Uncover Risk Drivers for Wildfire Occurrence with Machine Learning Approach

William Liu
Princeton High School
Princeton, NJ, United States
wliu2princeton@gmail.com

Mentor: Dr. Robert Rouse
Assistant Professor, University of Cambridge
Cambridge, United Kingdom

***Abstract*—Wildfire prediction plays an important role in ecosystem restoration, hazards prevention, and environmental biodiversity enhancement. Accurate predictions rely on the uncovering of the drivers behind fire events, such as interactions of climate, vegetation, and land cover variables. In this work we adopted a data driven machine learning approach to uncover the impact on wildfire probability prediction of different risk drivers. We downloaded MODIX 6.1 daily active fire data from NASA FIRMs, daily weather data and vegetation coverage data from re-analysis product of ERA-5 Land, and gridded daily Normalized Difference Vegetation Index (NDVI) from the NOAA (National Centers for Environmental Information).**

In this research, we applied two tree-based classification algorithms to predict the fire danger for the region centered around the border between British Columbia, Canada and Oregon, US, an area that has seen several large scale wildfires in year 2018 to 2021. Random Forest and Extreme Gradient Boosting are two main tree-based supervised algorithms to solve classification problems. We compared the accuracy and F1-score for the Random Forest model v.s. the XGBoost model. XGBoost slightly outperforms Random Forest in terms of F1-score and accuracy measure on test data. Further, we used ROC (receiver operating characteristic) curve to compare the false positive rate with the true positive rate for a model performance. XGBoost also outperforms Random Forest in terms of ROC-AUC measure.

Besides the climate, the vegetation cover data, and the active fire data, we calculated three new variables that can be derived from the raw climate data and the satellite based NDVI index, i.e., total precipitation in the previous month, average temperature in the previous month, and the closest distance to the developed land. We compute the feature importance scores and the SHAP values to identify the marginal contribution of each feature. Sensitivity analysis was conducted to study how the model performance of Random Forest or XGBoost is affected by the most important feature.

Major fire events are commonly driven by either low relative humidity or high wind speed. Our results confirmed the significant marginal contributions from the surface pressure and the relative humidity. In addition, our model identified the proxy to human behavior, the distance to the developed land as the risk driver with potentially the most impact on wildfire predictions.

Index Terms*— *wild ire, ire prediction, Machine Learning, ire risk driver

I. Introduction

Wildfires are increasingly influenced by climate change and human behaviors [1]. Climate change exacerbates the occurrence of extreme droughts and heatwaves, increasing the frequency and intensity of large wildfires across the globe. Forecasting wildfire danger and uncovering the drivers behind fire events become central for understanding relevant climate-land surface feedback.

Understanding and predicting wildfire spread is critical due to its potentially devastating effects on the environment, human life, and property [2]. The impacts of wildfires extend beyond immediate destruction, influencing economic, ecological, and social spheres [3]. The exacerbation of wildfire events is closely linked to more frequent and severe weather events, prolonged droughts, changes in vegetation patterns, and increased fuel loads. Additionally, shifts in wind patterns and extreme temperature have made fire events more unpredictable [4].

The challenges of forecasting wildfire events are rooted in their intrinsic stochastic nature and complex

interplays of different categories of risk drivers, such as climate, weather condition, vegetation coverage, and land cover type. Traditionally, the next day's fire danger is predicted by indices like the Fire Weather Index (FWI) [5], which relies only on climate conditions and disregards other risk drivers related to vegetation coverage and human behavior factors [6]. Machine learning approaches for modeling wildfire risk are facilitated by increased availability of satellite remote sensing data [7].

Researchers explored several approaches for leveraging machine learning in wildfire predictions. Khanmohammadi et al. [8] explored multiple machine learning models on grassland fires.They collected a dataset of grassland fires in Australia combined with meteorological variables and fire behavior metrics and showed the effectiveness of linear support vector machines, exponential Gaussian process regression, boosted trees, and neural network models. Rubi et al. [9] led a study to predict the spread and behavior of wildfires in Brazil. They used Brazilian Federal District's open data with variables such as climate, vegetation, hydrographic, and other fire risk factors, to test the accuracy of multiple machine learning approaches: artificial neural network (ANN), support vector machines (SVMs), random forest, and AdaBoost. Additionally, Reichstein et al. [10] suggested Deep Learning methodology for modeling the complex variable interplays for wildfires and other Earth system problems.

In this research, we applied data driven Machine Learning Methods to predict the fire danger for the region centered around the border between British Columbia, Canada and Oregon, US, an area that has seen several large scale wildfires in years 2018 to 2021. We downloaded daily active fire data, from MODIS (Moderate Resolution Imaging Spectroradiometer) Collection 6.1 for years 2015 to 2020 provided by NASA FIRMs (Fire Information for Resource Management System), for the region that centers around the border of British Columbia, Canada and Oregon, US. The important features such as daily weather data and vegetation coverage data are downloaded from re-analysis product of ERA-5 Land [11] provided by the climate store of ECMWF (the European Center for Medium-Range Weather Forecasts). For land cover type the gridded daily Normalized Difference Vegetation Index (NDVI) is downloaded from the NOAA (National Centers for Environmental Information).

We uncover the major risk drivers by investigating feature importance and SHAP values for feature contribution to fire danger predictions. The SHAP value computes the marginal contribution of each feature. Positive SHAP means a contribution to positive (higher) fire probability, while negative SHAP means a contribution to negative (lower) fire risk. The major contribution of this research is the creation of three new variables that can be derived from the raw climate data and the satellite based NDVI index. ALL three calculated variables have high contributions to fire danger predictions, with the closest distance to developed land having the highest feature importance score. Sensitivity analysis on the impact of the most influential feature was studied to test the robustness of Random Forest or XGBoost predictions.

