A Machine Learning Approach for Predicting Wildfire Occurrence using Burn History and Topography

Sumedha Khatter*, Mai H. Nguyen†, Daniel Crawl†, Jessica Block‡, Ilkay Altintas†
*Computer Science & Engineering, †San Diego Supercomputer Center, ‡Qualcomm Institute
University of California, San Diego
La Jolla, CA U.S.A.
skhatter@ucsd.edu, {mhnguyen,crawl,altintas}@sdsc.edu, j.block@eng.ucsd.edu

***Abstract*— Wildfire occurrence and extent is growing all over the world, and it is increasingly threatening infrastructure, homes and human lives. Despite this growing hazard, fire is an integral part of fire-adapted ecologies, and understanding fire occurrence can help better understand fire hazard.**

There is an opportunity to learn from fire history and how the spatial characteristics of the environment influence rates and extents of fire occurrence. Learning from the historical fires of San Diego County in California, this paper presents a system that identifies areas susceptible to wildfires based on fire histories, fire modeling data and landscape features.

Raw data from Landscape Fire and Resource Management Planning Tools Project (LANDFIRE) and the State of California's Fire and Resource Assessment Program (FRAP) is processed and integrated to prepare the dataset used for predicting fire occurrence. Machine learning models are then built using this dataset to predict the likelihood of burning of an area on a pixel-by-pixel basis. Issues with variability of the data in the area of interest is addressed using various oversampling, over-undersampling and undersampling techniques. Comparison of results from various machine learning models and resampling techniques are presented.

I. INTRODUCTION

The United States has seen a significant rise in wildfire extent in the last fifty years [1]. Climate change is blamed for elements of that rise [2] and has shown to increase wildfire danger all over the world [3]. Despite this growing hazard, fire is an integral part of fire-adapted ecologies. As urbanization has encroached into these fire-prone regions, infrastructure, homes and human lives are in growing danger.

There is an opportunity to learn from fire history to understand its threat. The spatial characteristics of the environment influence rates and extents of fire occurrence. San Diego County is an excellent place to study because it experiences many fires annually and has experienced some of California's largest fires in just the last twelve years. In such a heavily urbanized region, we want to know if some areas are more prone to wildfire than others. Have some regions burned more frequently than others, and can we understand why they burn more frequently based on their ecological features?

Learning from the historical fires of San Diego County, this paper presents a system that identifies areas susceptible to wildfires based on fire history, fire model data and landscape features. We describe a machine learning approach for wildfire data analysis using historical burn and landscape data of a particular area of San Diego County. The main contributions of the paper are as follows:

- Collecting geospatial data from LANDFIRE [4] and FRAP [5], and formatting the data amenable to analysis. We provide a series of steps to map all the data, extract the relevant data, and prepare it for the area of interest (a part of San Diego) using Geo-Spatial Data Abstraction Library (GDAL) [6].
- The dataset prepared faces a class imbalance problem. The class of interest comprises less than 2% of the total data. Addressing the class imbalance via resampling techniques provided by Scikit-Imbalance [7] increase the AUC-ROC values over 15% points.
- The construction of a model using supervised learning provided by Scikit-learn [8] to predict the likelihood of fire occurrence in an area. The prediction of model can be used to analyze wildfires trends.

Outline. The rest of the paper is organized as follows: Section II summarizes related work. Section III presents in detail the steps to prepare the dataset. Section IV discusses the features used. Section V describes the modeling process, including splitting of the dataset into training and test sets in Subsection V-A, evaluation metrics in Subsection V-B, addressing class imbalance problem in Subsection V-C, machine learning models built in Subsection V-D, and modeling results in Section V-E. Finally, conclusions and future work are discussed in VI.

II. RELATED WORK

Cortez and Moraiz [9] explored a data mining approach to predict forest fires using a feature selection setup using the spatio-temporal Fire Weather Index (FWI) and meteorological data. They found the best results using an SVM model when given input of temperature, precipitation, humidity and wind speed. Stojanova, Panov, Kobler, Dzeroski and Taskova [10] predicted forest fires on a given day based on forest structure, predicted weather, and MODIS satellite data. Canadian Wildland Fire Information System (Canadian Wildland Fire Information System [11]) has a Canadian Forest Fire Weather Index (FWI) system to estimate the fire risk of a given day using lookup tables based on observations of temperature, relative humidity and 24 hour rainfall.

The system described in this paper is different from the rest as it is utilizing the topographic, fuel properties, and fire history data of a particular area and is predicting the fire occurrence of the next two years. The fire occurrence

of a pixel is defined as a function of a set of landscape and fire related features. And how the results change with various resampling techniques to address the class imbalance problem is also discussed.

III. DATA PREPARATION

This section describes the steps taken to prepare the dataset for the machine learning models. The data analyzed includes a set of five input features that describe each pixel in the study area: elevation, slope, aspect, 2012 surface fuel model, and 1992-2012 cumulative fire frequency. The target data a binary value denoting if the pixel burned during 2013-2014.

