.1 million acres of land. The average number of wildfires per year over the past decade has been around 70,000, with an average of about 7 million acres burned per year. In Texas, wildfire activity varies from year to year, with 3,700 wildfires in 2021 burning approximately 200,000 acres of land. The devastating impact of wildfires on both human communities and the environment has led to a growing need for data-driven approaches to wildfire research.

In this report, we discuss data science-based approaches to analyze and address these challenges and provide actionable advice to decision-makers based on our analysis. We believe that through data-driven approaches, we can improve our ability to predict the spread of large wildfires depending on the region of occurrence, ultimately reducing their impact on human communities and natural resources.

3. Datasets

3.1. Study Area

The study area comprises the Contiguous United States (CONUS) split into 11 Level I Ecoregions and 967 Level IV sub-Ecoregions (Omernik, 1987). The western regions of the study area typically experience a higher number of wildfire incidents and larger burn area as compared to the western United states (Nagy et al., 2018). This can be attributed to heterogeneity in landscape caused by human development and fragmentation of forest land cover areas (Malamud et al., 2005). Spatial variations in climate forcings, topographic attributes and land cover can also contribute significantly to the occurrence and severity of wildfires across different regions of the United States (Liu et al., 2013). Therefore, effective management of wildfire incidents requires a deeper understanding of the natural and anthropogenic factors influencing the occurrence and severity of wildfires over a large spatial extent.

3.2. Wildfire data

The GIS data for wildfire incidents locations and burn areas boundaries for large wildfires in the US was obtained from the Monitoring Trends in Burn Severity (MBTS) program (Eidenshink et al., 2007) (https://data.fs.usda.gov/geodata/edw/datasets.php?xmlKeyword=Burn). The program assesses the frequency, extent, and magnitudes of all large wildland fires in the United States. The thresholds for large wildfires are set to greater than 1,000 acres in the western US and 500 acres in the Eastern US. A period of 10 years between 2011 and 2020 was selected for analysis and the "prescribed wildfires" were removed from the dataset. A total of 4,538 wildfire incidents were used in the analysis covering 87,305 square miles of burn area.

Additionally, the Spatial wildfire occurrence data for the United States, 1992-2015 dataset (Short, 2017) was used to analyze the information related to the cause of large wildfires (https://data.fs.usda.gov/geodata/edw/edw_resources/meta/S_USA.FPA_FOD_4thedition.xml).

The point locations for occurrence of large wildfires between 2011 and 2020 in the contiguous US were spatially visualized to identify the potential hotspots for wildfire occurrence as shown in Figure 1.

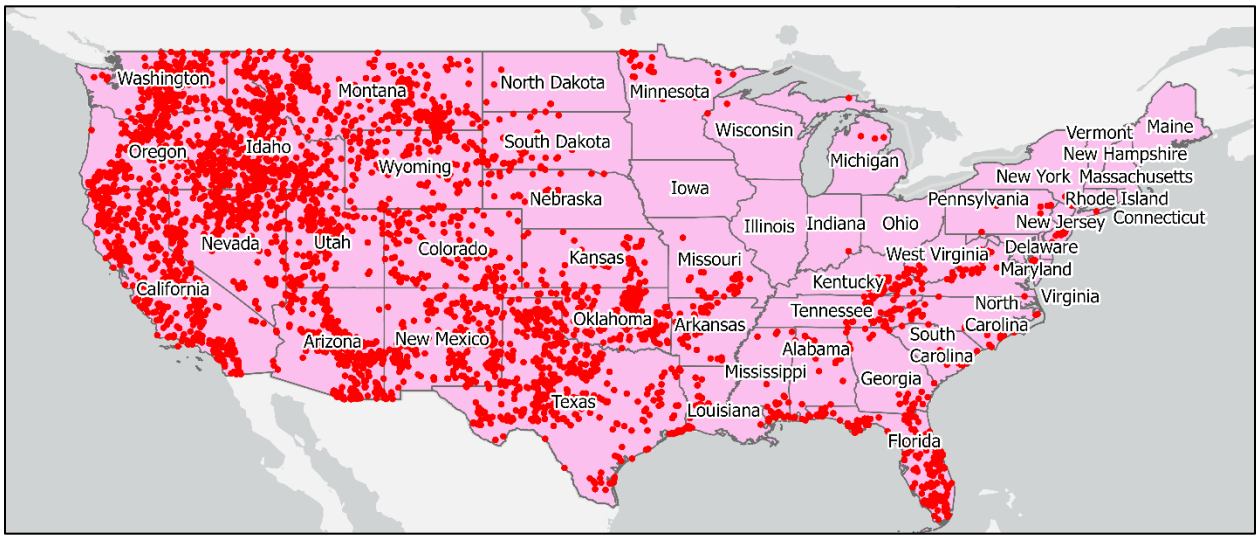

Figure 1. Large Wildfire incidents in the contiguous US between 2011 and 2020.

