Project Progress Report

**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](mailto:shalu.rani@sjsu.edu) | 012565046 |
| Nasrajan Jalin | [nasrajan.jalin@sjsu.edu](mailto:nasrajan.jalin@sjsu.edu) | 013784420 |
| Nithya Kuchadi | [nithya.kuchadi@sjsu.edu](mailto:nithya.kuchadi@sjsu.edu) | 013769665 |

*Table of Contents*

[*1. Introduction 3*](#_Toc7126565)

[*2. Related Work 3*](#_Toc7126566)

[*3. Selected Models 4*](#_Toc7126567)

[*4. Model Comparisons 8*](#_Toc7126568)

[*5. Model selection justifications 9*](#_Toc7126569)

[*6. Training and Test Data Preparation 11*](#_Toc7126570)

[*7. Model training process 14*](#_Toc7126571)

[*8. Case Study with Statistics 18*](#_Toc7126572)

# 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, 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 and weather 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, Faster R-CNN, Random Forest and Logistic Regression 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. Selected Models

**Model descriptions with references:**

**3.1 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 andTemporaldependenciesinanimagethroughthe 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.

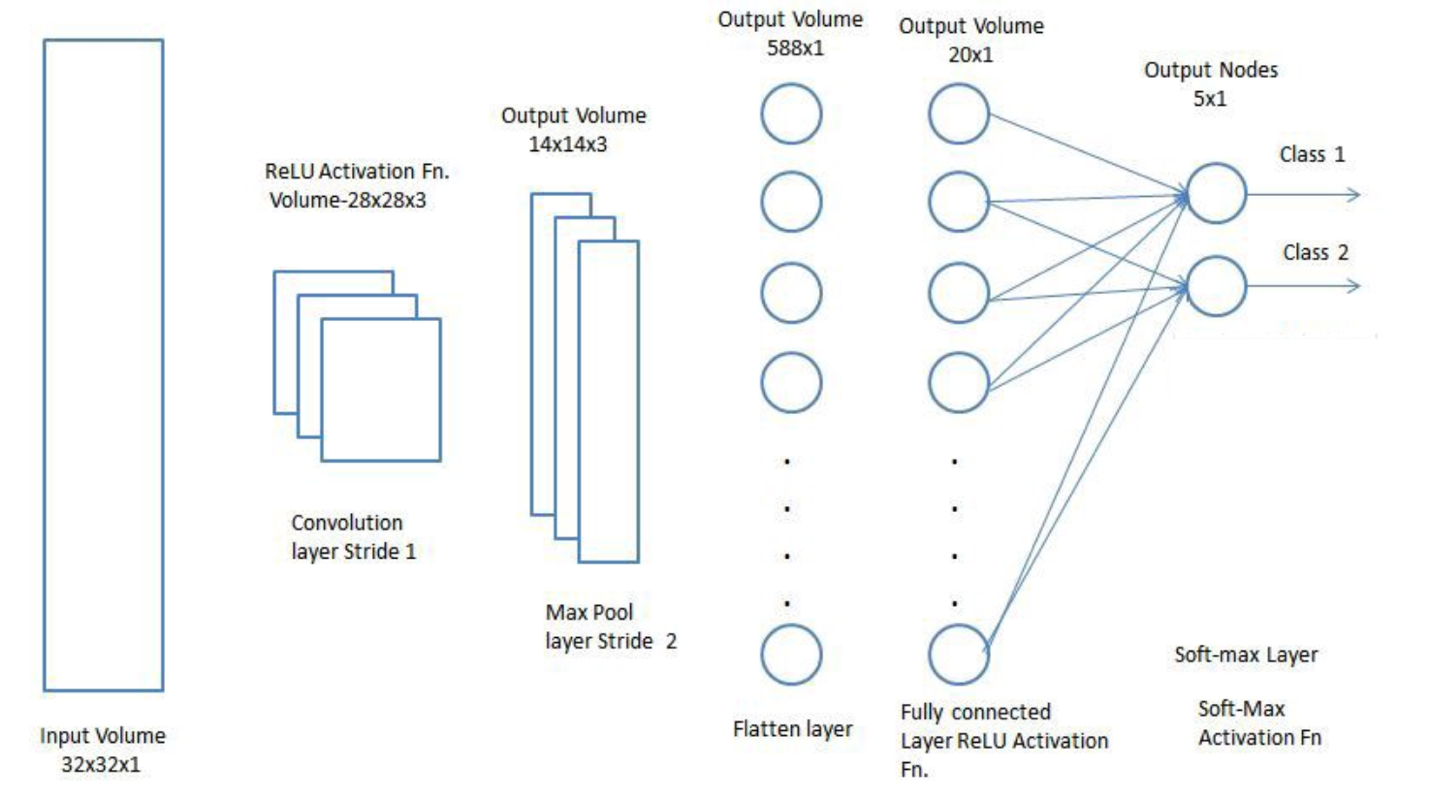
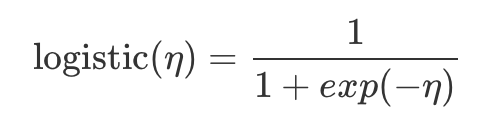


Fig 3.1.1: Convolutional Neural Network

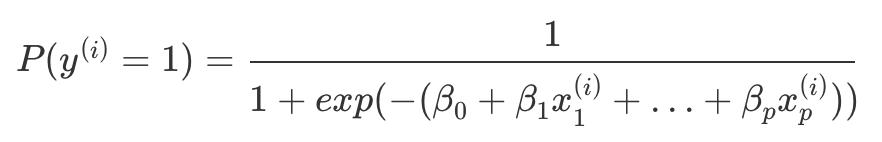
**3.2 Logistic Regression:**

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:



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.