A. Data Sources

Table I lists the data used and their sources. Details about these features are provided in Section IV.

TABLE I: Sources details of Dataset

Feature(X)	Source	Year
Elevation	LANDFIRE	2012
Slope	LANDFIRE	2012
Aspect	LANDFIRE	2012
Fuel Model	LANDFIRE	2012
Fire Frequency	FRAP	1992-2012
Label (Y)		
Fire Perimeters	FRAP	2013-2014

The system utilizes fire history data collected from the Fire and Resource Assessment Program (FRAP [5]). The FRAP analyzes the conditions of forests and rangelands and identifies the alternatives and guidelines. The landscape data are collected from LANDFIRE [4]. The Landfire program provides over twenty National geospatial layers, databases, ecological models and tools to review the models that are available to the public.

B. Region Selection for Dataset

The area of interest in this study was chosen to maximize the number of fires from the past twenty years and to have varied topography and vegetation/fuel properties. Figure 1 shows this area selection.

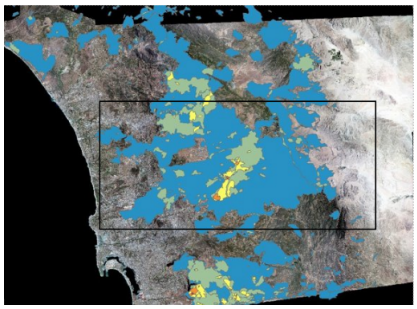

Fig. 1: The figure shows San Diego County and the study area within the boundary. The blue to yellow colored polygons show the cumulative fire frequency from 1992-2012. Yellow shows a cumulative frequency of five fires and blue denotes one.

C. Preparing Data from Fire Perimeters

Preparing Label of the Dataset. The steps to prepare the target data (2013-2014 fire perimeters) from the historical fire database are written below:

1) Reproject the original shapefile (for the system described in this paper, it is UTM Zone 11).
2) Select and export the records from the years 2013-2014 (2013 and 2014 both inclusive).
3) Rasterize the vector data saved in the previous step.
4) Clip the rasterized data to the study area.

Preparing Cumulative Fire Frequency Feature of the Dataset. The following explains the calculation of the cumulative fire frequencies between 1992 and 2012.

1) Change the projection of fire history dataset to the desired map projection.
2) Choose the records of the years 1992-2012.
3) Calculate the overlapping fire polygons.
4) Rasterize the vector file saved in the previous step.
5) Clip the rasterized data to the study area.

D. Preparing Data from Landscape File

The dataset prepared for this project uses first four bands of the file namely Elevation, Slope, Aspect and Fuel Model Number.

Steps to prepare dataset features using this file are written below:

1) Warping of the original file to reproject it into a desired common projection.
2) Translating the multi-band raster to individual first four raster band files.
3) Clipping of the four single band raster data to get data of a desired area.

After following the procedure above, a total of six files are created. Label file, four features using the landscape file and one feature using cumulative fire frequency. Thus, dataset is prepared. It is explored in the following section.

IV. EXPLORING FEATURES

A. Shape of the original dataset

As already discussed in Section III-A, the dataset has five features. The shape of the dataset is described in the first row of Table II. The majority of pixels are in the 'Non-burnt' class (98.53%) while the number of pixels in the 'Burnt' class is very small 1.47%. The dataset thus faces class imbalance problem, as the two classes are not equally represented.

B. Features of the Dataset

The file obtained for each feature is 1554×3002 pixels each representing 30m $\times$ 30m. The total area becomes $(1554 \times 3002 \times 30 \times 30)/(1000 \times 1000) = 4198.59$ km^2. The total area of San Diego County is 11720 km^2 [12]. The area chosen for the preparation of the dataset is 35.82% of the total area of San Diego County. The error in the dataset is calculated by counting the total number of '0' values in the Fuel Model Number feature (refer Figure 6) by the feature

TABLE II: Splitting of dataset

Set	Total Samples	Non-burnt (0)	Burnt (1)
Original	4,665,108 (100%)	4,596,878 (98.53%)	68,230 (1.47%)
Training	3,732,086 (80%)	3,677,502 (98.54%)	54,584 (1.46%)
Test	933,022 (20%)	919,376(98.53%)	13,646 (1.47%)

size and is equal to 0.72%, The following subsections explain the five features of the dataset in detail using histograms and statistics.

1) Elevation: Elevation represents land height, in meters, above mean sea level (LANDFIRE: Topographic [13]).