The remainder of this paper is organized as follows.

Section II reviews related studies with different machine learning methodologies. Section III reviews the data source and the process of cleaning and transforming input features. Section IV details the proposed machine learning methods and related metrics for model performance. Section VI presents the discussion on feature selection, feature importance, and SHAP values of marginal contributions by input features. Finally, Section VII concludes with key findings and future directions.

II. Related Work

Machine learning and deep learning techniques have emerged as novel alternatives to conventional simulator or index based wildfire modeling approaches, such as FARSITE [12], FWI (Fire Weather Index) [5]. ML algorithms automatically learn patterns from spatial and temporal dataset, improving the accuracy of predictions and supporting proactive wildfire management [13]. Machine Learning (ML) methods are promising in modeling through the complexity in a data-driven way. The use of ML is facilitated by the increasing amount of data related to fire drivers, especially remote sensing products [7]. Several supervised ML approaches were studied

by Jain et. al. (2020) [14] in wildfire-related research. Supervised learning involves training models on datasets with an explicitly defined target variable such as a binary indicator of fire occurrence or not. Logistic regression is a commonly used and effective supervised learning technique for binary classification tasks. The purpose of logistic regression (LR) in the context of fire probability modeling is to identify the significant set of independent variables such as meteorological conditions to predict the probability of fire occurrence. [15]. The study conducted by Zhang et. al. (2016) [16] used the logistic regression model taking land cover, topographic data, vegetation indices, and socio-economic variables as input features to predict wildfire occurrences in the Southeastern Australia region. Another study by Price et al. (2015) investigated the spread of wildfires to the urban area of Sydney, Australia by utilizing a Binomial regression model, achieving a 98% predictive accuracy on test data. Overall, the LR model has been proven to be effective in predicting fire events, but with a potential weakness of overfitting. As logistic regression is parametric model, it is less robust than decision trees and ensembled learning algorithms based on trees, given the non-linear relationship between fire probabilities and predictors.

The Decision tree classifier can intuitively show the rules or criteria that make a target class highly likely. This learning technique is a powerful supervised learning algorithm that was developed by Quinlan (1986) [17] and further improved by Breiman et. al. (2017) [18]. Decision tree algorithm allows for predicting, explaining, and classifying outcomes in a tree structure. It is also used as base learners for ensemble algorithms that combine multiple decision trees to improve predictions. This ensemble learning can be performed using Boosting, Bagging, and Random Forest algorithms [19]. Random Forest and Extreme Gradient Boosting (XGBoost) are two major tree-based ensemble learning algorithms to solve classification problems. Boosting decision tree works by iteratively reweighting training data and focusing on the misclassified data. Random Forest works by training multiple decision trees on different subsets of training data with a random selection of features for each tree.

One of the main drawbacks of decision tree learning is its susceptibility to overfitting when a tree is excessively complex so that it fits well on training data but can not generalize to make an accurate prediction on new data. A study by Sulova and Jokar Arsanjani (2020) [20] investigated the performance of various ML algorithms, including Random Forest and Classification and Regression Tree (CART). The various ML algorithms could predict Wildfire occurrences with an accuracy of over 90%. Another study by Al-Bashiti and Naser (2022) [21] provides an in-depth look at the severity of wildfires using decision tree learning and ML techniques. Both Deep Learning and decision tree algorithm have been found to be successful in accurately predicting wildfire occurrences. The decision tree algorithm is still subject to overfitting and bias that can be addressed by boosting or ensemble learning.

The Random Forest (RF) model has emerged as a robust ensemble learning algorithm that presents high accuracy in classification problems such as wildfires prediction. Breiman (2001) [22] developed an RF model that combined multiple decision trees during the training phase and aggregated their outputs to improve accuracy while controlling overfitting bias. A recent study by Dahan et al. (2024) [23] employed the RF model to predict the impact of climate change on wildfire spread. By utilizing climate variables such as precipitation, temperature, relative humidity, wind speed, and soil moisture, the RF model showed high accuracy in predicting future wildfire activity under different climate scenarios. The construction of the RF model involves creating multiple decision trees using bootstrap samples of the training data, with each tree built using a random subset of features. This approach enhances the model's robustness and reduces overfitting. One of RF model's strength is its ability to provide insights into feature importance. Analyzing the contribution of each feature helps to identify the most influential factors driving wildfire occurrences. However, the RF model's performance still depends on the quality and quantity of input data, especially in regions with limited climate data. In general, the RF model integrates multiple climate variables and provides insight into the importance of features, offering high

accuracy and robustness in wildfire predictions.

Gradient boosting machines (GBM) are powerful supervised learning algorithms that iteratively build models by combining weak learners to minimize errors [24]. This technique is particularly useful for addressing imbalanced datasets and capturing complex nonlinear relationships in wildfire data. Extreme Gradient Boosting (XGBoost) is an optimized version of GBM with improved performance through regularization techniques. To predict the spread of wildfires, Singh et. al. [25] tested different machine learning algorithms, including decision tree regression, XGBoost regression and artificial neural network (ANN). They trained the models on the Next Day Wildfire Spread dataset that includes satellite images, weather, and geography conditions aggregated across the United States from 2012 to 2020, with elevation data, wind direction and velocity, minimum and maximum temperatures, humidity and precipitation, drought index, vegetation type, population density, energy release component, and previous fire events.

