Project Design Document

Project Option #2: Wildfire Risk Analysis and Prediction

Team Project - CMPE 257

Instructor: Dr. Jerry Gao
Semester: Spring, 2019

Project Group #6
Team Members:

Name	Email	SJSU ID
Akshay Kumar Gyara	akshaykumar.gyara@sjsu.edu	013836277
Shalu Rani	shalu.rani@sjsu.edu	012565046
Nasrajan Jalin	nasrajan.jalin@sjsu.edu	013769665
Nithya Kuchadi	nithya.kuchadi@sjsu.edu	013769665

Table of Contents

1. Introduction

Wildfires are the most destructive and dangerous accidents that occur across the US especially in California. The numbers of catastrophic wildfire incidents are increasing every year causing a lot of damage to the environment, lives, and property. The nation has been spending more than $1 billion to suppress such accidents from the past 10 years.

Wildfires are both difficult to fight and predict. Difficulty in prediction is due to a number of different factors. Combination of fire, vegetation, climate, humans, and weather make it difficult to predict where wildfires might start and how they may spread. During summers large quantities of dried leaves, grass, deadwood, and small plants fuel the fires in the forest. The wind and the terrain are also key factors that help in spreading the fire across the forest. Relative humidity readings and air temperature forecasts are general conditions that can be an indication of fire danger across the entire fire ground.

With such a high level of danger, social and economic loss caused by the uncontrolled wildfire motivated us to take up this project. This project aims to develop real-time, simulation data-driven prediction and visualization of wildfire. The major objective of this project is to obtain a very high accuracy to predict fire in a county or region given with no of inputs parameters like temperature, humidity, and spatial imagery of that particular county etc.

Most of the existing models use either satellite data or sensor data of weather and terrain for fire prediction but we have chosen to train the model using all the 3 dimensions i.e. satellite image data, weather data and terrain data of a specific county which will yield more accurate results in the prediction of fire. As smoke is a key element of fire, it can be used for early detection of fire. In this project, we are going to use spatial images of a particular county which has both signs of smoke and no sign of smokes to train the model.

We have chosen 4 different classification models Convolution Neural Network, Support Vector Machine, Residual Neural Networks, and Random Forest which has been discussed in detail in section 4.

The outcome of this project is to predict possibility of fire for a county, on a given day.

2. Related Work

This section provides a review on the articles that propose various algorithms that predict wildfire risk. The data for most of the research till now is limited to one or two dimensions and also to an accuracy up to 75%.

In the paper '*Forest smoke detection using CCD camera and spatial-temporal variation of smoke visual patterns'* by Joon Young Kwak, Byoung Chul Ko, and Jae-Yeal Nam, a new forest smoke detection method is proposed. In this paper, spatial-temporal visual

features are extracted from camera images and the moving regions are detected by analyzing the frame difference between two consecutive key frames. They have extracted intensity, wavelet coefficients, and motion orientation as visual features and used random forests for smoke verification process with four smoke classes and they have achieved better detection performance. The limitation for this paper is they have considered only one dimension i.e. smoke.

In the paper *A Data Mining Approach to Predict Forest Fires using Meteorological Data* by Paulo Cortez and Anibal Morais, good accuracy is achieved for the prediction of small fires. In this paper, they have considered sensor data of weather and used SVM model. But this paper has a limited scope for large fires.

In the paper '*Satellite-Based Prediction of Fire Risk in Northern California*' by Caroline Famiglietti, Natan Holtzman, and Jake Campolo, fire risk prediction and analysis is done based on local climate and land surface conditions. In this paper, they have applied logistic regression with the forward stepwise selection, decision trees with gradient boosting, and a multilayer perceptron to a robust dataset of ecological, hydrological, and meteorological variables derived from remote sensing and achieved an accuracy of 75-80%. The limitation for this paper is they have considered only satellite data and two dimensions i.e., weather and land surface conditions.

3. Training and Test Data Preparation

3.1 Data Types considered:

To analyze the data in full scope in all the dimensions, we will be considering following kinds of data:

3.1.1 Satellite Images:

As satellite data covers large areas and helps in the identification of extremely large features. Satellite images provide insights about the vegetation changes over season and the change in human activity over a terrain. These types of data contains information related to the state of crops (NDVI: Normalized Difference Vegetation Index), meteorological conditions (LST: Land Surface Temperature) as well as the fire indicator "Thermal Anomalies". These aspects can be leveraged to predict the possibility of fire in a place at a particular time.