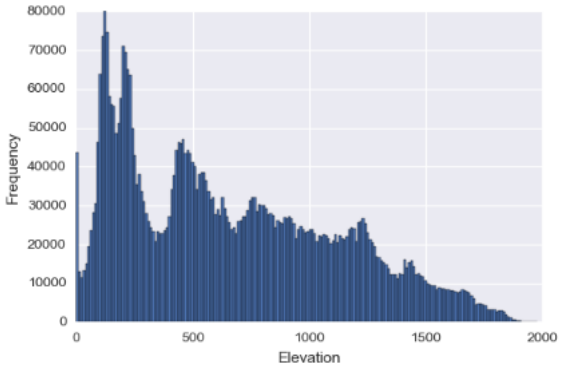

Fig. 2: Histogram describing the statistics of Elevation

Table III, shows the area chosen for the dataset has an average elevation of 663.61m above sea level. The maximum elevation goes to 1,977. Figure 2 shows the histogram describing the statistics of Elevation.

2) Slope: Slope is the relationship of vertical rise to a horizontal run, expressed as a percentage change of Elevation over a specific area. It is expressed in degrees (LANDFIRE: Topographic [13]).

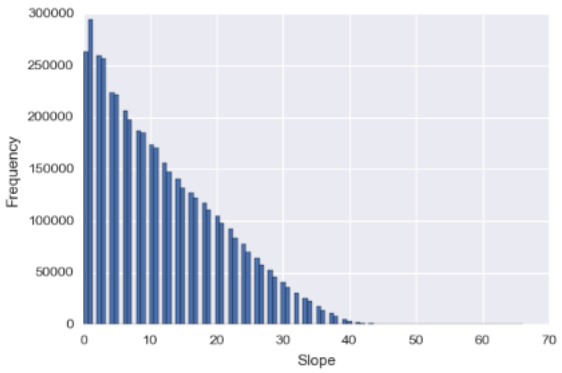

Fig. 3: Histogram describing the statistics of Slope

Figure 3 represents the histogram describing the statistics of Slope. Looking at the histogram, it can be inferred that the higher frequencies of Slope lie in the level, gentle, and moderate scales. From Table III, it can be seen that the median value is 10 and mean is 11.62.

3) Aspect: Aspect identifies the direction of the Slope with values from 0 to 359 degrees indicating the clockwise compass direction.

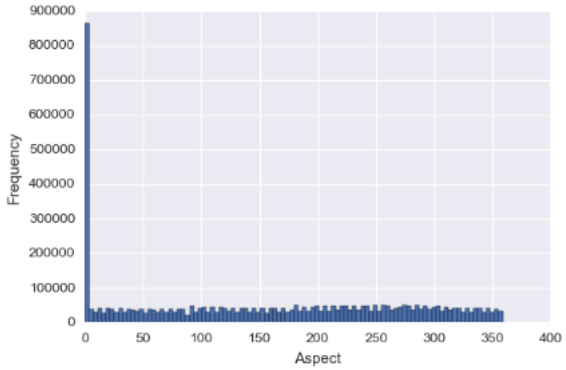

Fig. 4: Histogram describing the statistics of Aspect

The National Avalanche Center describes aspect is most important with respect to the sun in mid-latitude regions [14] such as San Diego. In the northern hemisphere, south facing slopes receive more heat as compared to north facing slopes which receive very little heat from sun. Thus, there will be a different climatic conditions on north facing as compared to south facing. North facing will be cooler, more shadier, moist and have different kind of vegetation as compared to a south facing slope. East facing slopes are colder than west facing slopes since the east side receive the sunlight during the morning hours while the west side receive the sunlight during the noon hours.

4) Fuel Model Number: Fuels are classified into four groups: grasses, brush, timber and slash [15]. Each fuel model is represented by the fuel load. Fuel load is the amount of fuel that is potentially available for combustion and the ratio of surface area to volume for each size class; the depth of fuel bed. A fuelbed is a combination of one or more fuel strata (Joe Scott) involved in the fire front; fuel moisture. Figure 5 represents the distribution of Fuel Model Number band. The orange color represents the non-burnable fuel. The blue color represents the Fuel Model Number corresponding to burnable category. On mapping Figure 1 with Figure 5, it can be said that the blue color (burnable fuel) of Figure 5 almost maps to the past twenty years of fire frequency layer of Figure 1. Thus, the orange colored area, can be considered as impervious to wildfires and represents the characteristic features of non-burnable class.

A brief description of Fire Behavior Fuel Model (FBFM) Numbers is written in Table IV. For more details, refer to Anderson [15].