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:

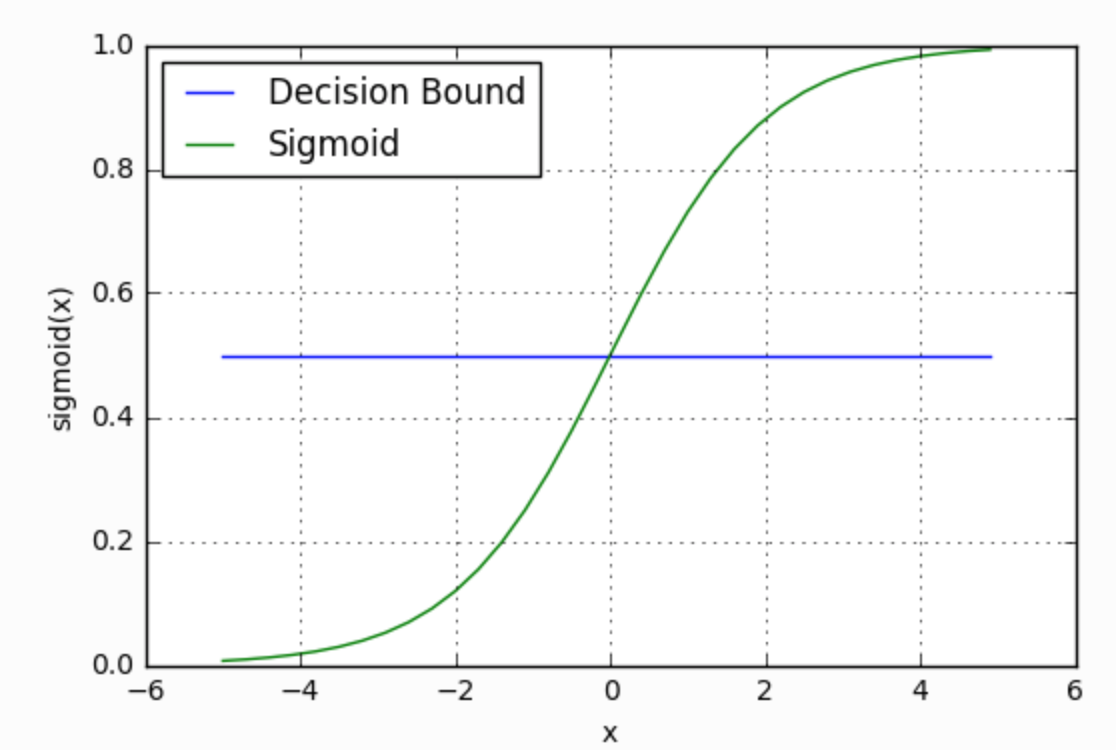


Fig 3.2.1: Logistic function

**3.3 Faster R-CNN with Residual Neural Network (ResNet):**

This model is used for the localized detection of fire and smoke in the satellite images. Most of the object detection algorithms are observed to fail if the size of the object to be detected is very small or varying in size. This is where Faster R-CNN works well. FR-CNN accomplishes faster object detection by employing RPN (Region Proposal Network). The feature map extracted using a ConvNet is passed through an RPN which returns object proposals. Here is a rough comparison of different CNN algorithms in the context of object detection (Reference - analyticsvidhya.com)

# 

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).

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

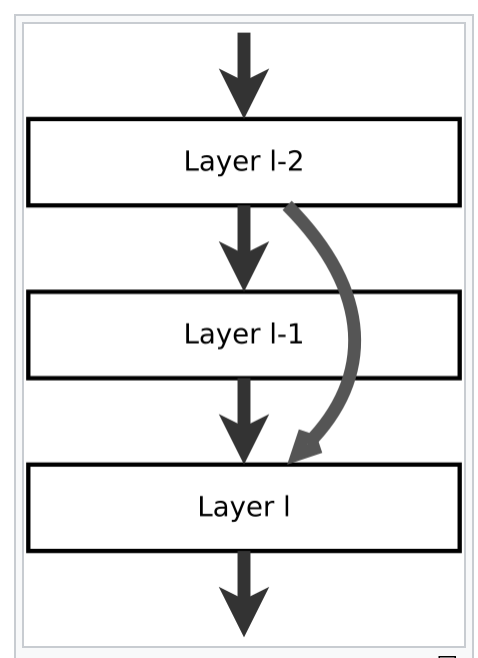
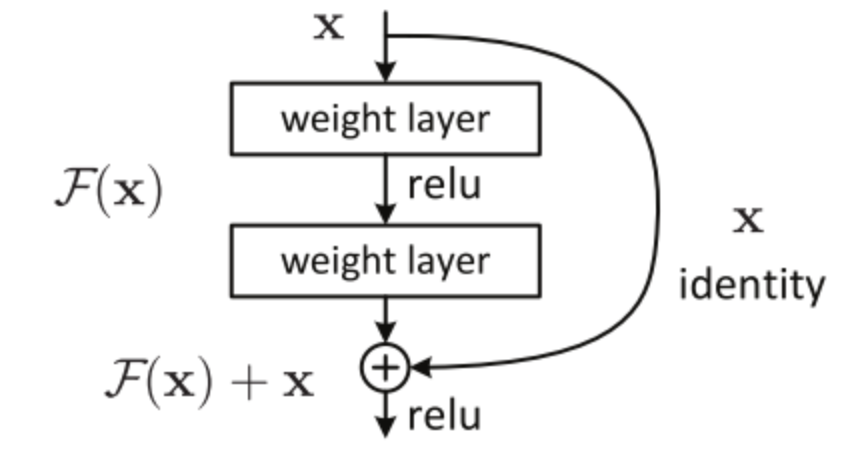
 