Fig 3.1.1.1: Satellite images of smoke and fire

Fig 3.1.1.2: Satellite image in grid format

3.1.2 Sensor Data:

Since smoke appears before flames, remote sensing that can detect smoke is a good indicator for early fire warning systems.

3.2 Data Pre-Processing:

1. **Handling missing values:**

   To handle missing values in case of sensor data/csv data, we'll be using imputer with appropriate strategy. In case of non-significant features, we can fill zeroes,

otherwise we'll be taking the mean value. Depending upon the type of feature, its significance and impact on final results, we may vary the strategy.

We will also be handling null and duplicate values.

2. **Dimensionality Reduction:**

   For better performance, we removed attributes that are least correlated with the data.

   Image data will be resized for dimensionality reduction. Feature reduction techniques like PCA and SVD will be used.

3. **Normalization:**

   The numerical attributes will be normalized using mean and standard deviation. Image data will also be normalized.

4. **Hold out:**

   80% of the data will be retained for training the models. 10% is will be used for validation and 10% will be kept aside as test data. We will be using Scikit-learn's train_test_split function for splitting the test and train data. If we see any chance of bias to get introduced, we'll be using the stratified sampling to make sure that the test set is representative. If any of the models has to be fine-tuned, k-fold cross validation will be used.

3.3 Different dimensions considered:

The frequency of large wildfires is influenced by a complex combination of natural and human factors. Temperature, soil moisture, relative humidity, wind speed, and vegetation (fuel density) are important aspects of the relationship between fire frequency and ecosystems.
The overall data set is a combination of satellite and sensor data depicting the below 3 dimensions -

1. Weather
2. Terrain
3. Fire- historic data

The following section briefly explains the data analysis and visualization done for the weather dimension.

3.3.1 Weather

https://interwork.sdsu.edu/fire/map_gallery/images/Fires_all-2003-2007.gif

Fig 3.3.1.1: Temperature variations with dates

The goal of this dimension is to provide data analytics and insights into the relationship between weather and the possibility of wildfire at San Diego County. The weather data were obtained from various sources. These are the steps followed in preparing the data set.

a) Collected data from various sources
b) Analyzed data for correctness
c) Cleaned data
d) Visualized data for correlations using Seaborn heatmaps and plots
e) Removed least correlated attributes
f) Labeled the data to 0 (“no fire”) and 1 (“fire”) using fire history of San Diego as a reference.

The structure of the weather data has been given below:

YEAR	MONTH	DAY	TEMP_MAX	TEMP_MIN	DEW_MAX	DEW_MIN	SLP_MAX	SLP_MIN	STP_MAX	...	WDSP_MAX	WDSP_MIN	MXSPD	GUST	MAX	MIN	PRCP	SNDP	FRSHTT	Class
2017	1	1	53.9	24	46.5	24	1013.6	22	1012.4	...	5.4	24	13.0	18.1	60.1	48.9	0.68G	999.9	10000	0
2017	1	2	56.4	24	45.4	24	1018.1	24	1017.1	...	3.9	24	8.0	999.9	60.1	48.9	0.00G	999.9	0	0
2017	1	3	56.4	24	46.3	24	1021.2	23	1020.3	...	1.3	24	6.0	999.9	62.1	52	0.00G	999.9	0	0