Figure 6 represents the Fuel Model histogram. Short Grass occupies 41%, Grass with Timber/Shrub Overstory occupies 17%, and then Young Brush. All these fuels have different

TABLE III: Statistics of Elevation, Slope, and Aspect

Statistics	Elevation	Slope	Aspect
Maximum	1977	66	359
Minimum	0	0	0
Mean	663.61	11.62	151.59
Median	578.0	10.0	150
Std. Deviation	454.39	9.175	116.32
Variance	206,466.67	84.185	13,532.35

TABLE IV: Fuel Model Number Description

Class	Description
0	No Data
FBFM1, FBFM2, FBFM3	Grass and Grass dominated
FBFM4, FBFM5, FBFM6, FBFM7	Chaparral and shrub fields
FBFM8, FBFM9, FBFM10	Timber litter
FBFM11, FBFM12, FBFM13	Slash
91, 92, 93, 98, 99	Non-Burnable

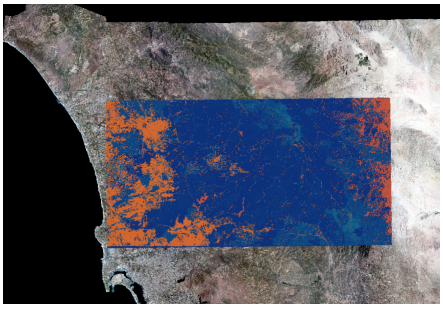

Fig. 5: Fuel Model Number band view

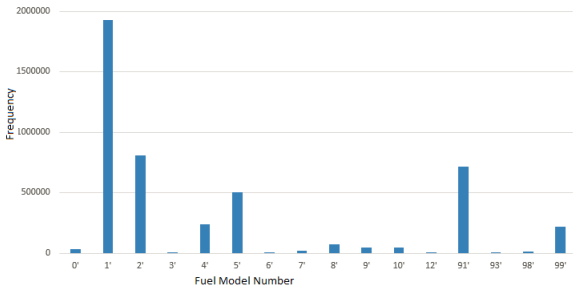

Fig. 6: Histogram describing the statistics of Fuel Model Number

fire behavior properties.

C. Cumulative Fire Frequency

This feature is prepared as discussed in Section III-C, and can help in understanding which areas are more susceptible to repeatedly burn, and which areas rarely burn. The value for each pixel represented by this feature is the number of times there was a fire in that pixel during 1992-2012.

Within the black boundary of Figure 1, the blue, green and yellow colored polygons represent the area which got burnt in past twenty years. The yellow color represents the highest value of fire frequency, five and the blue color represents the lowest value, one.

From the histogram shown in Figure 7, it can be deduced that around 50% of the area never got burnt in the past twenty years. 40.23% got burnt only once, 7.8% burnt only twice, 1.31% three times, 0.12% four times, and 0.004% five times.

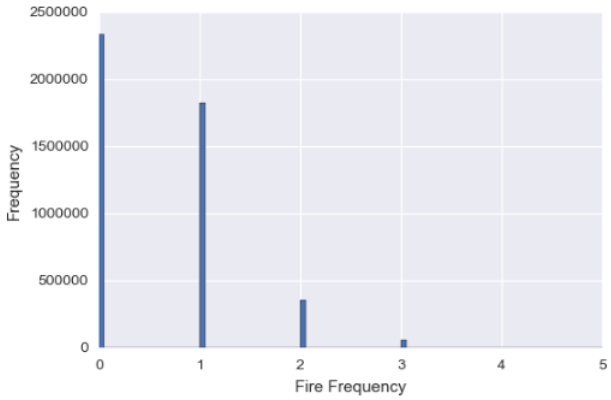

Fig. 7: Histogram describing the statistics of cumulative fire frequency

D. Target Data

The target data is one if fire occurred during 2013-2014 and zero if not. Figure 8 depicts this data: the orange area represents the area which had fires in 2013-2014 and blue represents the non-burnt area. From this image, it is evident that the dataset is imbalanced. The burnable class represents a very small percentage 1.47% of the total area (also refer first row of Table II).

Fig. 8: Target layer (fire occurrence during 2013-2014)

E. Correlation Between Variables of the Dataset

A correlation matrix [16] is a table showing correlation coefficients between sets of variables. Each variable X_i in the table is correlated with each other values in the table (X_j). This can help understand which pairs have highest, lowest or no correlation. And obviously, the diagonal elements are all 1 as it is a relationship reflecting how a variable is related to itself.

The features are represented as Ff (Cumulative Fire Frequency), E (Elevation), S (Slope), A (Aspect), FM (Fuel Model Number) and C (Target Class). Positive values represent positive relationship and vice versa. Figure 9 shows the pairwise relationships of all variables and the corresponding values. It can be seen that there is profound correlation between Elevation and Slope, Aspect and Slope, Slope and cumulative fire frequency. Higher elevations have higher slopes and higher slopes have faced higher number of fires.