III. Data Description

The active fire data from MODIS (Moderate Resolution Imaging Spectroradiometer) Collection 6.1 between 2015 and 2020 provided by NASA FIRMs (Fire Information for Resource Management System) gathers the time, spatial location of daily fire occurrences. We filtered out fire event type 0 to get rid of spurious fire detection other than wild fire events. The selected study area with latitude from 42 to 55 and longitude from -117 to -130 centers around the border of British Columbia, Canada and Oregon, US. This region features dense forests, mountains, and increasingly prolonged summer droughts. In 2018 BC's record-breaking season saw over 1.35 million hectares burned. Smoke blanketed the entire Pacific Northwest.

The important features for the prediction of wildfires include weather data, vegetation coverage, and land cover type represented by Normalized Difference Vegetation Index (NDVI). We downloaded daily weather data and vegetation coverage data from re-analysis product of ERA-5 Land [11] from the climate store of ECMWF (the European Center for Medium-Range Weather Forecasts). To align the fire data with the ERA-5 Land data (with a gridded resolution of latitude or longitude of 0.25, or 31 km in distance), we calculated the average input variable values for the four grid points around each active fire location. For the time period of 2015 to 2020 in the region we selected, there are 147222 daily fire event spatial-temporal data. For land cover type we downloaded gridded daily Normalized Difference Vegetation Index (NDVI) derived from the NOAA (National Centers for Environmental Information) Climate Data Record (CDR) of Visible Infrared Imaging Radiometer Suite (VIIRS) Surface Reflectance. This data is projected to 0.05 latitude and longitude resolution grid and spans from 2014 to 10 days before the present using data from NOAA polar orbiting satellites. We have converted the data from its netCDF-4 file format to daily .csv files that contain NDVI values for locations represented by latitude and longitude.

Below variables are used as input features for the prediction of wildfire events in the active fire data.

1) 10 m wind speed (aggregated 10 m vertical speed and 10 m horizontal speed into directional speed).
2) 2 m temperature and 2 m dew point temperature (calculated relative humidity from these two measures of temperature).
3) Daily average precipitation for the previous month.
4) Daily average temperature for the previous month.
5) Vegetation coverage such as Leaf Area Index (LAI) for both low (shrubs and grasses) and high (forests) vegetation.
6) Calculated distance to the closest developed land using Normalized Difference Vegetation Index (NDVI).

Metrics are crucial in wildfire spread prediction because they are critical to the performance evaluation of models [26]. In wildfire events classification, accuracy and F1-score are commonly used metrics to evaluate the performance of models when predicting binary outcomes [26]. Accuracy measures the overall percentage of correctly classified with fire and with no fire events. Through accuracy the quality of produced solution is evaluated based on the percentage of correct predictions over total instances. As there are much less wild fire cases than no fire cases, the input data set is largely

imbalanced. It is well known that for an imbalanced dataset, the accuracy is not a good metric for the quality of a classification, as some random guess may beat the model estimation in terms of accuracy when the testing data is extremely imbalanced. The F1-score, which is the harmonic mean of the precision and recall of the confusion matrix of a classification model, provides a better metric for imbalanced data modeling. The precision is the proportion of correctly predicted positive instances out of all predicted positives, while the recall is the proportion of actual positive fire occurrences correctly detected.

Further, we used ROC (receiver operating characteristic) curve to compare the false positive rate with the true positive rate for a model performance. The true positive rate represents the proportion of correctly forecast fire events, whilst the false positive rate represents the proportion of forecast fires in which no fire occurs [27]. The area under the ROC curve (AUC) is the metric for assessing a model's effectiveness in distinguishing between a correctly predicted fire event v.s. an erroneously predicted fire event. An AUC score that is close to 1 signals a highly skilled prediction model.

IV. Methodology

With the input features described in the last session, we treat the wildfire probability forecast problem as a binary classification problem. The input features taken from the re-analysis product (ERA5-Land [11]) include daily mean precipitation, 10 m wind speed, 2 m dew point temperature, and relative humidity that is calculated from 2 m dew point temperature and 2 m temperature. Fuel characteristics variables such as vegetation coverage, Leaf Area Index, surface pressure were used following McNorton and Di Giuseppe (2023) [27]. We further calculated the average precipitation in the previous month and the average temperature in the previous month as the prediction variables, as high temperature and low precipitation are significant conditions for wildfire occurrences. As demonstrated in studies by McNorton et.al. (2023) [27], the land cover map for urban fraction is an important factor for wildfire predictions. In this research we calculated the closest distance of fire locations to developed land area that were identified by NDVI index. This results in 10 input features, 6 directly from ERA5-land data, 4 calculated from ERA5 climate data and NDVI land cover data.

Random Forest [22] and Extreme Gradient Boosting (XGBoost) [28] are two main tree-based supervised algorithms to solve classification problems. Both RF and XGBoost methods are supervised algorithms which involve training an ensemble of decision trees to resolve the classification problems. The Random Forest approach randomly assigns features to a number of decision trees and then ensembles the majority votes from these decision trees to form the final prediction [27]. XGBoost improves the performance of decision tree with Gradient Boosting, emphasizing on building the subsequent base learners based on the misclassifications from the previous learner [27].