Fig 3.3.1.2: Class distribution - Fire and no fire in the year 2017

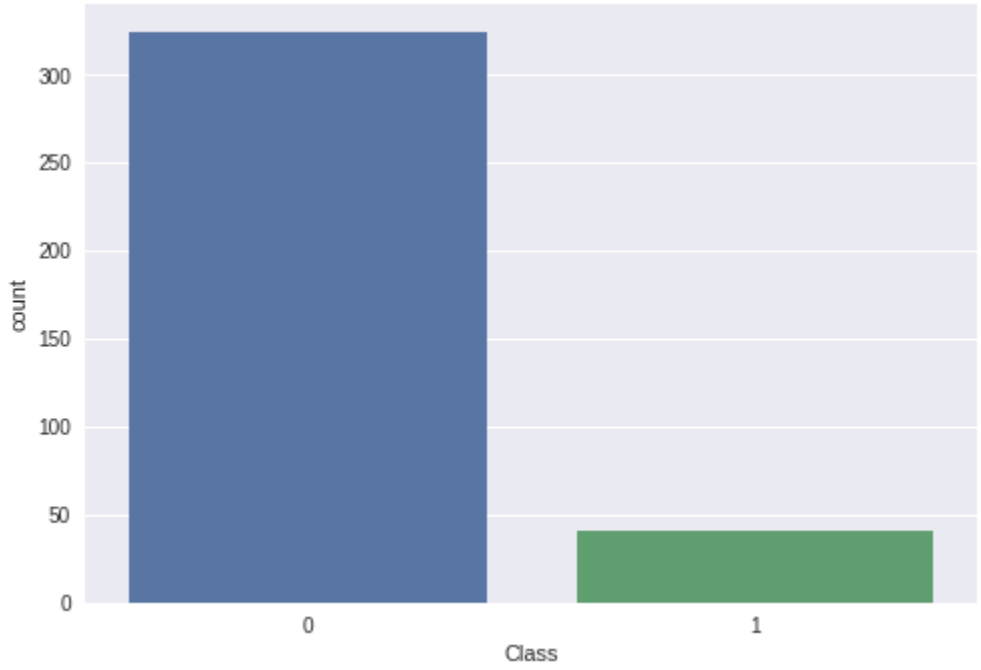


Fig 3.3.1.3: Class count statistics (1 represents fire; 0 represents no-fire)

The below heat map shows different data correlations. Each feature has been mapped with the other features for the same.

Fig 3.3.1.4: Data correlation matrix

The above mapping provides clarity regarding the unimportant features. The below matrix shows heat Map after removing less correlated attributes.

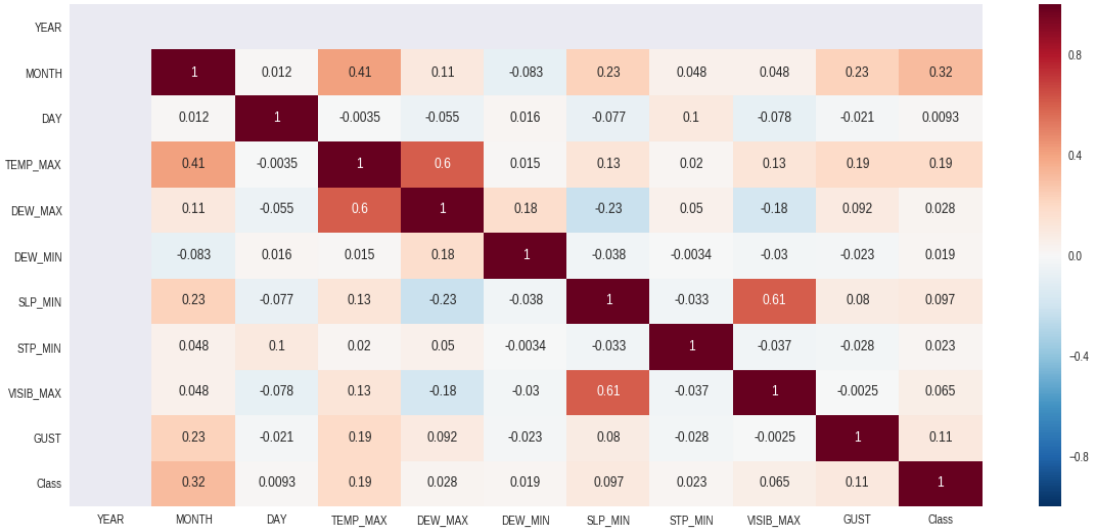


Fig 3.3.1.5: Data correlation matrix after removing non-correlated features

Relation between Wildfire and weather:

According to the data driven analysis, 3 major parameters are most correlated with wildfire.

1. Maximum Temperature
2. Maximum Gust
3. Month of the year

Temperature

For a county like San Diego, temperature plays a major role in causing fire. Majority of the fire incidents that we analyzed, happened when the temperature was above 60F. Higher temperature and drier conditions due to global warming is a fuel for wild fire.