3.3. Meteorological and topography data

The meteorological datasets were obtained for the period of analysis and at wildfire occurrence locations from sources listed in Table 1. Monthly climate attributes including total monthly precipitation, mean monthly temperature, and maximum and minimum vapor pressure deficit were obtained from the PRISM dataset. The Palmer Drought Severity Index (PDSI) was obtained from GRIDMET and indicates the relative dryness in the region. The index generally ranges from -10 (dry) to +10 (wet) (Alley, 1984). The 4-kilometer raster grid values for the climate attributes corresponding to the location and time of wildfire occurrence were used as features in the model.

The land cover data was obtained from the National Land Cover Dataset (NLCD). The 30-meter NLCD raster for year 2016 was used to obtain land cover percentages around a 4-kilometer buffer at the point of wildfire occurrence. The 4-kilometer radius was selected based on the mean burn area of all wildfires in the dataset to represent the amount of forest and shrubland available near the fire area that could potentially increase the extent of wildfires.
The vegetation indices included Normalized Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI) obtained from the Moderate Resolution Imaging Spectroradiometer (MODIS) satellite dataset (MOD13A3). The dataset was obtained at monthly timescale and 1 kilometer resolution at the point of occurrence of wildfires.
Elevation data was obtained from the United States Geological Survey (USGS) Digital Elevation Model (DEM) dataset at 100-meter spatial resolution.
The datasets were spatially and temporally linked to each of the 4,538 wildfires that occurred in the contiguous US between 2011 and 2020 using R and ArcGIS.

Table 1. List of datasets used in the study to model burn area in large US wildfires.

Category	Dataset	Variables	Source	Resolution
Climate	PRISM	Precipitation, Temperature, Vapor Pressure Deficit (min, max)	https://prism.oregonstate.edu/	4000 m gridded, Monthly
	GRIDMET	PDSI, PET	https://www.climatologylab.org/gridmet.html	4000 m gridded, 5-day (PDSI), 1-day (PET)
Land Cover	National Land Cover Database (NLCD), 2016	Open Water, Developed, Barren, Forests, Shrub/Scrub, Hay/Pasture, Cultivated Crops, Wetlands	https://www.mrlc.gov/	30 m gridded
MODIS	MOD13A3 Version 6	NDVI, EVI	https://lpdaac.usgs.gov/products/mod13a3v006/	1000 m gridded, Monthly
Topography	USGS DEM	Elevation (m)	https://earthworks.stanford.edu/catalog/stanford-zz186ss2071	100 m gridded
Ecoregion Boundaries	US EPA Ecoregions	Level I and Level IV Ecoregions	https://www.epa.gov/eco-research/ecoregions	Shapefile

4. Data Exploration

The initial data analysis was performed to understand the spatial and temporal distribution of large wildfire incidents in the US. Large wildfires were determined by burn area greater than 500 acres in the Eastern United States and 1000 acres in the Western United States. The three states with the highest number of large wildfires between 2011-2020 occurred were California (448 incidents), followed by Texas (434 incidents), and Idaho (426 incidents) (Figure 2).

Figure 2 depicts the distribution of causes of large wildfires (2011-2015) obtained from the USDA wildfire dataset. About 43 percent of large wildfires were caused by lightning, followed by "Miscellaneous" (18%), unidentified (10%), arson (9%), equipment use (8%), and debris burning (6%). It is important to note that this data does not include small wildfires (500 acres) that are more frequently caused by human activities (Prestemon & Prestemon, 2013).