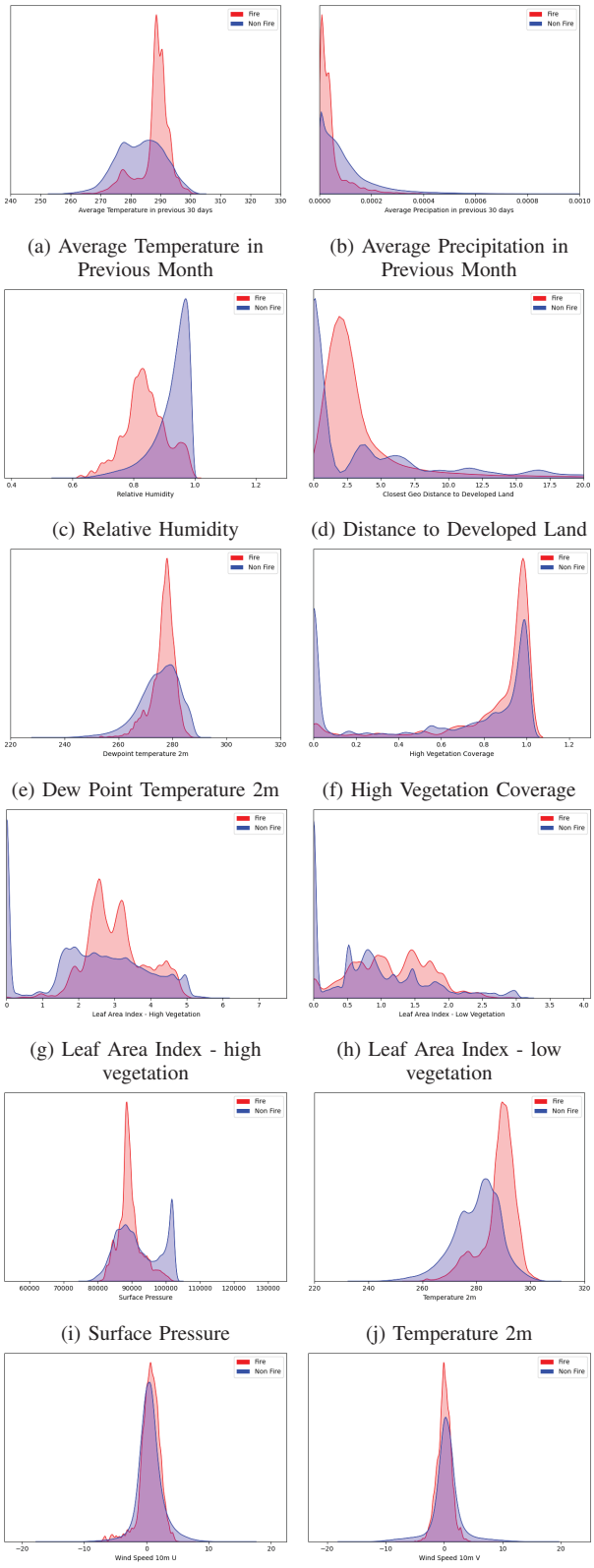
We first applied the decision tree to produce a set of rules that are easily interpretable and a flow diagram to demonstrate the rules. The sample decision tree in Figure 6 shows the important roles that the calculated variables, specifically the total precipitation in the previous month and the closest distance to the developed land, played in the classification of fire events v.s. no fire events.

For the binary classification setup, the target variable y is set to positive (y=1) for the fire burned grid cells, and negative (y=0) for the non fire burned area. As wildfire occurrence is intrinsically stochastic (Prapas et al. 2021) and relatively sparse compared to negative no fire data, we randomly sample two times more negatives from no fire data within 1 latitude and 1 longitude region from each fire data, to avoid incurring a highly imbalanced training data set. This means that the negative sampling is stratified by the location distribution of the positive samples to control the impact of land cover type on the wildfire probabilities.

In addition to random sampling, we also did a tempo-ral split for the evaluation. We applied 2015-2018 data as training data set and 2019 data as testing data set, keeping 2020 data for validation. In total, there are 201,721 training (107,201 non-fire, 94,520 fire), 33,736 validation (18,385 nonfire, 15,351 fire), and 28,963 testing (21,635 nonfire, 7,328 fire) sample data.

V. Results

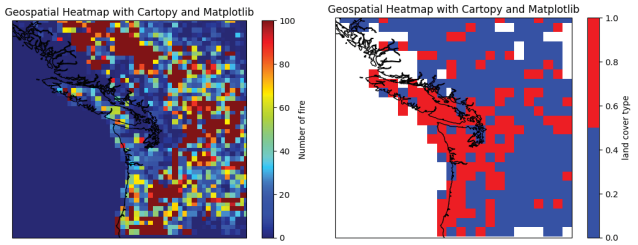
The probability density distributions of each dimension of the above data cube are demonstrated for the grid cells with wildfire burned area v.s. grid locations without wildfire events. There exist features for the prediction of wild fires, including weather data, vegetation coverage, and land cover type represented by the Normalized Difference Vegetation Index (NDVI). The daily weather data and vegetation coverage data are downloaded from the re-analysis product of ERA-5 Land [11], provided by the climate store of ECMWF (the European Center for Medium-Range Weather Forecasts. To align the fire data with the ERA-5 Land data (with a gridded resolution of latitude or longitude of 0.25, or 31km in distance), the average input variable values are calculated for the four ERA5 grid points around each active fire location. For the time period of 2015 to 2020 in the selected region, there are 147222 daily fire event spatial-temporal data. For land cover type, the gridded daily Normalized Difference Vegetation Index (NDVI) was derived from the NOAA (National Centers for Environmental Infor-mation) Climate Data Record (CDR) of Visible Infrared Imaging Radiometer Suite (VIIRS) Surface Reflectance. This data is projected to 0.05 latitude and longitude grid resolution and spans from 2014 to 10 days before the present, using data from NOAA polar orbiting satellites. This satellite based data has been converted from its netCDF-4 file format to daily .csv files that contain NDVI values for locations represented by latitude and longitude. As the fire events can occur at any latitude and longitude but the no fire data represented by ERA5 data can be only on 0.25 grid points, there exists bias when calculating distances to developed land. To further align fire data and non-fire data, the ERA5 data on the grid points are combined to get the non-fire data at the center of each grid cell. The distribution charts revealed the important contributions of relative humidity, the distance to developed land, and the temperature to distinguish fire events from non-fire data.