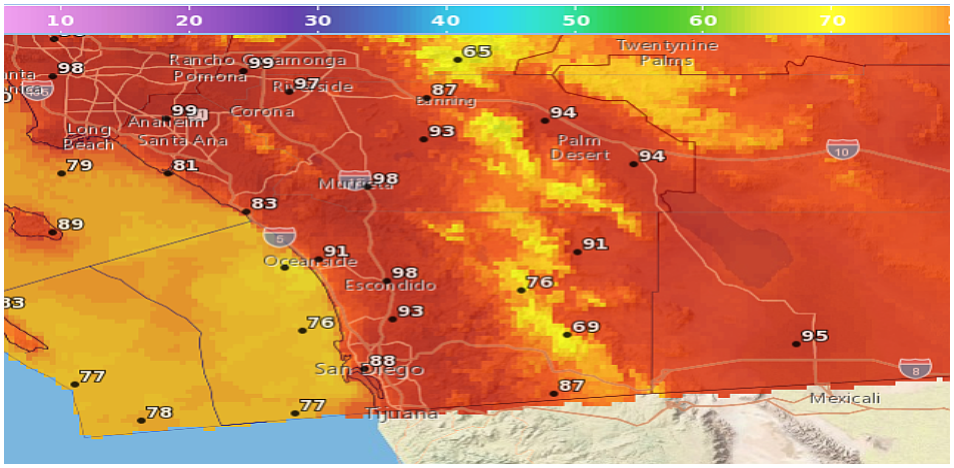


Fig 3.3.1.6: Temperature heatmap (directly impacts fire ignition)

Maximum Gust

Wind speed is another factor which directly impacts the chance of wildfire. Wind causes friction between natural elements like twigs and cause the initial ignition at times. It also accelerates the spread of wildfire, once ignited.

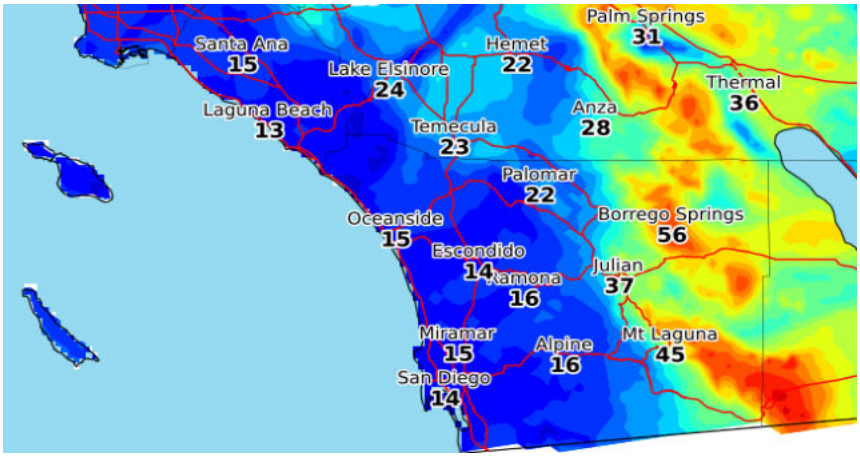
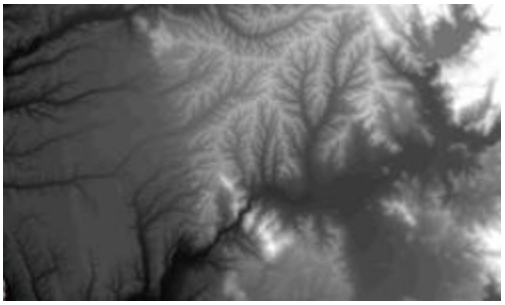


Fig 3.3.1.7: Wind speed factors (impacts fire spread)

3.3.2 Landscape and Vegetation

The accumulated biomass in the form of dead leaves, branches etc. makes a huge impact in the ignition and spread of fire. The landscape, soil moisture content, type of vegetation in the area influences the probability of fire occurrence.