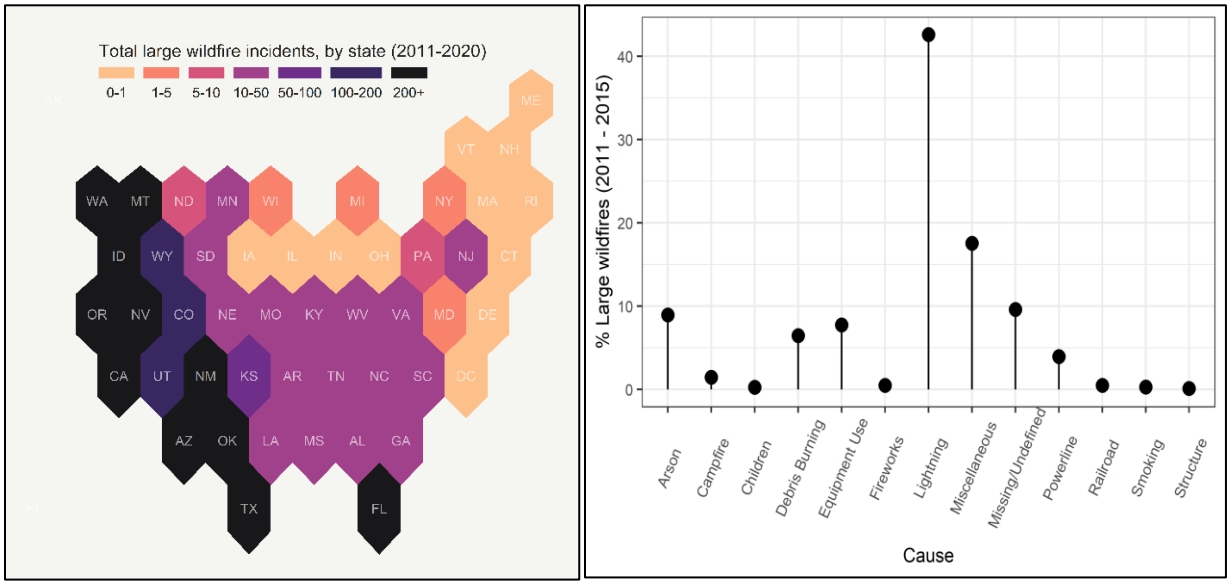

Figure 2. Left: Average annual large wildfire incidents by states in the US. Right: Cause of large wildfires (>500 acres) in the contiguous US between 2011-2015.

In order to identify potential hotspots for wildfires in the United States, the number of wildfire occurrences and burn area were also evaluated within each Level IV ecoregion. The percentage of burned area per ecoregion depicted in Figure 3 illustrates the severity of wildfires in various US ecosystems. The area consumed by wildfires was greatest in Mediterranean California, the Marine West Coast Forest, and North American Desserts, and smallest in Northern and Eastern Temperate Forests.

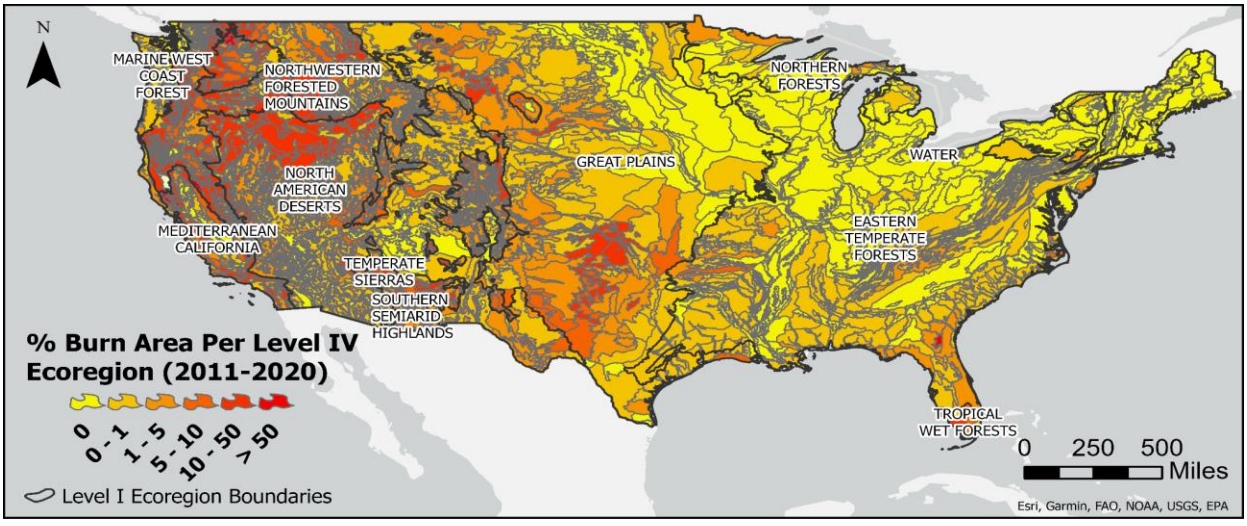

Figure 3. Percent of Level IV Ecoregion land burned in large wildfires between 2011-2020.

We investigated the relationship between land cover and wildfires by analyzing the NLCD land cover classes within the 4,538 burn area boundaries within the study period. Figure 4 demonstrates that most of the burn area consists of grassland, forest, and shrub/scrub land covers; therefore, including land cover as predictors could aid in determining the extent of a wildfire. A greater proportion of these land cover classes near the location of the wildfire would be positively correlated with the severity of the wildfire and, consequently, the burn area.