(a) Average Temperature in Previous Month
(b) Average Precipitation in Previous Month
(c) Relative Humidity
(d) Distance to Developed Land
(e) Dew Point Temperature 2m
(f) High Vegetation Coverage
(g) Leaf Area Index - high vegetation
(h) Leaf Area Index - low vegetation
(i) Surface Pressure
(j) Temperature 2m
(k) Wind Speed 10m Horizontal
(l) Wind Speed 10m Vertical

Fig. 1. The distribution of the values of the input features for burned and not burned cells. The y-axis represents the probability density.

To evaluate the relationship of fire danger with land cover type, the heat maps of number of fire events in the 5 year time period, and the NDVI (Normalized Difference Vegetation Index) were created and displayed side by side. The fire density was calculated by binning the active fire data into 31km by 31km grid cells and count the number of wildfire events in each grid cell over 2015 - 2020 time period. Figure 2 shows the fire density map together with the land cover type for the region in study. The land cover type was identified as developed land if the NDVI is between 0.0 and 0.5. Most of the wildfires were observed to occur near developed land with NDVI less than 0.5. The qualitative inspection of fire danger heatmap reveals the distance to human activities or developed land as a potentially important factor for wildfire risks.



(a) Fire Density Heatmap (b) Land Cover Type Heatmap

Fig. 2. (a) Fire density heatmap for 5 years from 2015 to 2020, calculated from the active fire data from MODIS Collection 6.1. Fire density is the number of fire event days within grid cells of 31km by 31km. (b) Land cover type heatmap is created from the average NDVI within grid cells of 31km by 31km. Areas with NDVI less t han 0.5 are categorized as developed land (blue) and areas with NDVI greater than 0.5 are categorized as undeveloped land (red).

Both Random Forest and XGBoost tree models performed relatively well in random sampling case, with F1-score greater than 0.87 and accuracy as high as 0.89. The grid search cross validation was applied to optimize the hyper-parameters of each model, splitting the sample space into 60% training, 20% testing, and 20% validation. The best performing hyper-parameters were selected as max-depth of 10, number of estimators of 10, and minimum sample for split of 25 to show the result of Random Forest algorithm and XGBoost algorithm. The AUC for the Random Forest model was 0.906, which is lower than that of XGBoost 0.915.

Below table 1 shows the comparison of accuracy and F1-score for the best Random Forest model v.s. the best XGBoost model: It shows that XGBoost slightly outperforms Random Forest in terms of F1-score and accuracy measure on test data.

Table I also shows the comparison of ROC and AUC measure for XGBoost compared with Random Forest model. XGBoost also outperforms Random Forest in terms of ROC AUC measure.

TABLE I. MODEL PERFORMANCE: RANDOM SAMPLING

Method	Accuracy	F1-Score	ROC-AUC
Random Forest	0.87	0.86	0.906
XGBoost	0.887	0.882	0.915

To select the best set of model hyper-parameters, such as the depth of the underlying decision trees, the number of underlying learners to use, and the minimum number of samples allowed in a tree node, the grid search cross validation algorithm was applied to split the training data into five folds, and use the fifth fold to cross validate model hyper-parameters trained on the first four folds of training data. Below table demonstrates the selection process for the major hyper-parameters of the Random Forest model. The most important hyper-parameter that affects model accuracy is the maximum depth of the underlying decision trees. The sensitivities of model performance to the minimum samples in tree node and the number of estimators are not shown below as they are less significant compared to the sensitivities to the maximum depth of the decision trees.

TABLE II. GRID SEARCH CROSS VALIDATION: RANDOM FOREST HYPER-PARAMETERS

MaxDepth	Min Sample	#Estimators	Accuracy
2	25	10	0.745
4	25	10	0.795
6	25	10	0.82
8	25	10	0.845
10	25	10	0.875

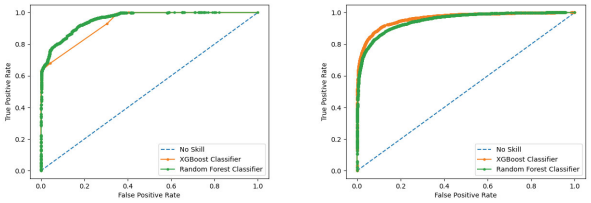
As the pattern for wildfires tends to repeat under similar climate and land cover conditions, we further evaluated the forecast impact of a temporal splitting of data. We used the data from 2015 to 2018 as training set, the data for 2019 as testing set, and the data for 2020 as validation set. The logic behind this sample splitting is to

investigate whether the climate and land cover pattern for previous year active fires can help to predict the coming fire season risks. Table II shows the model performance measures in such a temporal split of samples.