Vegetation Types and structural characteristics



Topographic Information

Fig 3.3.2.1: Satellite Terrain images

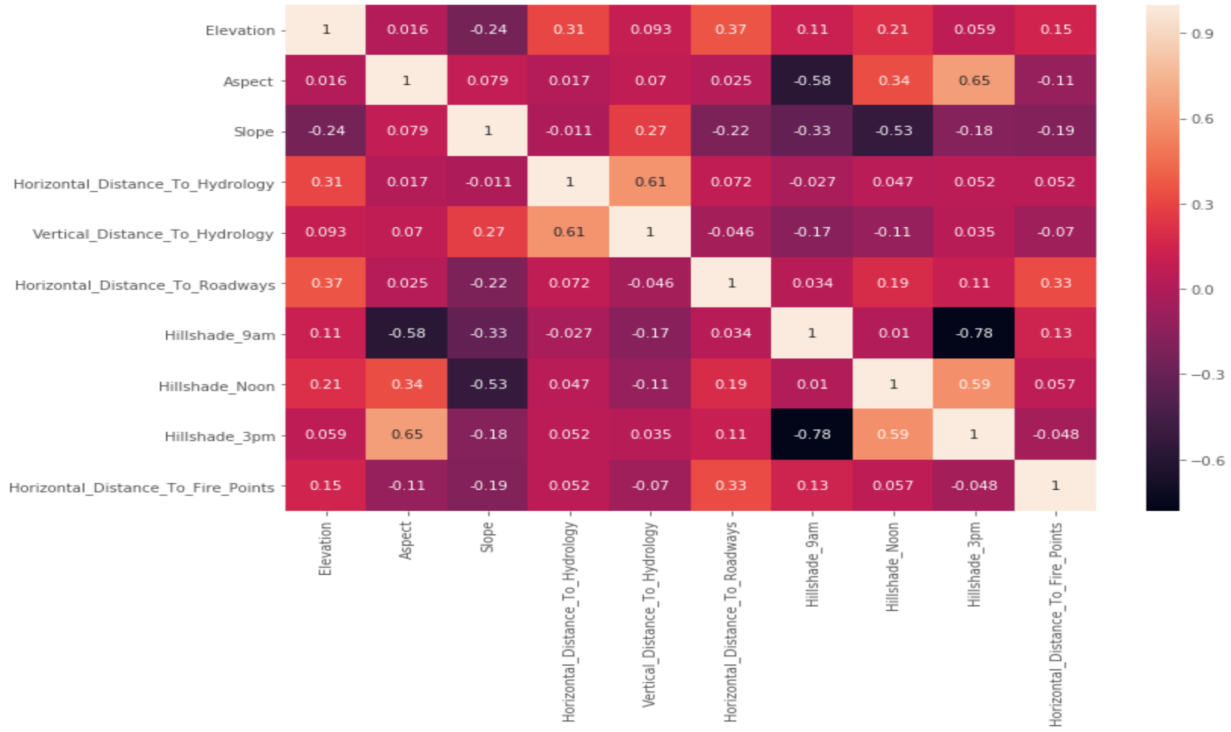


Fig 3.3.2.2: Correlation matrix with terrain, vegetation features

3.3.3 Fire-historic data

For analyzing and labeling any fire, the historic data is essentially required. National center for environmental information, ca.gov etc. websites provide historical data about

fire incidents. The historic data will be used for labeling the dataset in different dimensions, especially fire. One special insight we obtained by examining the history was, every fire was followed by a severe weather alert.

3.3.4 Human Activity

Human beings have a great impact on fire regimes because they alter ignition frequency and fuel fragmentation and suppress fires. The analysis of human factors in fire ignition is widely recognized s of critical importance. This dimension includes rate of urbanization, deforestation, presence of fire fueling parameters in the area etc. Data for the human variables includes fine-resolution maps of the WUI (wildland-urban interface) produced using housing density and land cover data.