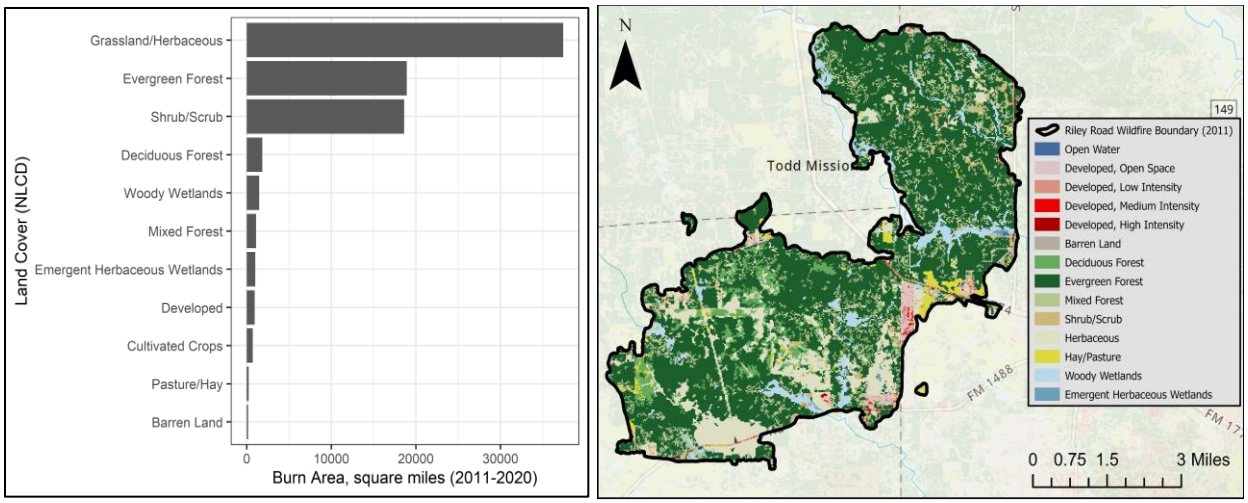

Figure 4. Left: Burn area by NLCD land cover in large wildfires between 2011-2020 in the Contiguous US. Right: Example of NLCD land cover within the burned area in the September 2011 Riley Road wildfire northwest of Houston burning 19,000 acres of land.

The occurrence and severity of wildfires are substantially influenced by climate characteristics. Higher temperatures and extended droughts are directly related to the occurrence of wildfires. Figure 5 depicts the number of large fire incidents in the United States and the average monthly temperatures, with the highest number of incidents occurring in July and August, which are typically the hottest and driest months. In order to model the burn area, it would be necessary to account for climate attributes such as precipitation, temperature, and drought index.

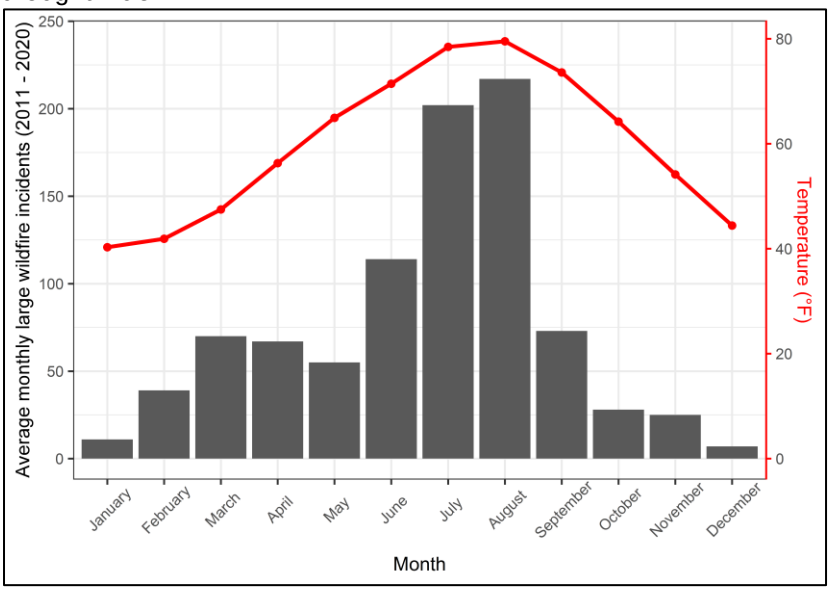

Figure 5. Plot showing the relationship between average monthly large wildfires (primary y-axis) in the contiguous US and the mean monthly temperature (line).

5. Methodology

The aim of this study is to predict the burn area of large wildfires in the US based on climatological and geological attributes surrounding the point of wildfire origin. Therefore, we used the attributes described in Section 3 as features in a Deep Neural Network model.