Fig 3.3.1: Canonical form of ResNet Fig 3.3.2: Residual block

**3.4 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.

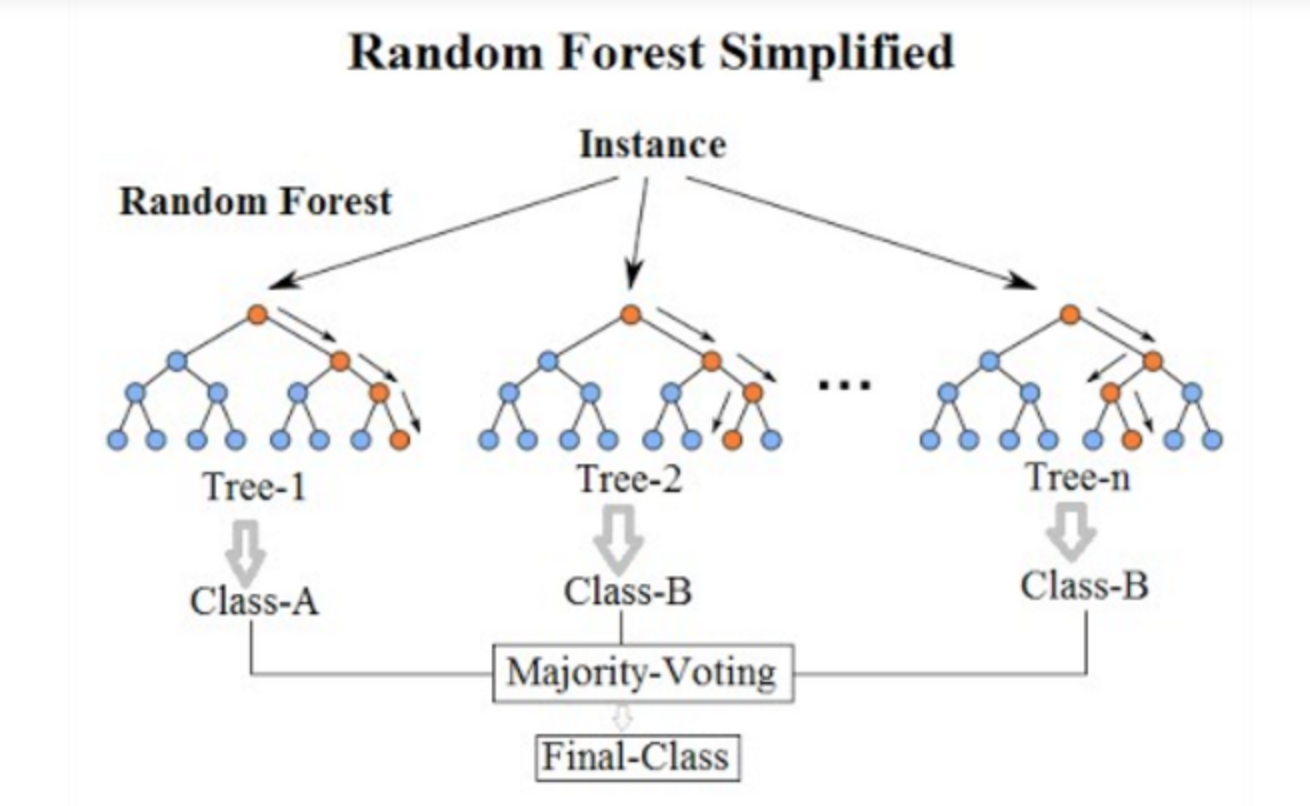


Fig 3.4.1: Random forest representation

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.

# 4. Model Comparisons

Given below are the outline of selected models and current status:

|  |  |  |
| --- | --- | --- |
| **Model** | **Dimension** | **Metrics - Status** |
| Faster R-CNN | Satellite Images | Accuracy for bounding boxes: 80% |
| CNN | Satellite Images | roc value = 0.7883 |
| Logistic regression | Weather | roc value = 0.82 |
| Random Forest | Weather | roc value = 0.7325 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithm** | **Decision Boundary** | **Model Complexity Reduction** | **Limitations** |
| Logistic Regression | linear | L2  Regularization | Do not work well with image data |
| Random Forests | wiggly contours with complex shapes | Low bias and low variance | Accuracy is less compared to CNN |
| CNN | Non-linear | Reduce number of hidden layers, regularization, early stopping | High Computational Cost |

# 5. Model selection justifications

**5.1 Faster R-CNN Algorithm with ResNet**

This model is used for the localized detection of fire and smoke in the satellite images. Most of the object detection algorithms are observed to fail if the size of the object to be detected is very small or varying in size. This is where Faster R-CNN works well. FR-CNN accomplishes faster object detection by employing RPN (Region Proposal Network). The feature map extracted using a ConvNet is passed through an RPN which returns object proposals. Here is a rough comparison of different CNN algorithms in the context of object detection (Reference - analyticsvidhya.com)

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithm** | **Features** | **Prediction time / image** | **Limitations** |
| CNN | Divides the image into multiple regions and then classifies each region into various classes. | – | Needs a lot of regions to predict accurately and hence high computation time. |
| R-CNN | Uses selective search to generate regions. Extracts around 2000 regions from each image. | 40-50 sec | High computation time as each region is passed to the CNN separately. Also, it uses three different models for making predictions. |
| Fast R-CNN | Each image is passed only once to the CNN and feature maps are extracted. Selective search is used on these maps to generate predictions. Combines all the three models used in R-CNN together. | 2 sec | Selective search is slow and hence computation time is still high. |
| Faster R-CNN | Replaces the selective search method with region proposal network (RPN) which makes the algorithm much faster. | 0.2 sec | Object proposal takes time and as there are different systems working one after the other, the performance of systems depends on how the previous system has performed. |

**5.2 Convolutional Neural Networks:**

This model is used for the prediction of fire in the satellite images.