3.4 Design Workflow

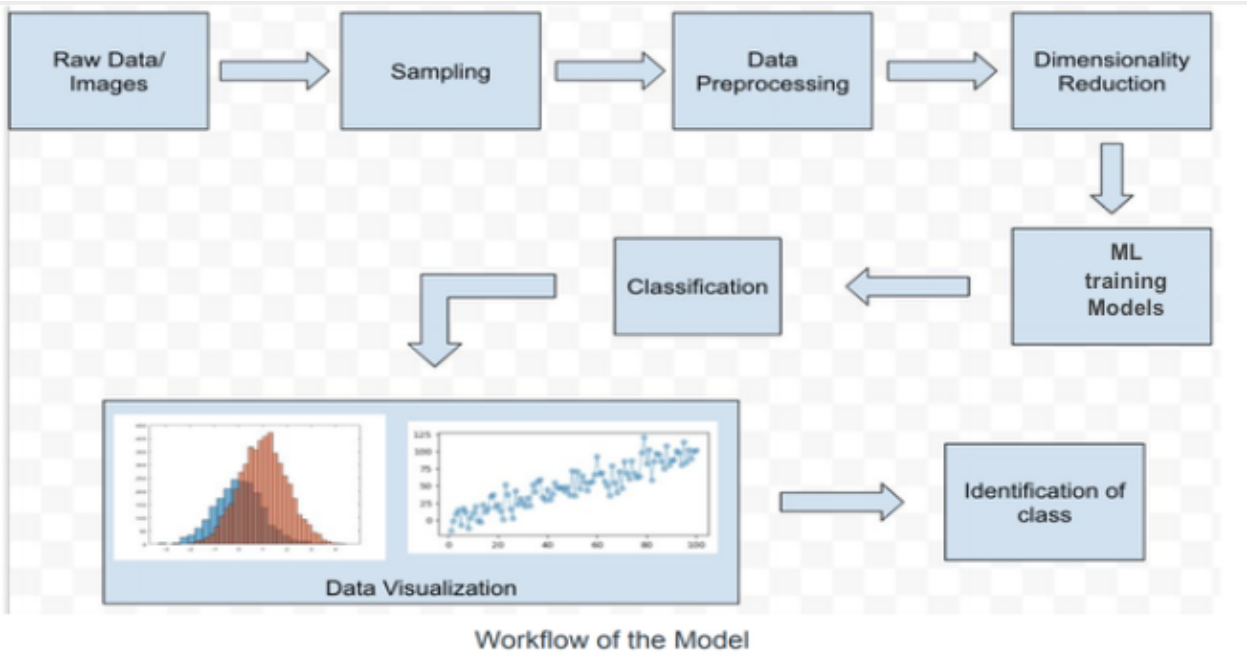


Fig 3.4.1: Design Workflow

4. Machine Learning Models and AI Algorithms

This section presents the selected machine learning models and AI algorithms.

4.1 Convolutional Neural Network (CNN):

Convolutional Neural Network is a class of deep, forward-feed artificial neural networks, most commonly applied to analyzing visual imagery. CNNs use a variation of multilayer perceptron designed to require minimal preprocessing. It takes the input image and assigns learnable weights and biases to the various features in the image. A convolutional network is able to successfully capture the Spatial and Temporal dependencies in an image through the application of relevant filters. The filter is used to detect patterns and slides over the entire image to capture relevant data.

In CNN, input from one layer is passed to the next, wherein each layer is treated as an object that feeds data to the next layer. Below diagram depicts the architecture of a CNN based classification model.

First layer is always convolutional layer, which, based on defined filter detects low-level features such as edges, color, gradient orientation, etc. Subsequent convolutional layers provide and wholesome understanding of images. The convolutional layer output is passed through the activation function ReLU (rectified linear unit) which replaces negative pixels with 0.

Similar to the Convolutional Layer, the Pooling layer reduces dimensionality of each feature map and retains the most important features. A dropout layer may be introduced in between which prevents overfitting. Before sending the output from dropout layer, to dense layer, the multi-dimensional array needs to be flattened using flatten layer. The final layer is the dense layer i.e. fully connected layer where every input is connected to every output by weight. Softmax activation function provides probabilistic distribution in different classes.