For data preparation, we used Pearson's correlation plots to identify highly correlated features (Figure 6), and then merged the various forest percentages to reduce the number of correlated model features. Similarly, all classes of developed land cover were merged, as were emergent and wooded wetlands. Also, wildfires with missing attributes were removed from the study resulting in a total of 4,536 observations for model development.

Prior to being used in the DNN model, the wildfire acres were log-transformed to account for the skewness in the observed data and normalize the distribution of the target. The architecture for the DNN model (Figure 6) is further described in Section 6.

The description of 17 features used in the model along with their minimum and maximum values in the dataset are shown in Table 2.

Table 2. Features used in the DNN model to predict large wildfire burn area with minimum and maximum values in the dataset.

Feature	Description	Min	Max
LATITUDE	Latitude coordinates of wildfire occurrence (decimal degrees)	25.2	49
LONGITUDE	Longitude coordinates of wildfire occurrence (decimal degrees)	-124.1	-72.8
DOY	Wildfire ignition day of year	1	365
ppt	Total monthly precipitation for month of wildfire ignition	0	1063.2
tmean	Average monthly temperature for month of wildfire ignition	-5.3	36.8
vpdmax	Maximum vapor pressure deficit for month of wildfire ignition	2.7	81.8
vpdmin	Minimum vapor pressure deficit for month of wildfire ignition	0	35.3
PDSI	Palmer Drought Severtiy Index during ignition date	-8.1	7.6
Developed	% NLCD developed around 4-kilometer buffer of wildfire ignition	0	64.2
Forests	% NLCD forests around 4-kilometer buffer of wildfire ignition	0	99.8
Shrub	% NLCD shrub/scrub around 4-kilometer buffer of wildfire ignition	0	100
grass	% NLCD grasslands/herbaceous around 4-kilometer buffer of wildfire ignition	0	100
Pasture	% NLCD hay/pasture around 4-kilometer buffer of wildfire ignition	0	74
Wetlands	% NLCD wetlands around 4-kilometer buffer of wildfire ignition	0	100
NDVI	Normalized Difference Vegetation Index for month of wildfire occurrence	0.1	0.9
EVI	Enhance Vegetation Index for month of wildfire occurrence	0	0.7
Elevation	Elevation of wildfire occurrence	-2	3507

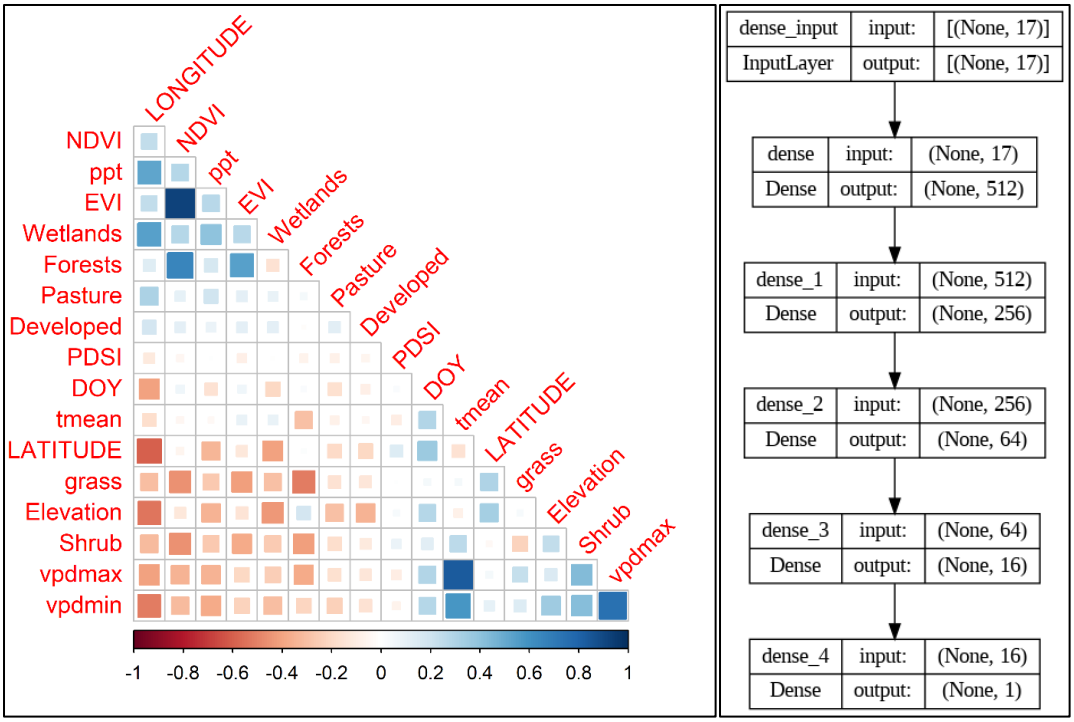

Figure 6. Left: Correlation plot for features used in the study. Right: Model architecture for DNN model used to predict wildfire burn area.