* CNNs work well with data having spatially recurring patterns i.e., images, speech.
* CNNs are able to learn relevant features from an image at different levels similar to a human brain.
* Works well with large amount of data
* Gives good accuracy in image recognition problems
* Most of the other machine learning and AI algorithms do not work well for large amount of image data.

**5.3 Random Forest:**

* Random Forests require almost no input preparation. They can handle binary features, categorical features, numerical features without any need for scaling.
* Random Forests perform implicit feature selection and provide a pretty good indicator of feature importance.
* Random Forests are very quick to train. The random feature sub-setting that aims at diversifying individual trees, is at the same time a great performance optimization.
* Random Forest are not very sensitive to hyper-parameters
* Random Forests can be easily grown in parallel. The same cannot be said about boosted models or large neural networks.

**5.4 Logistic Regression:**

* Logistic regression is used when the dependent variable is binary (0/ 1, True/ False, Yes/ No) in nature.
* In logistic regression, the outcome (dependent variable) has only a limited number of possible values.
* The output of the logistic regression model is more informative than other classification algorithms.
* It is more robust: the independent variables don’t have to be normally distributed, or have equal variance in each group
* It does not assume a linear relationship between the IV and DV
* It may handle nonlinear effects
* You can add explicit interaction and power terms
* The DV need not be normally distributed.
* There is no homogeneity of variance assumption.
* Normally distributed error terms are not assumed.
* It does not require that the independents be an interval.
* It does not require that the independents be unbounded.

# 6. Training and Test Data Preparation

**6.1 Data Types and dimensions considered:**

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

**6.1.1 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 6.1.1.1: Satellite images of smoke and fire

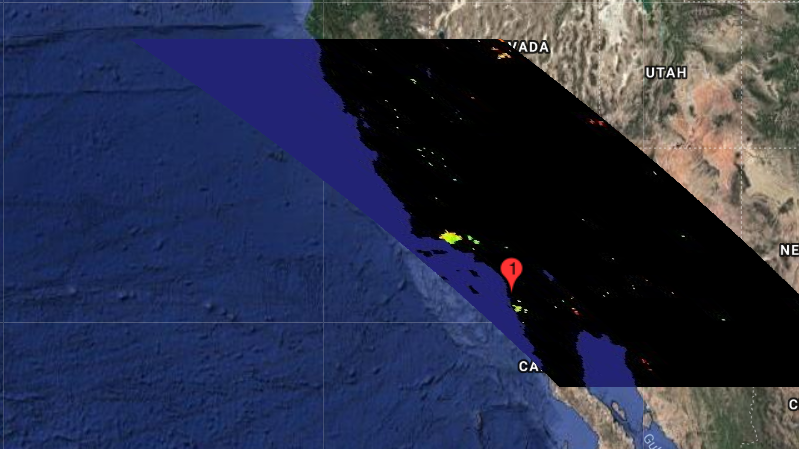
****

Fig 6.1.1.2: Satellite image in grid format

**6.1.2 Weather CSV Data:**

Historical data in CSV format is considered for applying machine learning for predicting fire. The data is fetched from US government weather data site for implementing the initial model. 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.

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.

1. Collected data from various sources
2. Analyzed data for correctness
3. Cleaned data
4. Visualized data for correlations using Seaborn heatmaps and plots
5. Removed least correlated attributes
6. Labeled the data to 0 (“no fire”) and 1 (“fire”) using fire history of San Diego as a reference.

**6.2 Data Pre-Processing:**

1. **Data Labeling:** Every dataset considered was labeled to indicate fire. The satellite data was
2. **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 handled null and duplicate values.

1. **Dimensionality Reduction:**

For better performance, we removed attributes that are least correlated with the data. Image data were resized for dimensionality reduction. Feature reduction techniques like PCA and SVD were used.

1. **Normalization:**

The numerical attributes were normalized using mean and standard deviation. Image data was not normalized because we need to retain all the bands and colors in the satellite data for training purposes.

1. **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.