	Ff	E	S	A	FM	C
Ff	1.000000	0.083812	0.246805	0.152234	-0.346138	0.018798
E	0.083812	1.000000	0.306614	0.149148	-0.359786	0.077342
S	0.246805	0.306614	1.000000	0.310003	-0.240521	0.071563
A	0.152234	0.149148	0.310003	1.000000	-0.162514	-0.021614
FM	-0.346138	-0.359786	-0.240521	-0.162514	1.000000	-0.046330
C	0.018798	0.077342	0.071563	-0.021614	-0.046330	1.000000

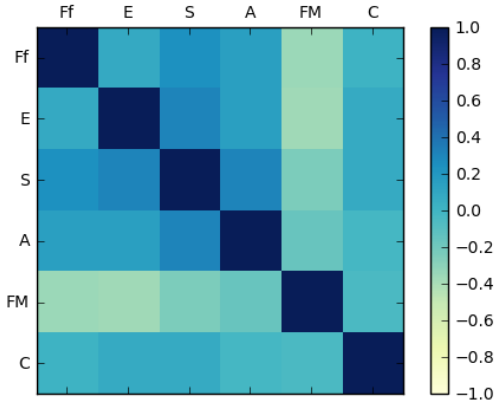

Fig. 9: Correlation Matrix of different features and the target variable

V. MODELING

This section presents the modeling process, including evaluation metrics used, addressing class imbalance, machine learning models built, and modeling results.

A. Splitting of Dataset

Table II shows the splitting of original dataset into a training set (80%) and a test set (20%). The data is split using a stratified sampling using the class labels. This is because, when subpopulations within an overall population vary, it is advantageous to sample each subpopulation or stratum independently [17]. From Table II, it is also clear that the percentage of non-burnt samples and burnt samples ratio is the same in all three sets: 98.53% non-burnt samples and 1.47% burnt samples.

B. Scores to Evaluate the Model

There are various evaluation metrics that can be used to analyze a model's performance. We discuss the most commonly used metrics below.

A confusion matrix is used to evaluate the quality of the output of a classifier. The diagonal elements represent the number of points for which the predicted labels is equal to the true label, while off-diagonal elements are those that are mislabeled by the classifier [18] [19]. Table V shows a 2*2 confusion matrix for binary dataset describing what each cell of matrix represent.

Accuracy score is another metric to evaluate a data model and is given by sum of total true values of both the classes divided by total values [19]. It can be written as:

$$AccuracyScore = (TP + TN)/(TN + FP + FN + TP)$$

For this dataset, which is very large and skewed, the number of samples that belong to (FN + TP) is 1.5% of the total dataset, and the number of samples that belong to (TN + FP) is 98.5% of the total dataset. Therefore, the following can be written:

$$(TP + TN) \approx TN$$

$$(TN + FP + FN + TP) \approx (TN + FP)$$

$$AccuracyScore \approx (TN)/(TN + FP)$$

From the above equation, it is clear that Accuracy Score does not reflect the values of FN or TP, which represents the minority class. Hence it cannot be used to evaluate this dataset correctly.

Consider two operating characteristics: true positive rate (TPR) and false positive rate (FPR). TPR [20] is given by the following equation:

$$TPR = TP/(TP + FN)$$

and FPR [21] is given by the following equation:

$$FPR = FP/(FP + TN)$$

Both TPR and FPR are rates for their corresponding class and hence the would not be affected by the size of the dataset. The Receiver Operating Characteristic, or ROC, curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various settings [22] [23]. The ROC curve is not affected by the size and skewness of the dataset. Hence, we use the confusion matrix and ROC curve as evaluation metrics in our experiments.

For the confusion matrix, higher values of the diagonal elements are desirable, indicating many correct predictions. TPR and (1- FPR) should be close to one for a better model. The area under the ROC curve (abbreviated as AUC) is often used as a metric to evaluate prediction results. An AUC value of 1 indicates perfect prediction, and 0.5 represents random guessing. In a ROC curve plot, such as Figures 10 and 11, $AUC = 0.5$ is represented as a dashed line. An AUC value greater than 0.8 is considered as a good prediction score [24].

TABLE V: Confusion Matrix

Class	Prediction (0)	Prediction (1)
Actual (0)	True Negative (TN)	False Positive (FP)
Actual (1)	False Negative (FN)	True Positive (TP)

C. Addressing Class Imbalance

From Table II, it is apparent that there is a class imbalance problem with this dataset. Resampling techniques can be used to address class imbalance. We tested several resampling techniques on our dataset and present the results in this section. The various resampling techniques are oversampling, over-undersampling, and undersampling techniques. We tested these different techniques using a decision tree classifier [25] [26].

Oversampling is tested using Synthetic Minority Oversampling with Replacement (SMOTE). This technique generates new samples of the minority class by applying operations like tweak, rotation along the line segments joining nearest neighbors [27].