6. Modeling and Analysis

6.1. Deep Neural Network

We developed the Deep Neural Network (DNN) model using the keras and tensorflow libraries in python. The dataset was split into training and testing data using a 80/20 split with 80 percent of the data for training and validation and 20% for testing the accuracy of the DNN model. Further, the data was split three times to generate multiple random samples of training and test data to evaluate the accuracy over multiple test set combinations.

The features were scaled using a standard scaler and fed as inputs to a DNN model with five layers. The DNN layers had 512, 256, 64, 16 and 1 neurons respectively. ReLU was used as the activation function for each of the five DNN layers.

The DNN model was trained using root mean square optimizer and 0.001 learning rate. Callbacks were used to monitor validation loss. Mean squared error was the loss function used and performance metric was mean absolute error. The model was trained for 200 epochs with batch size of 32 and validation split of 0.2.

For each of the three values of random seed that was used for generating train and test sets, plots for training loss and validation loss were convex in nature.

The error rate for the test data was determined as shown below:

$$Error\ rate = \frac{\sum_1^N |\, y_{obs} - y_{pred} \,|}{\sum_1^N |y_{obs}|}$$

For test sets generated in each of the three values of random seed, the error rate was found to be between 0.055 and 0.006. The plots for model training and validation loss are shown in Figure 7.

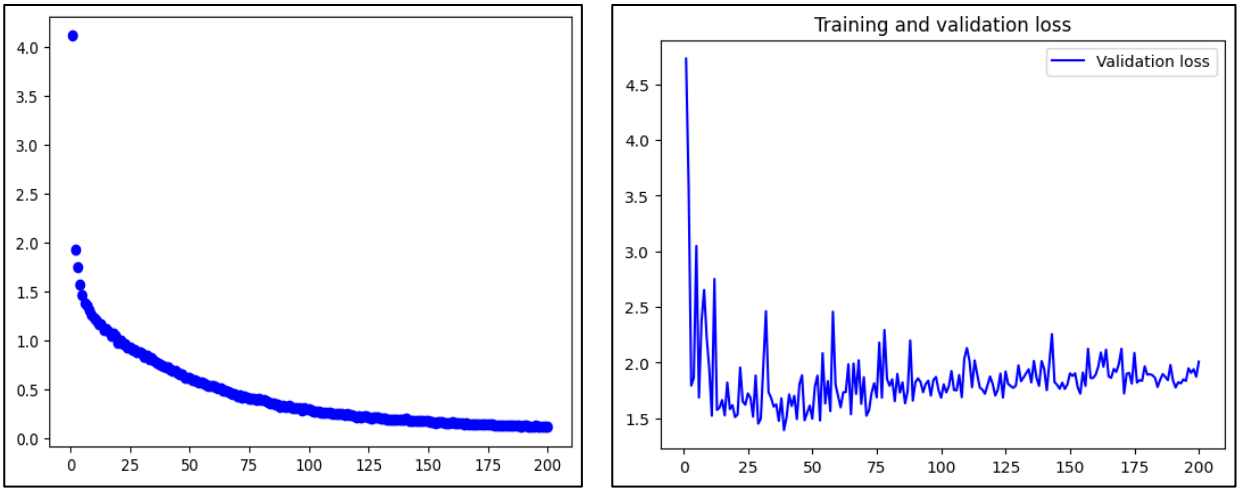

Figure 7. Left: Plot of model training loss. Right: Plot of model validation loss.

6.2. SHAP Analysis

SHAP (SHapley Additive exPlanations) is a surrogate (model independent) explanation method for ML models. It computes values that quantify the contribution of each feature to a prediction based on cooperative game theory (Lundberg & Lee, 2017). SHAP values can be used to determine the significance of a feature at both the global model scale and each local prediction scale. We use SHAP values to determine the relative importance of each model feature and partial dependence to evaluate the relationships between model features and burn area.

Simply, positive SHAP values indicate an increase in the model's prediction due to the feature, whereas negative SHAP values indicate a decrease in prediction. Model prediction equals the sum of all feature SHAP values and the average prediction.

7. Visualization and Interpretation

For efficient communication of the findings and analysis in this report, we used R shiny to develop a mobile friendly website that includes interactive leaflet maps and plots made during the competition time frame. The website can be accessed using the link – https://shubhamjain.shinyapps.io/Wildfires/. Further work on this web-tool could involve real time data visualization and regular updates to the tool to efficiently communicate the wildfire datasets in a user-friendly approach to the community.