**6.3 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 2 dimensions -

1. Weather
2. Satellite Images

**6.4 Project Workflow**

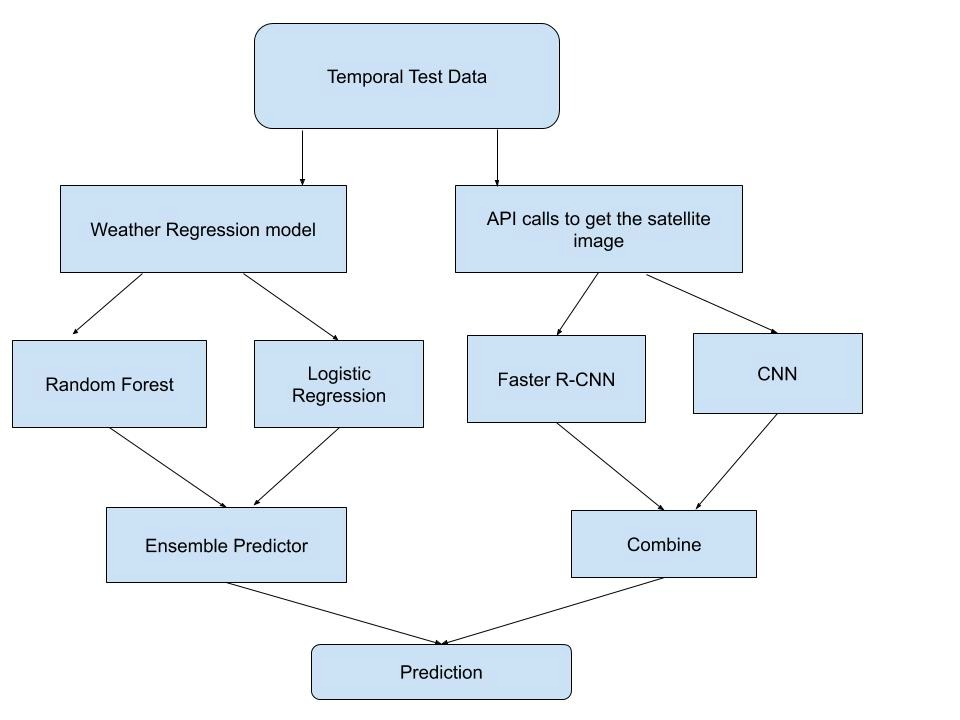
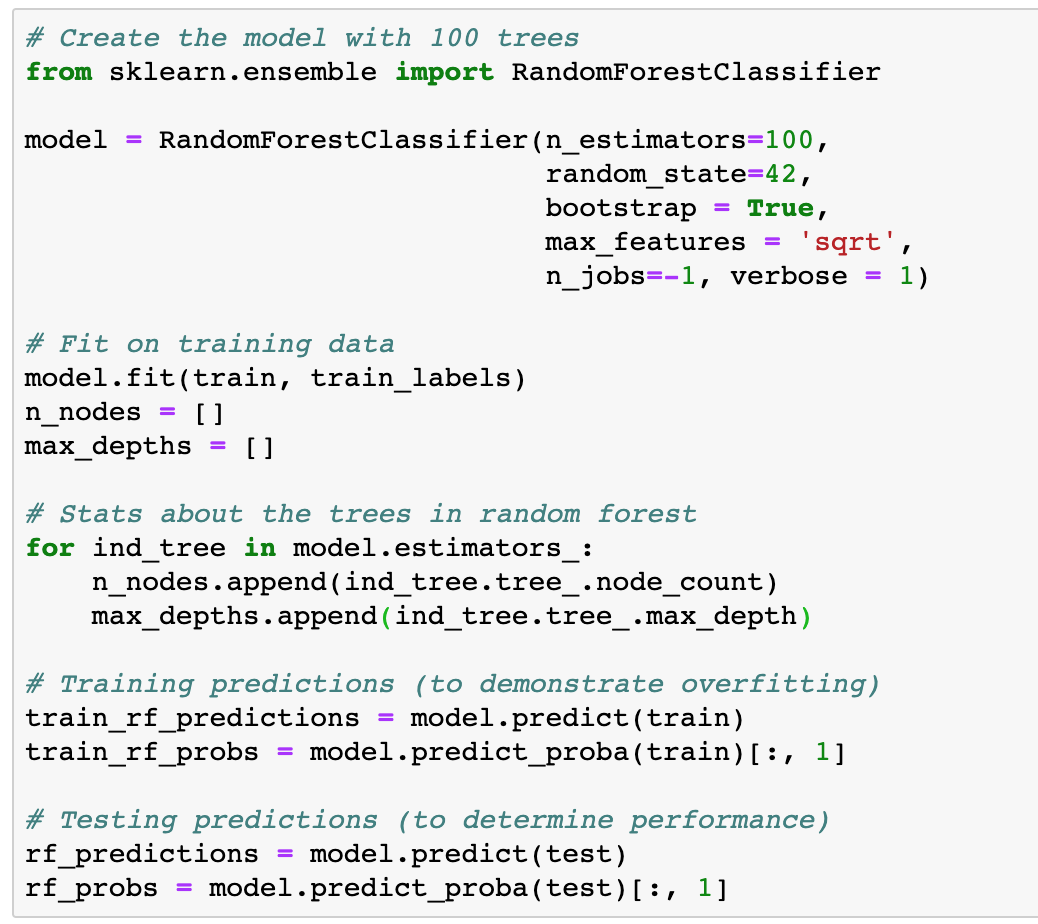
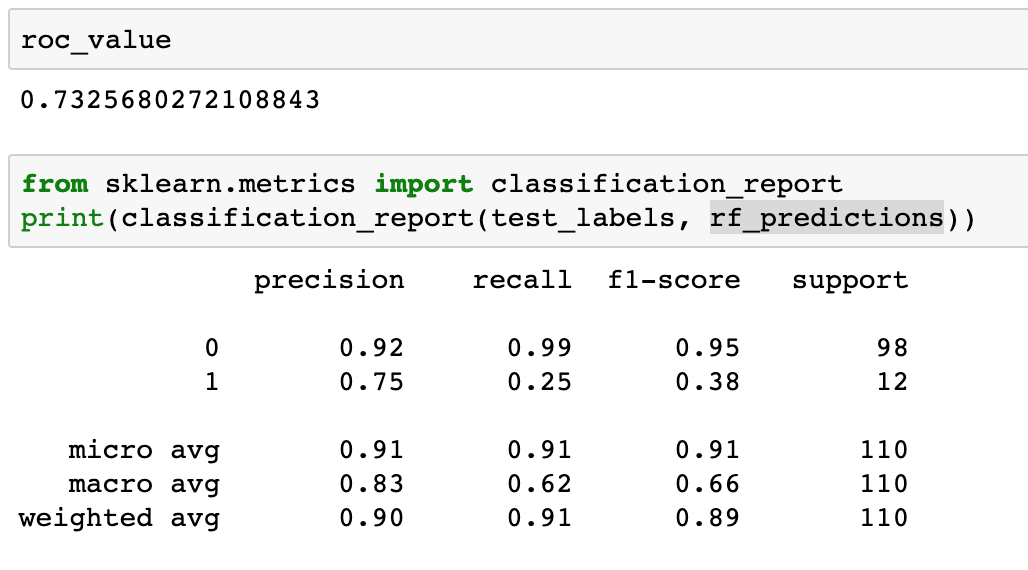