TABLE III. MODEL PERFORMANCE: 2015-2018 TRAINING, 2019 TEST, 2020 VALIDATE

Method	Accuracy	F1-Score	ROC-AUC
Random Forest	0.903	0.775	0.956
XGBoost	0.906	0.777	0.932

A receiver operating characteristic (ROC) curve was used to show the comparison between the false positive rate and the true positive rate of different model performance. The true positive rate represents the proportion of correct prediction, while the false positive rate represents the proportion of wrong forecast of fire events [27]. The ROC curve for XGBoost and Random Forest models are displayed in below Figure 3. For the temporal splitting case, Random Forest outperforms XGBoost in terms of ROC-AUC measure.

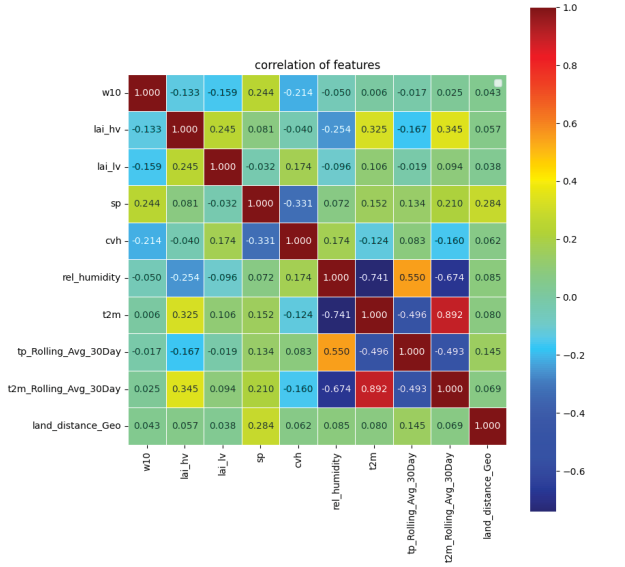


(a) training 2015-2018, testing 2019, validate 2020

(b) Random Sampling 60%, 20%, 20%

Fig. 3. ROC-AUC (Receiver Operating Characteristic curve) c ompares the probabilistic false positive rate with the true positive r ate for prediction Models at 16km grid daily resolution. (a) ROC-AUC measure for temporal data split by using 2015-2018 data as training, 2019 data as testing, and 2020 data as validation set. (b) Random sampling split 60% as training, 20% as testing, and 20% as validation.

The correlations of input features were further examined to identify any potential over-fitting problems caused by multi-collinearity. As shown in Fig. 4, the pairwise correlations among the input features are mostly less than 50%, indicating that no features need to be removed as the tree-based models such as Random Forest are robust to multi-collinearity and the model's predictive accuracy is marginally impacted by correlated features. However, the presence of correlated features in random forests may cause the instability and potential bias in standard feature importance measures (like Gini importance or mean decrease in accuracy). The feature importance score will be diluted or "shared" between the correlated variables, making it harder to determine which variable is truly more influential. In addition to feature importance scores, the importance of input features was investigated by looking at the Shapley values (SHAP). The SHAP value computes the marginal contribution of each feature. Positive SHAP means a contribution to positive (higher) fire probability, while negative SHAP means a contribution to negative (lower) fire risk. Fig. 5 shows the SHAP of the 10 most important features.

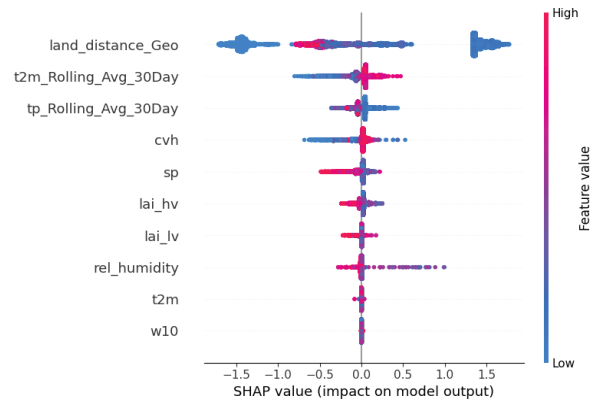


(a) Correlations of Input Features

Fig. 4. The correlation matrix for all the input features shows no significant correlations except a correlation of -0.74 between 2m temperature and relative humidity. Relative humidity is calculated from 2m temperature and 2m dew point temperature.

The SHAP summary plot shows several interesting features. Most of active fires are associated with lower distance to developed land. The higher temperature and lower precipitation in the previous month contribute to increased fire dangers. Higher surface pressure and higher humidity contribute to lower fire risks. We do not observe significant marginal contributions from wind speed as reported by other authors [29].

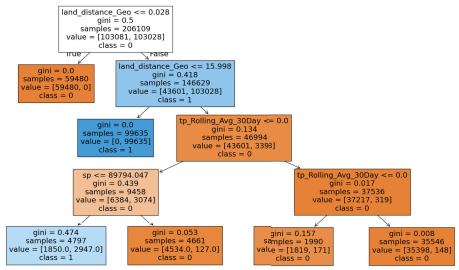
Fig. 6 illustrates the decision making process rendered by tree-based classifiers such as two models we adopted, one underlying learner for the Random Forest



(a) SHAP Values of Input Features

Fig. 5. The SHAP value denotes the importance of any given feature in a model, with a positive SHAP value indicating a positive impact on the model prediction and a negative SHAP value a negative impact on the model prediction. As shown in the figure, higher temperature, lower humidity, and lower distance to developed land contribute to higher fire probabilities.

or XGBoost, two ensemble learning algorithms based on decision trees.