The SHAP values were used to interpret the model and determine the potential drivers for predicting the burn area of large wildfires. As shown in Figure 8, the overall feature importance was determined by calculating the mean absolute sum of all SHAP values for each feature in the entire dataset. The land cover classes around 4-kilometer buffer at the point of occurrence including percentage of Grasslands/Herbaceous, percentage of Forests, and percentage of Shrublands were found to be the most influential in predicting wildfire burn area within a 4 km radius of the point of wildfire occurrence. The location of wildfires, as represented by latitude and longitude, was also important in predicting burn area, as our initial analysis revealed that the Northwestern United States had the largest burn areas.

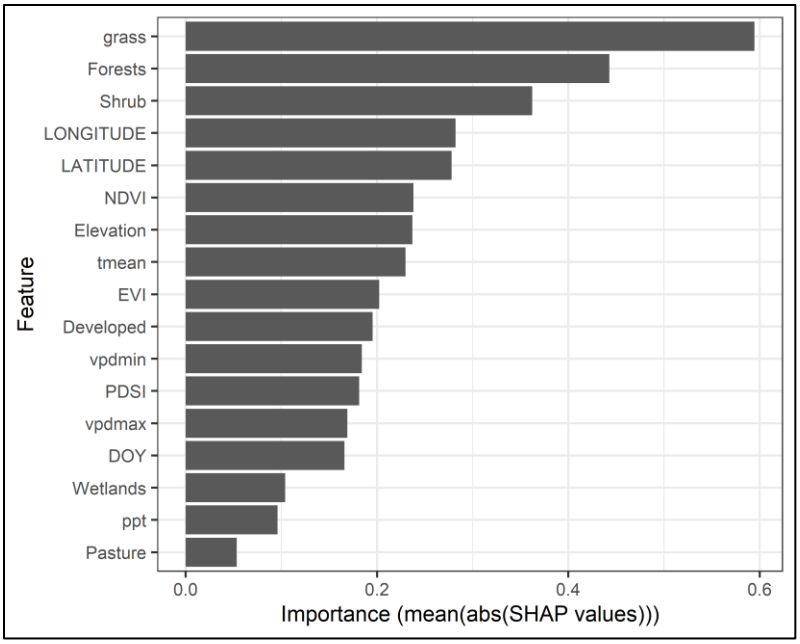

Figure 8. Feature importance in the DNN model obtained from SHAP values.

Figure 9 illustrates the non-linear relationships between features and the predicted burn area using partial dependence plots derived from SHAP values. An increase in forest area within a 4-kilometer buffer zone surrounding the point of wildfire occurrence was correlated positively with the burn area. A forest cover of 30 percent or more increased the predicted burn area above the mean. Similarly, extremely high elevation values had positive SHAP values, indicating that regions with a high elevation had a larger burn area.