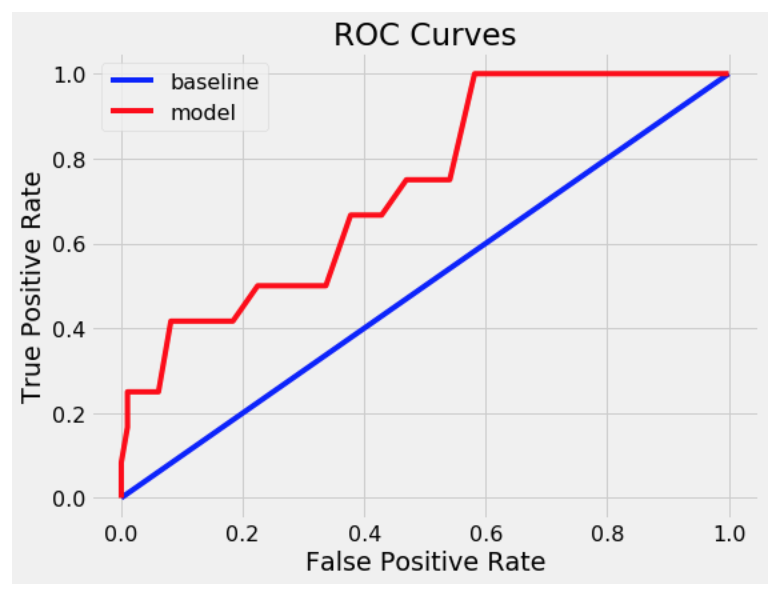
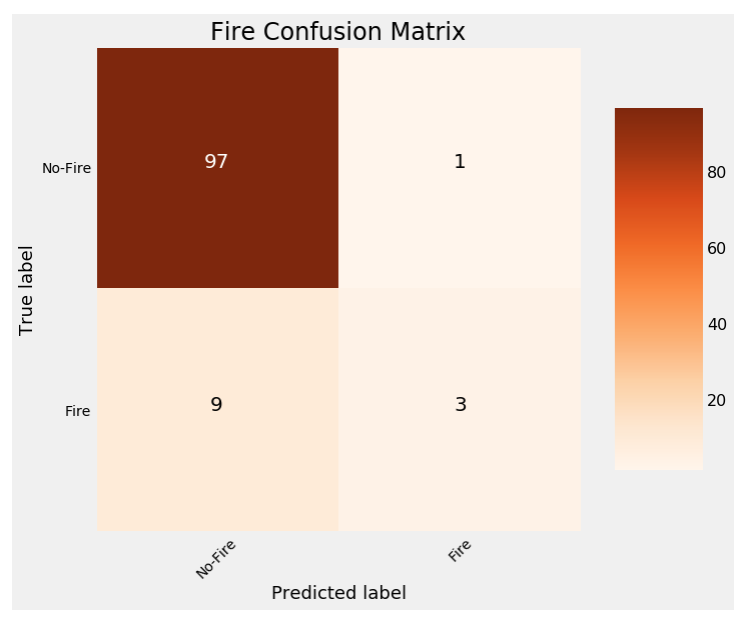
Fig 6.4: Project Workflow