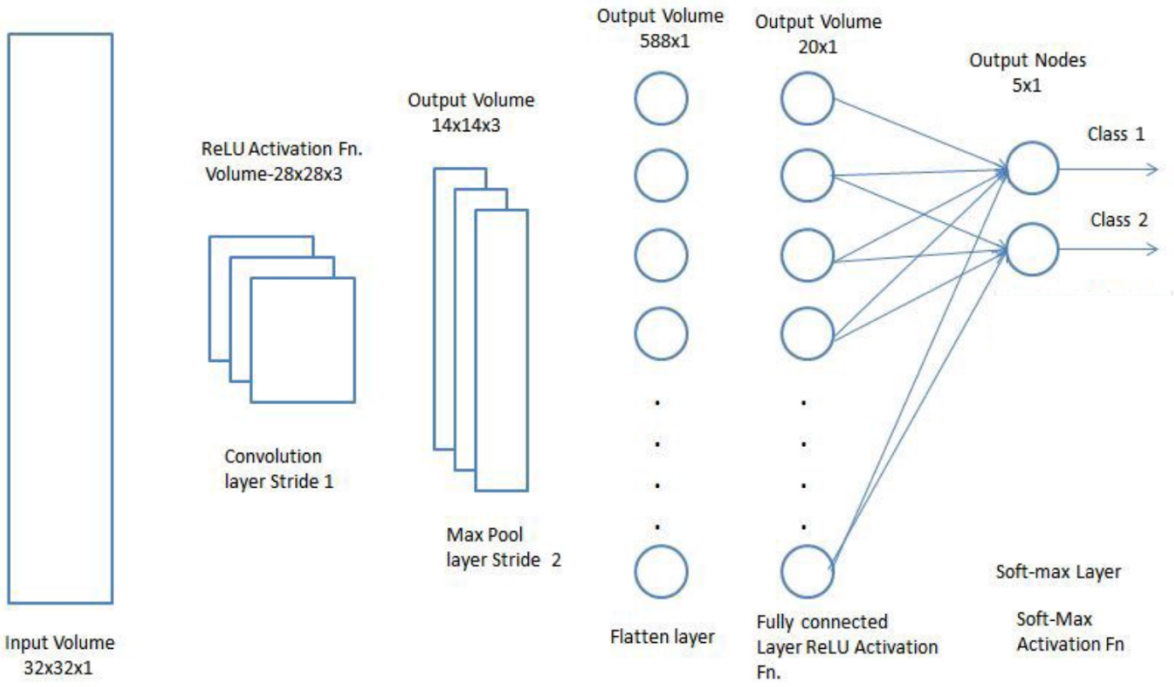


Fig 4.1.1: Convolutional Neural Network

4.2 Logistic Regression:

Logistic regression models the probabilities for classification problems with two possible outcomes. It's an extension of the linear regression model for classification problems.

In regression analysis, logistic regression is predictive analysis used when dependent variable(target) is categorical and dichotomous (e.g. yes/no; fire/no-fire etc.). Logistic regression is named for the function used at the core of the method, the logistic function (also called sigmoid function). In simple words, logistic regression is the estimation of parameters of the logistic model. The actual representation of the model stored in file/memory are the coefficients in the equation. Logistic regression function is defined as below:

$$\text{logistic}(\eta) = \frac{1}{1 + exp(-\eta)}$$

For classification, we prefer probabilities between 0 and 1, so we wrap the right side of the equation into the logistic function. This forces the output to assume only values between 0 and 1.

$$P(y^{(i)} = 1) = \frac{1}{1 + exp(-(\beta_0 + \beta_1 x_1^{(i)} + \ldots + \beta_p x_p^{(i)}))}$$

The coefficients (Beta values) of the logistic regression algorithm must be estimated from the training data. This is done using maximum-likelihood estimation. The best coefficients would result in a model that would predict a value very close to 1 (e.g. fire) for the default class and a value very close to 0 (e.g. no-fire) for the other class.

Graphically, the logistic function is represented as below:

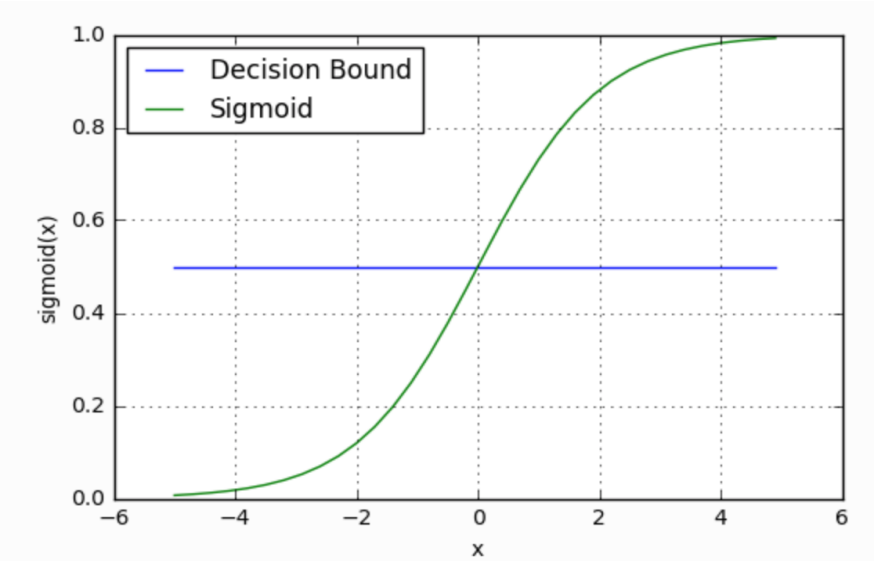


Fig 4.2.1: Logistic function

4.3 Residual Neural Network (ResNet):

A residual neural network is an artificial neural network of a kind that builds on constructs known from pyramidal cells (type of multipolar neurons found in the areas of brain including cerebral cortex, the hippocampus, and the amygdala). Residual neural networks do this by utilizing skip connections. A typical ResNet model will utilize single-layer skips.