Over-undersampling is tested using two techniques: SMOTE+ENN and SMOTE+Tomek. The intuition behind SMOTE+ENN is to do oversampling using SMOTE followed by Edited Nearest Neighbor (ENN) to remove any example that is misclassfied by its three nearest neighbors [28]. SMOTE+Tomek is similar to SMOTE+ENN apart from the fact that instead of ENN, tomek links are identified and removed. A pair of two points (x,y), where x belongs to class A and y belongs to class B, is a tomek link if for any point z, $d(x,y) < d(x,z)$ or $d(x,y) < d(y,z)$, where d(x,y) is the distance between points x and y and so on. And if any two examples are tomek links, then either one of them is noise or both of them could be located on the boundary classes [28].

Undersampling is tested using random-undersampling. This technique under-samples the majority class by randomly selecting samples with or without replacement [7].

Table VI shows the class distribution of the training set after applying four resampling techniques. It is worth noting that the size of training set increases while using SMOTE, SMOTE+ENN and SMOTE+Tomek and decreases while using random-undersampling. This can be verified by comparing these values to those in Table II).

Table VII presents the normalized confusion matrix values and the AUC scores on the test data using each resampling technique. Figure 10 displays the ROC curves. It can be seen that the random-undersampling provides the best results as this technique yields the highest AUC scores and normalized values of diagonal elements of the confusion matrix (True Negative (TN) and True Positive (TP)).

The oversampling and over-undersampling techniques do not work well for this dataset and give scores almost similar to the case without using any resampling technique. It can be due to the fact that not much variability of features is increased by synthesizing new samples. On the other hand, undersampling selects few samples from the majority class and brings equal variability for both majority and minority classes, thus providing better modeling results.

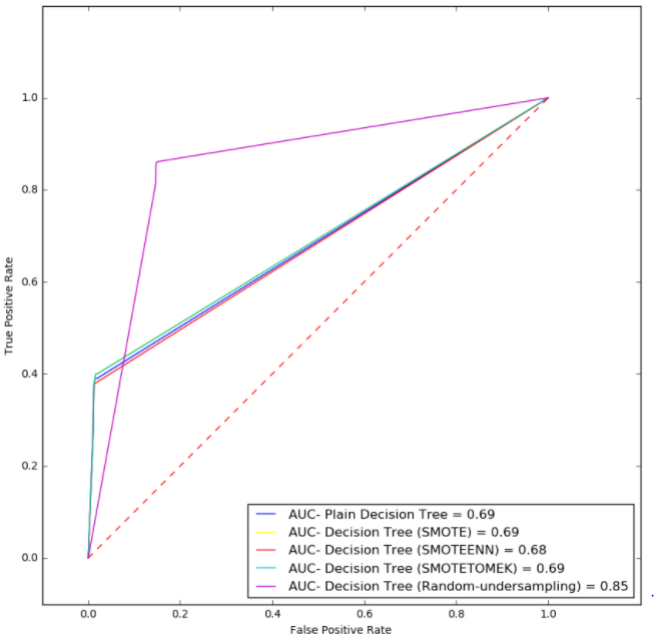

Fig. 10: Area Under Curve (AUC)- Receiver Operating Characteristic (ROC) corresponding to class imbalance problem, refer Table VII

D. Predictive Models

This section describes the use of various supervised learning data models to predict the occurrence of fires. These models are trained using the training set resampled with random-undersampling. Logistic Regression [29] [8], SGDClassifier [8] with 'log' loss, decision tree [25] [26], and random forest [30] [31] are used.

Decision tree and random forest models are also tuned to achieve better results. Tuning for decision tree is done across the following hyper-parameters: number of features to consider for split (max_features), maximum depth of the tree (max_depth), minimum number of samples required to split an internal node (min_samples_split), minimum number of samples required in a leaf node (min_samples_leaf) and function to measure the quality of split Gini index or the entropy (criterion). Tuning for random forest is done across the number of trees in a forest (n_estimators). Table VIII shows the hyper-parameters chosen for these models and the best values obtained for each feature.

E. Results

Table IX shows the normalized confusion matrix values and the AUC scores on the test data using each of these classifiers. Scores for the untuned and tuned decision tree models are also included. Similarly for random forest. Figure

TABLE VI: Class distribution of the training set after applying various resampling techniques

Technique	Non-Burnt (0)	Burnt (1)
SMOTE	3,677,502 (50%)	3,677,502 (50%)
SMOTE+ENN	3,677,502 (51.14%)	3,512,678 (48.85%)
SMOTE+Tomek	3,677,502 (50.05%)	3,669,641 (49.94%)
Random-undersampling	54,584 (50%)	54,584 (50%)

TABLE VII: Scores using various resampling techniques

Classifier	True Negative	False Positive	False Negative	True Positive	AUC-ROC
Plain Decision Tree	0.99	0.01	0.68	0.32	0.69
SMOTE	0.99	0.01	0.61	0.39	0.69
SMOTE+ENN	0.99	0.01	0.63	0.37	0.68
SMOTE+Tomek	0.99	0.01	0.61	0.39	0.69
Random-undersampling	0.85	0.15	0.14	0.86	0.85

TABLE VIII: Hyper-parameters Grid

Decision Tree Feature	Values	Best Value
max_features	1,2,3,4,5	4
max_depth	1,5, 10, 20	20
min_samples_split	2,10, 50, 100, 200, 500	50
min_samples_leaf	2, 10, 50, 100, 200, 500	10
criterion	gini, entropy	gini
Random Forest Feature	**Values**	**Best Value**
n_estimators	20,30,40,50,60,70,80, 90, 100	100

11 displays the corresponding AUC curves. It can be observed that the SGDClassifier performed poorly compared to the other classifiers. Logistic regression and untuned decision tree achieve the same AUC score. However, the decision tree has higher values of the diagonal elements of normalized confusion matrix as compared to logistic regression.