# 7. Model training process

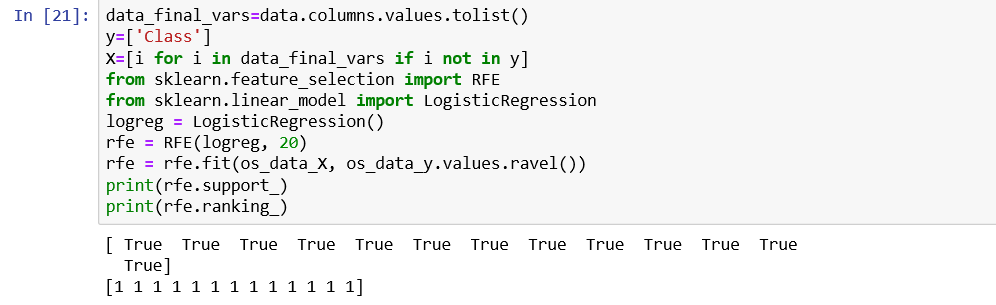
**7.1 Random Forest implementation**

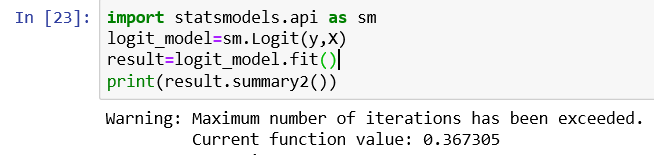


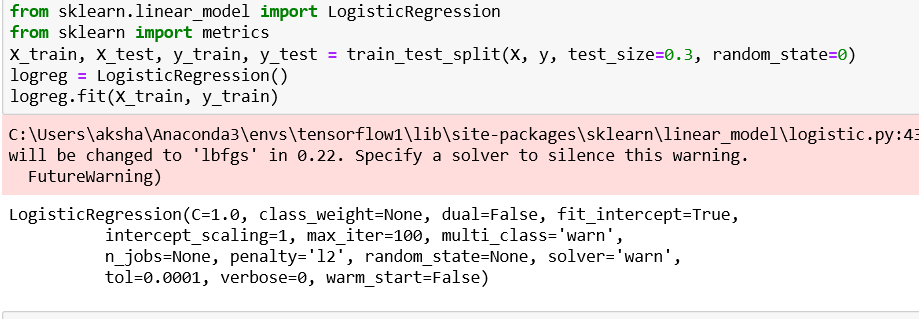


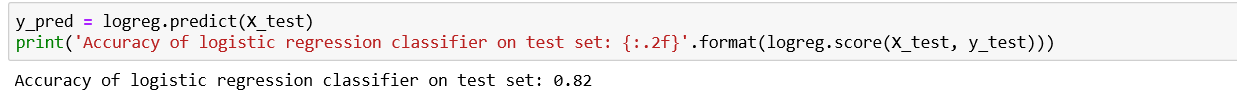
 

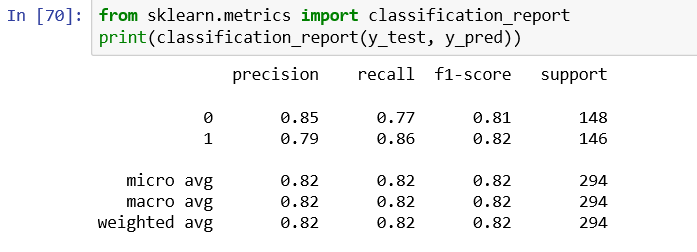
**7.2 Logistic Regression implementation**