ResNet can give high performance even with thousands of layers. ResNet addresses the vanishing gradient problem wherein performance gets saturated.

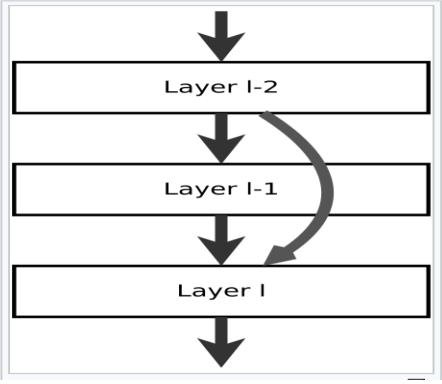


Fig 4.3.1: Canonical form of ResNet

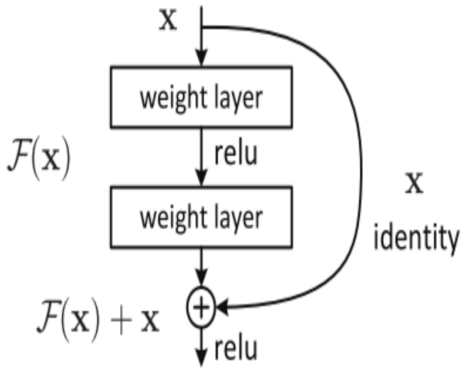


Fig 4.3.2: Residual block

4.4 Random Forest:

Random Forest builds an ensemble of decision trees, by applying the bagging method.

Random Forest is one of the most widely used algorithms and works well with high dimensional data. It is flexible and gives good performance even without tuning the hyperparameters.

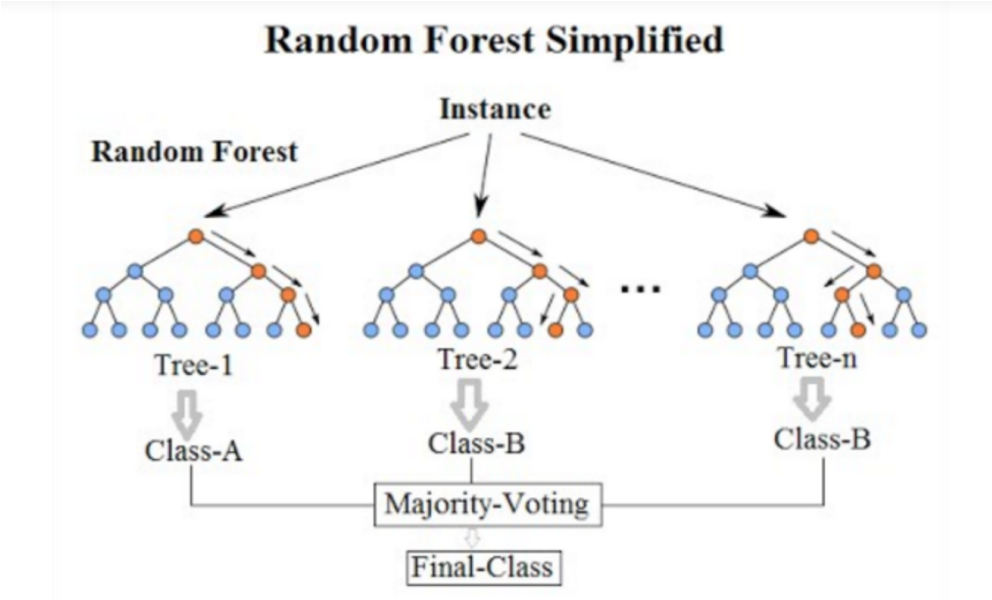


Fig 4.4.1: Random forest representation

Unlike decision trees where the most important feature is used for splitting the node, random forest searches for the best features among a random subset of features.

References:

1. https://en.wikipedia.org/wiki/Residual_neural_network
2. https://earthexplorer.usgs.gov/
3. http://cs229.stanford.edu/proj2018/report/210.pdf
4. https://ieeexplore.ieee.org/document/6054103
5. https://ieeexplore.ieee.org/document/7414773
6. https://www.scientificamerican.com/article/heres-what-we-know-about-wildfires-and-climate-change/