(a) One base learner of Random Forest

Fig. 6. Decision Tree Model shows one set of splitting rules for the classification of data into positive fire and negative no fire target classes.

VI. Discussion

Wildfire occurrence is intrinsically stochastic and sporadic. For the region in this study, there is only a small sample of fire events (147222 spatial-temporal records) compared to the number of grid points with no fire events on a grid resolution of 31 km by 31 km (more than 500 million records). As positives are limited, the negative sampling is stratified by the location distribution of fire sample data, collected from within latitude range of +1 or -1 and longitude range of +1 or -1 from the location of fire events at the same time scale. For example, if on a date 11/2/2020 there is a wildfire event at latitude 43.5 and longitude -123.5, then we sample all the grid points with no fire events at latitude from 42.5 to 44.5 and longitude from -124.5 to -122.5 on the date 11/2/2020 as negative data points. This sampling strategy largely reduced the imbalance issue of training data set. The next step of data processing is to filter out sample data with invalid NDVI index or unreasonable distance to developed land. These two steps make sure the fire data and no fire data are with similar land cover type so to remove the bias brought by land cover type or human behavior factors.

In this research, we collected the climate data, the vegetation cover data, and the active fire data for the region at the border of British Columbia, Canada and Oregon, US. We provided three new variables that can be derived from the raw climate data and the satellite based NDVI index, i.e., total precipitation in the previous month, average temperature in the previous month, and the closest distance to the developed land. To identify which variables drive the predictions of fire danger, we compute the feature importance scores and the SHAP values to compare the marginal contribution of each feature. ALL three calculated variables have high contributions to fire danger predictions, with the closest distance to developed land having the highest feature importance score. As NVDI index and ERA5-Land grid points are at different resolutions, the NDVI index grid resolution is 0.05 latitude and longitude, while the ERA5-Land climate data are at 0.25 grid resolution, the no fire data tend to be distributed at the lower (close to 0.0) and higher end (higher than 30) of the probability density curve of the distance to closest developed land. This is a bias brought by the data availability and data resolution differences.

It is interesting to study how the model performance of Random Forest or XGBoost is affected by the most important feature, the distance to the closest developed land. If this distance feature is removed, the model accuracy is decreased to 0.806 and the F1-score is reduced to 0.797. Below figure compares the model performance with or without the distance feature.

TABLE IV. MODEL PERFORMANCE: WITH DISTANCE TO CLOSEST DEVELOPED LAND

Method	Accuracy	F1-Score	ROC-AUC
Random Forest	0.87	0.86	0.906
XGBoost	0.887	0.882	0.915

TABLE V. MODEL PERFORMANCE: WITHOUT DISTANCE TO CLOSEST DEVELOPED LAND

Method	Accuracy	F1-Score	ROC-AUC
Random Forest	0.806	0.797	0.879
XGBoost	0.817	0.81	0.892

Major fire events are commonly driven by either low relative humidity or high wind speed [29]. Our results confirmed the significant marginal contributions from the surface pressure and the relative humidity (Fig. 5). In addition, our model identified the proxy to human behavior, the distance to the developed land, as the input feature with potentially the most impact on fire prediction model performance. As converting satellite based land cover images to NDVI index values requires large memory and storage resources, the calculation of the distance to the developed land is very time consuming and subject to resource limitations. Each distance measure for a single fire location requires filtering of a daily NDVI index file for all the available latitudes and longitudes of a specific date. For further research, it would be interesting to investigate whether the impact of the distance to the developed land on active fire dangers persists to other fire-prone areas, such as Eastern Mediterranean or Southern Australia.

VII. CONCLUSION

Wildfire predictability is limited by its stochastic essence and complex interactions among different climate and land cover factors. This paper contributes to existing data-driven research by applying NASA active fire data and ERA-5 Land data to uncover the major risk drivers behind wildfire events in the region of the British Columbia and Oregon border. We downloaded satellite-based climate data, vegetation cover, active fire data, and Normalized Difference Vegetation Index (NDVI) data in different grid solution. To train machine learning models we further processed and combined data into the same 16km x 16km grid cells to calculate the distance to developed land for each fire event spatial-temporal data.

For the two versions of tree-based algorithms we trained, XGBoost provided slightly better prediction accuracy. The relationships emerging from the machine learning models reveal the importance of fuel-related variables such as distances to developed land, NDVI, and relative humidity. The models also show that the average temperature and total precipitation in the previous month are significant for predicting wildfires. The attribution of input variables to the probability of fire can be derived using evaluation tools such as SHAP [30]. Higher temperature in previous month and lower humidity are associated with higher probability of wildfires. Higher distance to developed land, higher leaf area index, and higher surface pressure are associated with lower probability of fires.

As more variables become available for the future forecast, the addition of more input variables that indicate human interaction activities could further improve the model performance. This paper tried to indicate human activities by deriving the distance to developed land. Most of wildfire events in the region we studied occurred close to the developed land, indicating the importance of human behavioral factors in the wildfire prediction. Investigating the impact of human behavioral factors on wildfire predictions could be the major direction of further studies.