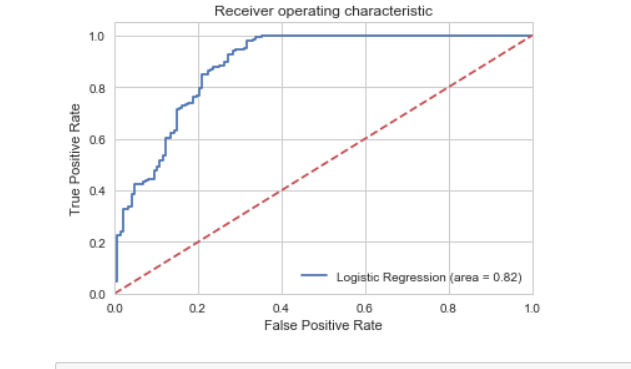












**7.3 CNN implementation**

No of Epochs: 50

Batch Size:50

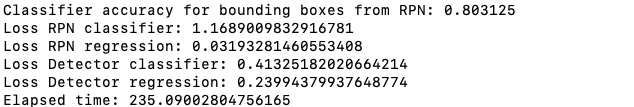


**7.4 Faster R-CNN**

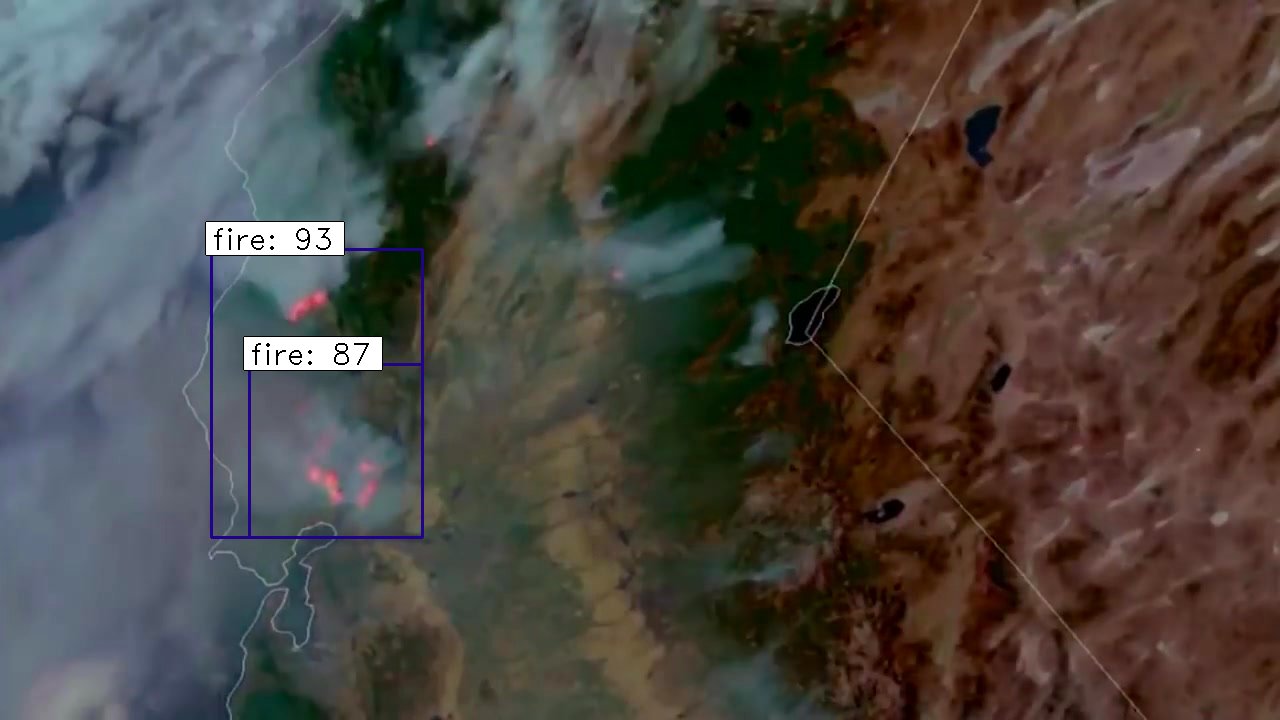
No of Epochs: 50

Epoch Length: 10

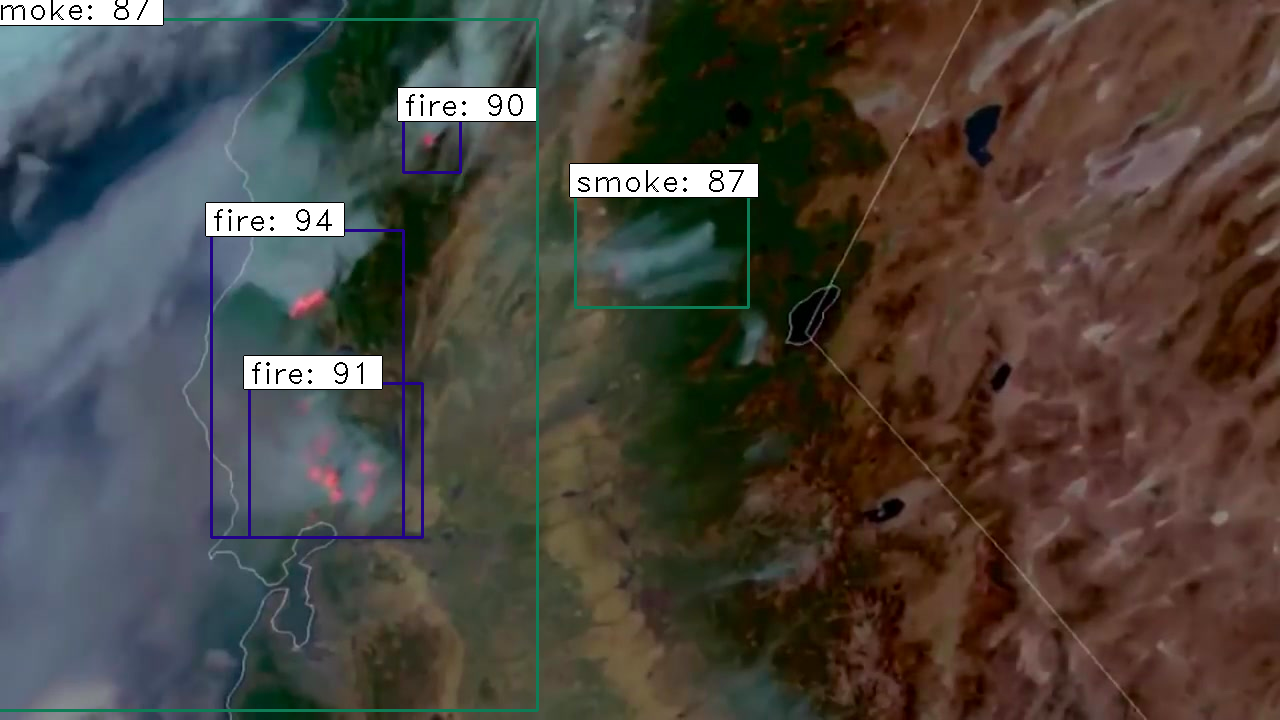
Base Network: Resnet 50



1. Fire localization



2. Smoke localization



3. No Fire

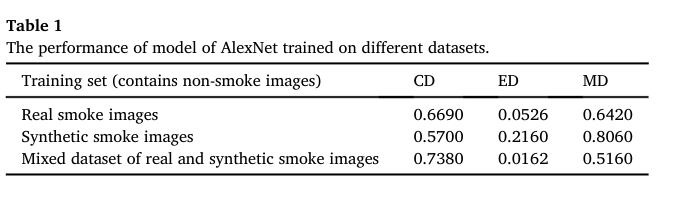


# 8. Case Study with Statistics

**Case Study - Early Fire detection using weather data:** Detecting and predicting fire at early stages is very crucial because fire spreads at a very fast pace. The most effective way to minimize the damages caused by the forest fires is the early detection of forest fires and a fast-appropriate reaction. Many countries have their own indices to measure and react to the wildfire occurrences. One of the well-known indices is Canadian Fire Weather Index. It uses fire weather modeling, as a basis for modeling and displaying of the possibility of fire occurrence in the forests. The calculation of the components is based on consecutive daily monitoring of the temperature, the relative humidity, the wind speed, and the 24-hour rainfall.

Since USA do not have a fire weather index unlike other countries, the dependency between different weather parameters need to be analyzed with the help of machine learning models and needs to reach the best predictor.

**Case Study - Smoke Detection in Satellite Data:** While detecting satellite data, it is very easy to mistake between smoke and cloud distribution. Since smoke appears before fire, it is indeed a good indication of early fire prediction too. To distinguish between smoke and cloud, a good method to adopt is, to train smoke patterns with vegetation or land as backdrop. The paper *Deep domain adaptation-based video smoke detection using synthetic smoke images**by Gao Xu, Yongming Zhang, Qixing Zhang, Gaohua Lin, Jinjun Wang uses* synthetic smoke on green backdrops to simulate the effect of smoke on vegetation. Following is the statistics of their approach.



The test accuracy obtained while using a localized smoke identification with Faster R-CNN with vegetation backdrop is 80.3%.

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/>
7. <https://medium.com/rants-on-machine-learning/the-unreasonable-effectiveness-of-random-forests-f33c3ce28883>