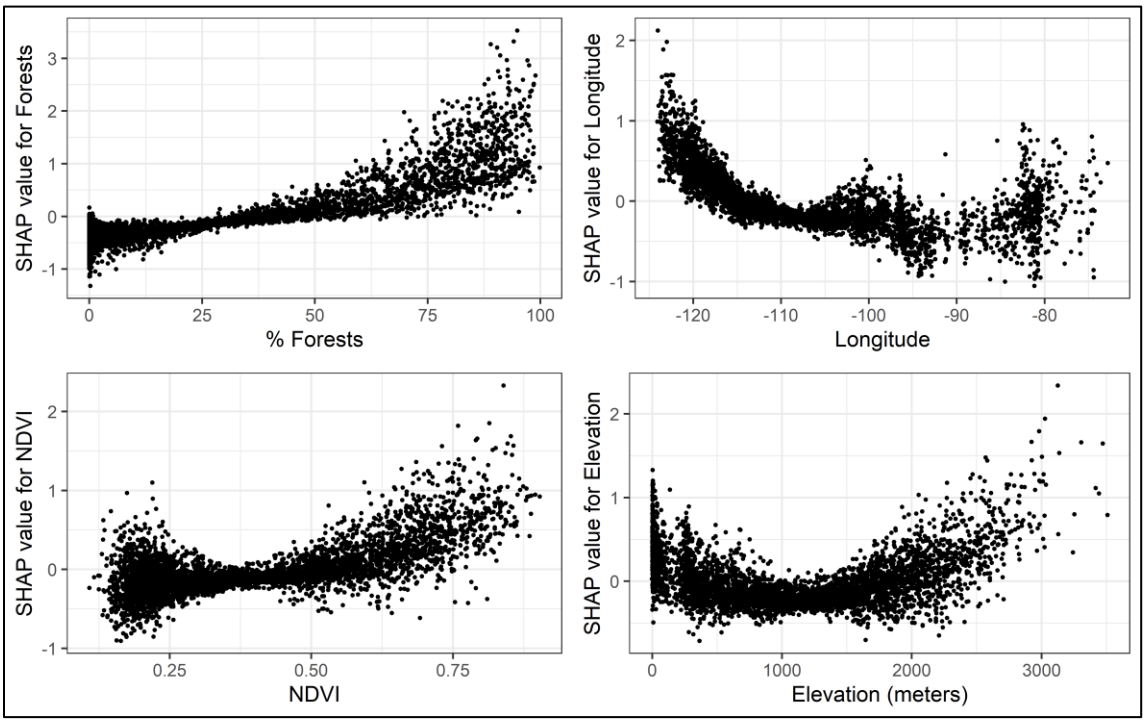

Figure 9. Partial Dependence Plots show the interactions between features and burn area using SHAP values.

8. Conclusions and Recommendations

In this study, the burn area of large wildfires across the contiguous United States from 2011 to 2020 was analyzed using a predictive Deep Neural Network model with climatological and geological features as model covariates. We found that land cover and the location of wildfire occurrence were the most influential factors in determining the severity and extent of wildfires in the United States. Using the predictive model, the extent of future wildfires can be predicted based on climate and land cover changes. Future work in this analysis could involve using datasets for all extents of wildfire occurrences to determine the probability of wildfire occurrence and predict the extents of all wildfires. In addition, a variety of characteristics, such as soil characteristics and 100-hour fuel moisture, could be utilized to improve the accuracy of the DNN model. Overall, our data science project provides important insights into wildfires and the need for a multidisciplinary approach to tackle this issue. Our findings highlight the importance of data collection, data modelling, and on-ground action as key areas to work on in order to develop an effective strategy to mitigate the impact of forest fires. We believe that to scale this project on a practical level, it is essential to improve the quantity and quality of data by collecting data on climate, drought index, land cover, forest cover, and topography using both in-situ and remote methods at higher resolution.

The data modelling team has the responsibility to make the insights easily readable and user-friendly for on-ground personnel, ideally by creating a suitable mechanism such as an app that can be used on smartphones. On-ground personnel require additional training to interpret the data and communicate their insights effectively. Moreover, a two-way channel between these departments is necessary to incorporate new reasons for forest fires as identified by on-ground personnel and incorporate novel ways to collect these new features by the data collection department. By strengthening interdisciplinary departments and working with minimal friction between them, we can effectively resolve the issue of wildfires and mitigate their impact on society.

9. References

Alley, W. M. (1984). The Palmer drought severity index: limitations and assumptions. *Journal of Applied Meteorology and Climatology*, *23*(7), 1100–1109. https://doi.org/10.1175/1520-0450(1984)023<1100:TPDSIL>2.0.CO;2

Eidenshink, J., Schwind, B., Brewer, K., Zhu, Z.-L., Quayle, B., & Howard, S. (2007). A project for monitoring trends in burn severity. *Fire Ecology*, *3*(1), 3–21. https://doi.org/10.4996/fireecology.0301003

Liu, Y., Goodrick, S. L., & Stanturf, J. A. (2013). Future US wildfire potential trends projected using a dynamically downscaled climate change scenario. *Forest Ecology and Management*, *294*, 120–135. https://doi.org/10.1016/j.foreco.2012.06.049

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*.

Malamud, B. D., Millington, J. D. A., & Perry, G. L. W. (2005). Characterizing wildfire regimes in the United States. *Proceedings of the National Academy of Sciences*, *102*(13), 4694–4699. https://doi.org/10.1073/pnas.0500880102

Nagy, R. C., Fusco, E., Bradley, B., Abatzoglou, J. T., & Balch, J. (2018). Human-related ignitions increase the number of large wildfires across US ecoregions. *Fire*, *1*(1), 4. https://doi.org/10.3390/fire1010004

Omernik, J. M. (1987). Ecoregions of the conterminous United States. *Annals of the Association of American Geographers*, *77*(1), 118–125. https://doi.org/10.1111/j.1467-8306.1987.tb00149.x

Prestemon, J. P., & Prestemon, J. P. (2013). *Wildfire ignitions: a review of the science and recommendations for empirical modeling*. US Department of Agriculture, Forest Service, Southern Research Station ….

Short, K. C. (2017). *Spatial wildfire occurrence data for the United States, 1992-2015 [FPA_FOD_20170508]*