VIII. REFERENCES

[1] J.G. Pausas and J.E. Keeley. Wildfires and global change. Frontiers in Ecology and the Environment, 19, 2021.
[2] M.G. Eliott, T.J. Venn, T. Lewis, M. Farrar, and Srivastava S.K. A prescribed fire cost model for public lands in south-east queensland. Forest Policy and Economics, 132, 2021.
[3] S.H. Doerr and C. Santin. Global trends in wildfire and its impacts: Perceptions versus realities in a changing word. Philosophical Transactions of the Royal Society B: Biological Sciences, 371(1696), 2016.
[4] M.W. Jones, J.T. Abatzoglou, S. Veraverbeke, N. Andela, G. Lasslop, M. Forkel, A.J.P. Smith, C. Burton, R.A. Betts, and G.R. van der Werf. Global and regional trends and drivers of fire under climate change. Reviews of Geophysics, 60(3), 2022.
[5] C.E. Van Wagner. Structure of the canadian forest fire weather index. Environment Canada, Canadian Forest Service, 1333, 1974.
[6] S. Kondylatos, I. Prapas, M. Ronco, I. Papoutsis, G. Camps-Valls, M. Piles, M.A. Fernandez-Torres, and N. Carvalhais. Wildfire danger prediction and understanding with deep learning. AGU Advancing Earth and Space Science, 2022.

[7] M. Forkel, W. Dorgio, G. Lasslop, I. Teubner, E. Chuvieco, and K. Thonicke. A data-driven approach to identify controls on global fire activity from satellite and climate observations. Geoscientific Model Development, 10, 2017.
[8] S. Khanmohammadi, M. Arashpour, E.M. Golafshani, M.G. Cruz, and A. Rajabifard. An artificial intelligence framework for predicting fire spread sustainability in semiarid shrublands. International Journal of Wildland Fire, 32:636–649, 2023.
[9] J.N.S. Rubi, P.H.P. de Carvalho, and P.R.L. Gondim. Application of machine learning models in the behavioral study of forest fires in the brazillian federal district region. Engineering Application of Artificial Intelligence, 118:105649, 2023.
[10] M. Reichstein, G. Camps, Valls, B. Stevens, M. Jung, J. Denzler, N. Cavalhais, and Prabhat. Deep learning and process understanding for datadriven earth system science. Nature, 566, 2019.
[11] J. Munoz-Sabater, E. Dutra, A. Agusti-Panareda, C. Albergel, G. Arduini, and G. Balsamo. Era5- land: A state-of-the-art global reanalysis dataset for land applications. Earth System Science Data Discussions, 13, 2021.
[12] M.A. Finney. Farsite, fire area simulator-model development and evaluation. U.S. Department of Agriculture, Forest Service, 1998.
[13] H. Singh, L.M. Ang, T. Lewis, D. Paudyal, M. Acuna, P.K. Srivastava, and S.K. Srivastava. Trending and emerging prospects of physics-based and ml-based wildfire spread models: A comprehensive review. Journal of Forestry Research, 35, 2024.
[14] P. Jain, S.C.P. Coogan, S.G. Subramanian, M. Crowley, S. Taylor, and M.D. Flannigan. A review of machine learning applications in wildfire science and management. Environmental Reviews, 28(4), 2020.
[15] M.H. Nami, A. Jaafari, M. Fallah, and S. Nabiuni. Spatial prediction of wildfire probability in the hyrcanian ecoregion using evidential belief function model and gis. International Journal of Environmental Science and Technology, 15(2), 2018.
[16] Y. Zhang, S. Lim, and J.J. Sharples. Modeling spatial patterns of wildfire occurrence in southeastern australia. Geomatics, Natural Hazards and Risk, 7(6), 2016.
[17] J.R. Quinlan. Introduction of decision trees. Machine Learning, I(1), 1986.
[18] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. Classification and regression trees. Routledge, 2017.
[19] D. Stojanova, P. Panov, A. Kobler, S. Deroski, and K. Taskova. Learning to predict forest fires with different data mining techniques. 2006.
[20] A. Sulova and J. Jokar Arsanjani. Exploratory analysis of driving force of wildfires in australia: An application of machine learning within google earth engine. Remote Sensing, 13(1), 2020.
[21] M.K. Al-Bashiti and M.Z. Naser. Machine learning for wildfire classification: Exploring black-box, explainable, symbolic, and smote methods. Natural Hazards Research, 2(3), 2022.
[22] L. Breiman. Random forests. Machine Learning, 45, 2001.
[23] K.S. Dahan, R.A. Kasei, R. Husseini, M. Sarr, and M.Y. Said. Analysis of the future potential impact of environmental and climate changes on wildfire spread in ghana's ecological zones using a random forest (rf) machine learning approach. Remote Sensing Applications: Society and Environment, 2024.
[24] A. Afshar, G. Nouri, S. Ghazvineh, and S.H. Hosseini Lavassani. Machine-learning applications in structural response prediction: A review. Practice Periodical on Structural Design and Construction, 2024.
[25] Singh, R. Yadav, G. Sudhamshu, A. Basnet, and R. Ali. Wildfire spread prediction using machine learning algorithms. 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 6-7 July 2023, 2023.
[26] Henintsoa S. Andrianarivony and Moulay A. Akhloufi. Machine learning and deep learning for wildfire spread prediction: A review. Fire, 7, 2024.
[27] J.R. McNorton, Di F. Giuseppe, E. Pinnington, M. Chantry, and C. Barnard. A global probabilityof-fire (pof) forecast. AGU Advancing Earth and Space Sciences, 2024.
[28] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016.
[29] J. Ruffault, T. Curt, V. Moron, R.M. Trigo, F. Mouillot, and N. Koutsias. Increased likelihood of heat-induced large wildfires in the Mediterranean basin. Scientific Reports, 10, 2020.
[30] S.M. Lundberg and S.I Lee. A unified approach to interpreting model predictions. Advances in Neural Information Processing System, 30, 2017.