It is worth noting that a substantial increase in prediction performance is achieved when the decision tree's hyper-parameters are tuned, but the performance increase is much more modest with the random forest. From the results in Table IX, it can be concluded that the classifier with the best prediction performance on our dataset is the tuned random forest, which yields an AUC value of 0.95 and an average of 0.89 accuracy (average of True Negative and True Positive).

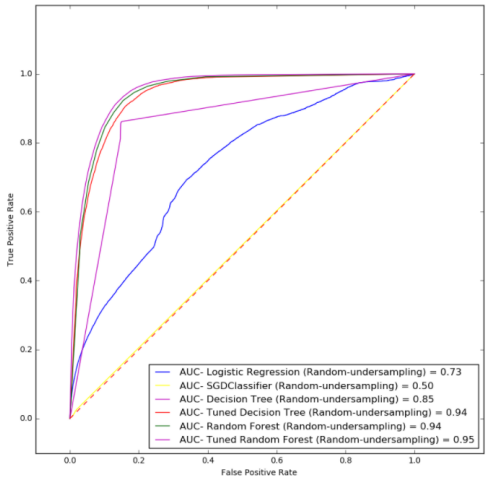

Fig. 11: Area Under Curve (AUC)- Receiver Operating Characteristic (ROC) corresponding to various classifiers, refer Table IX

VI. CONCLUSIONS & FUTURE WORK

The work presented here addresses a real-world problem using real data. A system is proposed which can analyze and predict the occurrence of wildfires in an area on a pixel-by-pixel basis. This is achieved with an AUC-ROC of 0.95 which can predict fire potential of an area when its landscape features and fire history are provided.

The steps in the data preparation process are described in detail. Preparing geospatial data for analysis is not a trivial task, requiring many complex steps. By providing the details of this process, we aim to offer a reference point for other researchers interested in working with this type of data.

The imbalance problem of the dataset is addressed by testing various resampling techniques, and it is determined that the undersampling technique works best for this dataset.

Several classification models, with different hyper-parameter configurations, are evaluated on the balanced dataset. The best performing model is tuned random forest which provides an AUC-ROC score of 0.95.

The proposed system aims to predict the occurrence of wildfires, and is a step towards assessing the risk of future wildfires. There are several ways to extend this research.

The first would be including more data from other sources into the dataset. One clear source is weather related. High temperature, low humidity, wind speed, etc. can create high fire-risk conditions. It would be useful therefore to include weather variables for prediction. Human proximity is another variable that should be studied. Many wildfires are started by human activity, so proximity to human artifacts, such as buildings and roads, can be a factor in the likelihood of fire ignition of an area.

Improvements can also be made in the techniques used in the system. We have shown here that undersampling is

TABLE IX: Scores using various classifiers (with undersampling)

Classifier	True Negative	False Positive	False Negative	True Positive	AUC-ROC
Logistic Regression	0.65	0.35	0.3	0.7	0.73
SGDClassifier (log loss)	0.98	0.02	0.98	0.02	0.50
Decision Tree	0.85	0.15	0.14	0.86	0.85
Decision Tree (tuned)	0.85	0.15	0.1	0.9	0.94
Random Forest	0.87	0.13	0.11	0.89	0.94
Random Forest(tuned)	0.87	0.13	0.09	0.91	0.95

effective in addressing the class imbalance problem. There are other methods for class imbalance such as class weighting that should also be investigated. Additionally, more complex models such as neural networks and gradient boosted trees should also be studied for prediction.

The system described in this paper was build using approximately 35.22% of San Diego County. Incoporating data from more areas, with different topography, vegetation, and weather conditions, would make the system more robust. Using data from different years will also integrate more variability into the training data, which should also increase prediction robustness.

The proposed system can also be used to analyze wildfire trends. Understanding which areas are susceptible to wildfires, and why, would be invaluable in formulating fire prevention strategies.

ACKNOWLEDGEMENTS

This work was supported in part by NSF 1331615 under CI Information Technology Research and SEES Hazards and NSF 1636879 under BD Spokes - Big Data Regional I programs.

REFERENCES

[1] National Interagency Fire Center, https://www.nifc.gov/fireInfo/fireInfo_stats_totalFires.html

[2] John T. Abatzoglou, A. Park Williams. Impact of anthropogenic climate change on wildfire across western US forests. Proceedings of the National Academy of Sciences, 2016; 201607171 DOI: 10.1073/pnas.1607171113

[3] Jolly WM, Mark A Cochrane, Patrick H Freeborn, Zachary A Holden, Timothy J Brown, Grant J Williamson, David MJS Bowman, (2015) Climate-induced variations in global wildfire danger from 1979 to 2013. Nat Communications

[4] Landfire- Data Downloads, Feb 2017, https://landfire.gov/viewer/

[5] FRAP- AboutUs, http://frap.fire.ca.gov/about/index

[6] GDAL. 201x. GDAL - Geospatial Data Abstraction Library, Open Source Geospatial Foundation, http://gdal.osgeo.org

[7] Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning, Journal of Machine Learning Research, 2017, vol 18, p1-5

[8] Scikit- learn: Machine Learning in Python, Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E., Journal of Machine Learning Research, 2011, p2825–2830

[9] P. Cortez and A. Morais. A Data Mining Approach to Predict Forest Fires using Meteorological Data. In J. Neves, M. F. Santos and J. Machado Eds., New Trends in Artificial Intelligence, Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence, December, Guimarães, Portugal, pp. 512-523, 2007. APPIA, ISBN-13 978-989-95618-0-9.

[10] Daniela, Stojano, Panče, P., Andrej, K., Sašo, D., Katerina, T., 2006. Learning to predict forest fires with different data mining techniques. In Conference on Data Mining and Data Warehouses (SiKDD 2006), Ljubljana, Slovenia.

[11] Canadian Wildland Fire Information System- Natural Resources Canada, Last modified at July 18, 2017 http://cwfis.cfs.nrcan.gc.ca/background/summary/fwi

[12] Wikipedia- San Diego County, 18 July 2017 https://en.wikipedia.org/wiki/San_Diego_County,_California

[13] LANDFIRE, U.S. Department of Agriculture and U.S. Department of the Interior, Accessed May, 2017 at https://www.landfire.gov/topographic.php

[14] National Avalanche Center- Aspect http://www.fsavalanche.org/aspect/

[15] Anderson, H. E. 1982. Aids to determining fuel models for estimating fire behavior. Gen. Tech. Rep. INT-122. Ogden, UT: U.S. Department of Agriculture, Forest Service, Intermountain Forest and Range Experiment Station.

[16] Correlation matrix. A.V. Prokhorov (originator), Encyclopedia of Mathematics. http://www.encyclopediaofmath.org/index.php?title=Correlation_matrix—&oldid=19066

[17] Hunt, Neville;Tyrrel, Sidney (2001)- Stratified Sampling, web page at Conventry University

[18] Kohavi, R., and Provost, F. 1998. On Applied Research in Machine Learning. In Editorial for the Special Issueon Applications of Machine Learning and the Knowledge Discovery Process, Columbia University, New York, volume 30.

[19] Stehman, Stephen V. (1997). "Selecting and interpreting measures of thematic classification accuracy". Remote Sensing of Environment. 62 (1): 77–89

[20] H Wang, H Zheng- "True Positive Rate", Encyclopedia of Systems Biology, 2013 - Springer New York

[21] Burke DS, Brundage JF, Redfield RR, et al. Measurement of the false positive rate in a screening program for human immunodeficiency virus infections.N Engl J Med 1988;319:961-4.

[22] Metz CE(1978) Basic Principles of ROC Analysis, https://www.ncbi.nlm.nih.gov/pubmed/112681

[23] Zweig, MH, Campbell, G. Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. Clin Chem 1993;

[24] Tape TG. Interpreting diagnostic tests, ROC Curves. [Accessed May, 2017];2010 [University of Nebraska medical center web site]. Available at: http://gim.unmc.edu/dxtests/roc3.htm.

[25] Quinlan, J.R., Simplifying decision trees. International Journal of Man-Machine Studies, 27, 1987.

[26] Loh, Wei-Yin. (2011). Classification and Regression Trees. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 1. 14 - 23. 10.1002/widm.8.

[27] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP: SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002, 16: 341-378.

[28] G. Batista, R.C. Prati, M.C. Monard A study of the behavior of several methods for balancing machine learning training data SIGKDD Explorations, 6 (1) (2004)

[29] McCullagh, P., Nelder, J.A. (1989). Generalized linear models. CRC press, isbn: 9780412317606, https://www.crcpress.com/Generalized-Linear-Models-Second-Edition/McCullagh-Nelder/p/book/9780412317606

[30] Breiman, L., Random Forests, Machine Learning (2001) 45: 5. https://doi.org/10.1023/A:1010933404324

[31] Ho, Tin Kam (1995). Random Decision Forests. Procedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC