Wildfire Severity Prediction using Spatio-Temporal Dataset Combining Wildfire Occurrence with Relevant Covariates and Machine Learning Models

Vasundhara Acharya[1], Professor Thilanka Munasinghe[2]

[1,2]Rensselaer Polytechnic Institute, Troy, NY, United States

Abstract

Fire risk management critically depends on the ability to model and predict fire spread. However, the development of predictive models is hampered by the scarcity of integrated datasets that correlate fire incidents with relevant environmental factors. In our work, we introduce an expansive dataset that combines historical fire incidents with corresponding covariates derived from satellite data. This research employs a suite of atmospheric and land cover variables, including temperature, precipitation, wind speed, and vegetation characteristics, to forecast the intensity of fire events as measured by Fire Radiative Power (FRP). We apply a range of machine learning algorithms—XGBoost, Random Forest, Lasso, Ridge, and Linear Regression, along with a Stacked Ensemble method—to address this predictive challenge. To mitigate the risk of overfitting, we meticulously partition our data into separate sets for training, validation, and testing. Model optimization for hyperparameters is systematically conducted through Optuna, and robustness is ensured by testing across various random seeds. The XGBoost regression model emerged as the most accurate, achieving a minimum Mean Absolute Error (MAE) of 1.1511 when predicting FRP levels on the test set. **Contrary to our preliminary assumption, excluding topographic data from the model inputs led to a slight decrease in performance, with the best MAE marginally increasing to 1.1550, thus underscoring the significance of topographical features in modeling fire spread**.

Introduction

A staggering $15 billion was spent on the disastrous Camp Fire in California in 2018, which left 88 people dead, displaced countless more, and destroyed over 18,500 properties. Wildfire management represents the challenging world of dynamic resource allocation in catastrophe response. The complexity of responding to natural disasters like wildfires, floods, and earthquakes is influenced by several factors. These events are dynamic in nature, defying simple modelling due to their complex and unpredictable system dynamics. Numerous factors, such as vegetation, fuel, altitude, and wind, have an impact on how a fire spreads. Given a dataset comprising relevant covariates, machine learning algorithms can be leveraged to forecast the propagation of wildfires [3]. In this work, we utilized a combination of machine learning models to harness their individual advantages: a Random Forest regressor for its robustness and handling of nonlinear data, a Linear Regression for simplicity and speed, a Ridge Regressor to address multicollinearity, an XGBOOST regressor for its gradient boosting performance, and finally, a stacked ensemble to integrate the predictions of all models. The effect of wildfire in terms of acres burned is shown in figure 1. In the future, we want to use mean field neural networks, a kind of neural network design that has shown success in processing sophisticated, high-dimensional data.

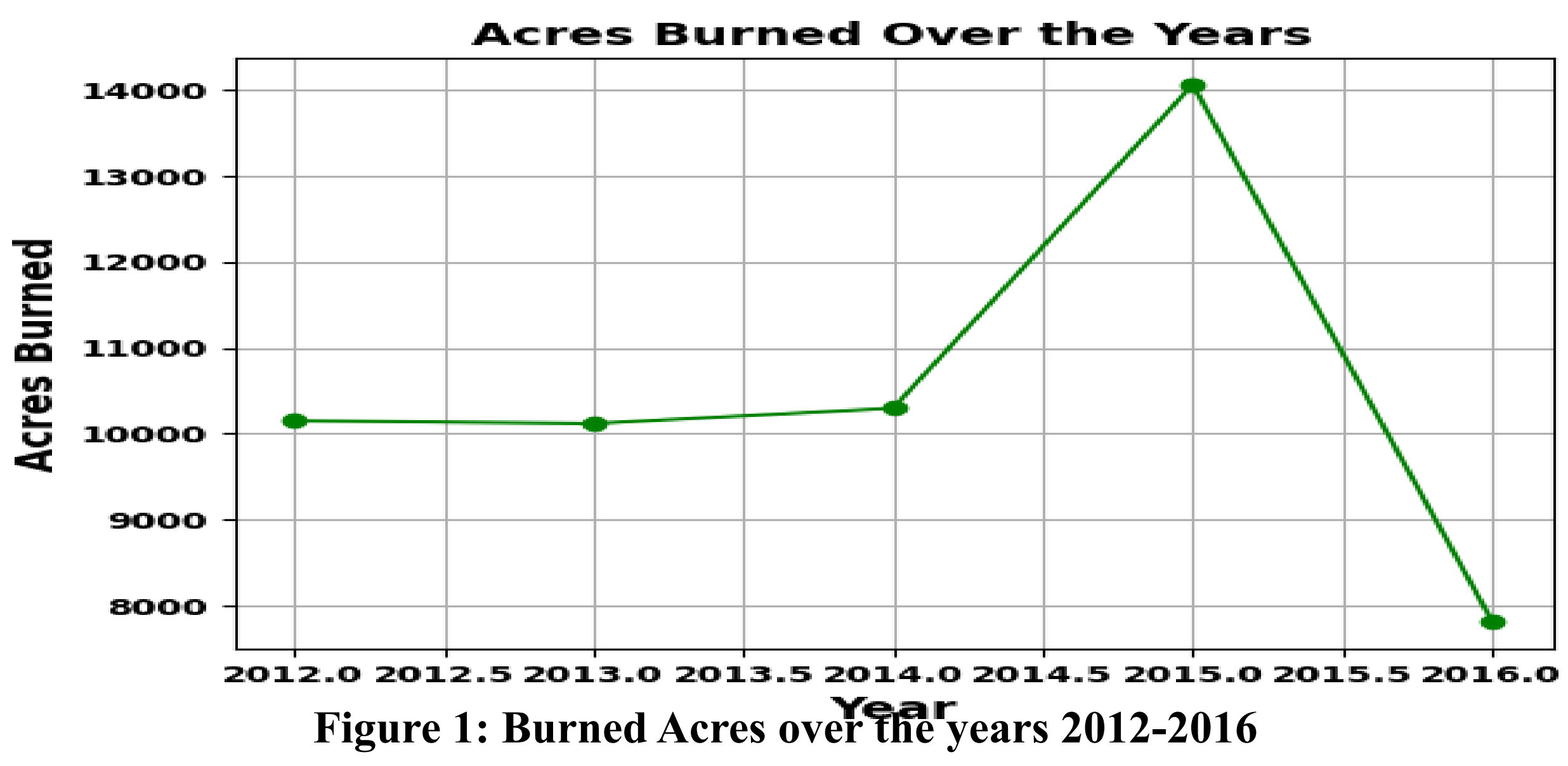

Figure 1: Burned Acres over the years 2012-2016

Data

For this proposed study, we have chosen 6,00,000 datapoints with 148 features. Every datapoint is made up of a polygon cell that is burning in one time step (a day), and the FRP of the neighboring polygon in the subsequent time step (the day after). In FRP, a zero value denotes the absence of fire in its Neighbor. The process of data generation is depicted in Figure 2. The table 1 shows the raster categories. For the training we use the instances acquired from years 2012-2016. To evaluate our model, the test set comprises exclusively of instances acquired in the year 2017.

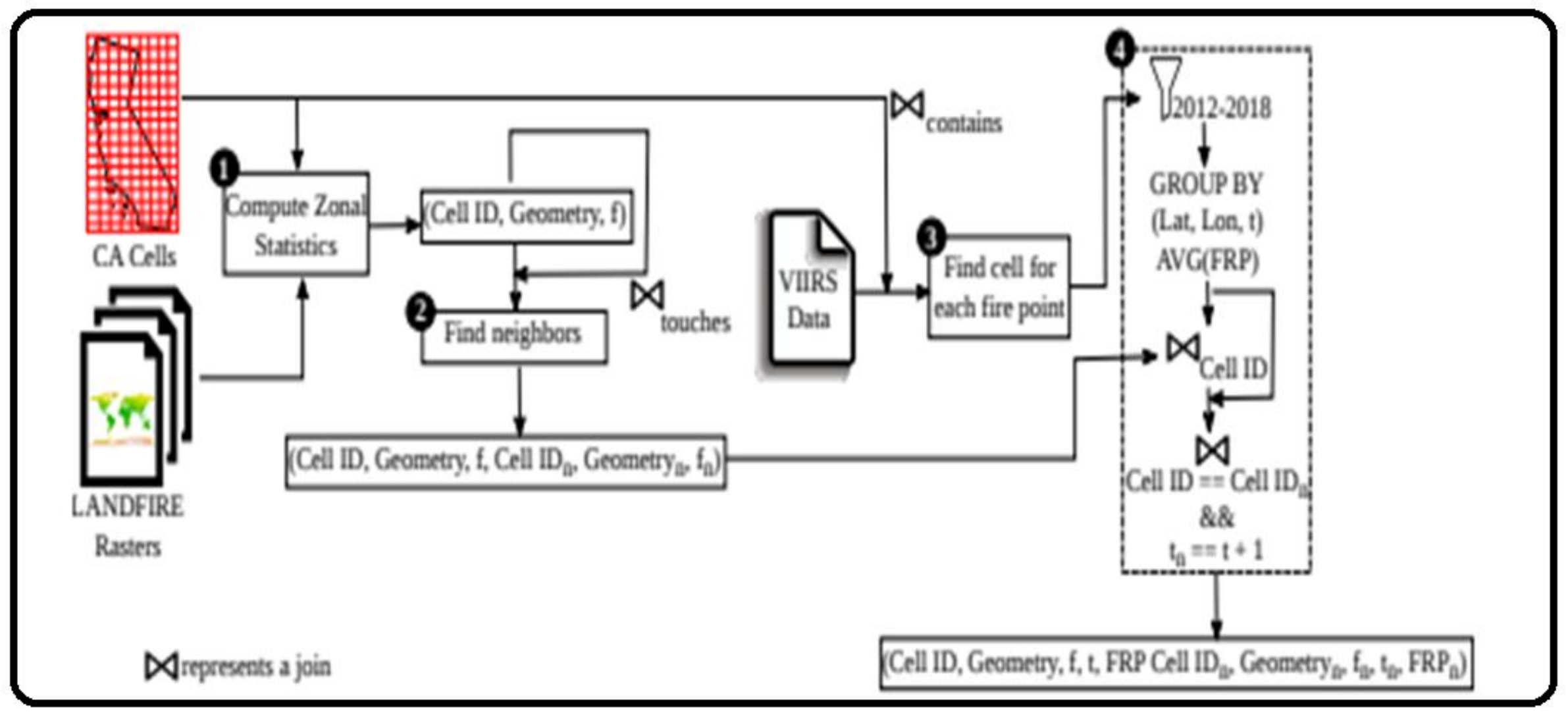
CA Cells
LANDFIRE Rasters
1
Compute Zonal Statistics
(Cell ID, Geometry, f)
2
Find neighbors
⋈ touches
(Cell ID, Geometry, f, Cell ID_n, Geometry_n, f_n)
VIIRS Data
⋈ contains
3
Find cell for each fire point
4
2012-2018
GROUP BY (Lat, Lon, t) AVG(FRP)
⋈ Cell ID
⋈ Cell ID == Cell ID_n && t_n == t + 1
(Cell ID, Geometry, f, t, FRP Cell ID_n, Geometry_n, f_n, t_n, FRP_n)
⋈ represents a join

Figure 2: Data Generation: Daily fire occurrences in years 2012 through 2016 [1]

Table 1: Raster Categories [1]

Name	Year(s)
Canopy Base Density	2012, 2014, 2016
Canopy Base Height	2012, 2014, 2016
Canopy Cover	2012, 2014, 2016
Canopy Height	2012, 2014, 2016
Existing Vegetation Cover	2012, 2014, 2016
Existing Vegetation Height	2012, 2014, 2016
Existing Vegetation Type	2012, 2014, 2016
Elevation	2016
Slope	2016

Methodology

EDA

The violin plot of variables is shown in figure 3. During the EDA, we did not eliminate the outliers. The distribution of the class label which is the fire severity is shown in figure 4. The boxplots of various variables is depicted in the figure 5.

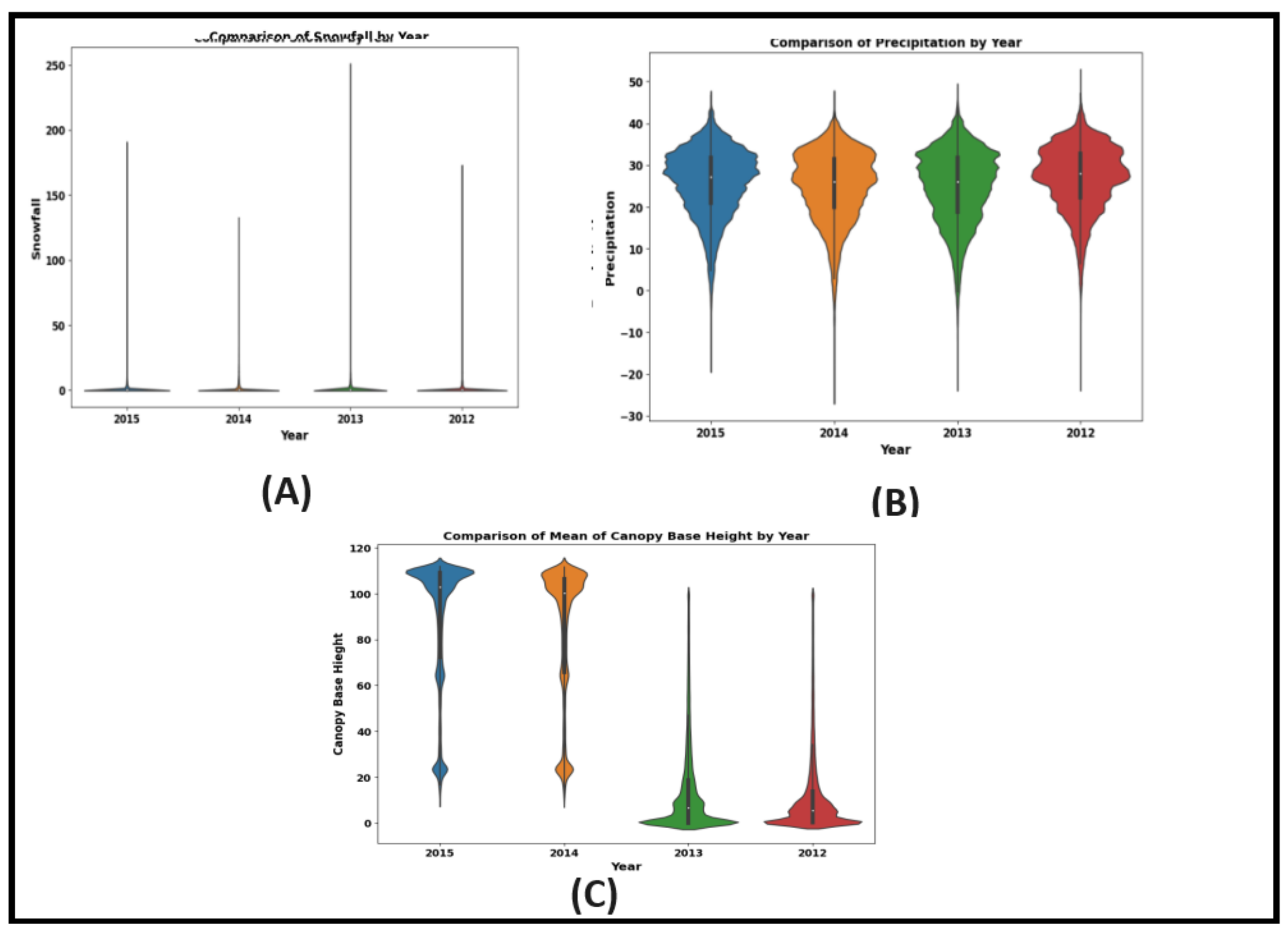

Figure 3: Comparison of the Violin Plot of Variables during the year 2012 and 2015. (A). Violin plot of Snow. (B). Violin Plot of Precipitation. (C). Violin Plot of Canopy Base Height

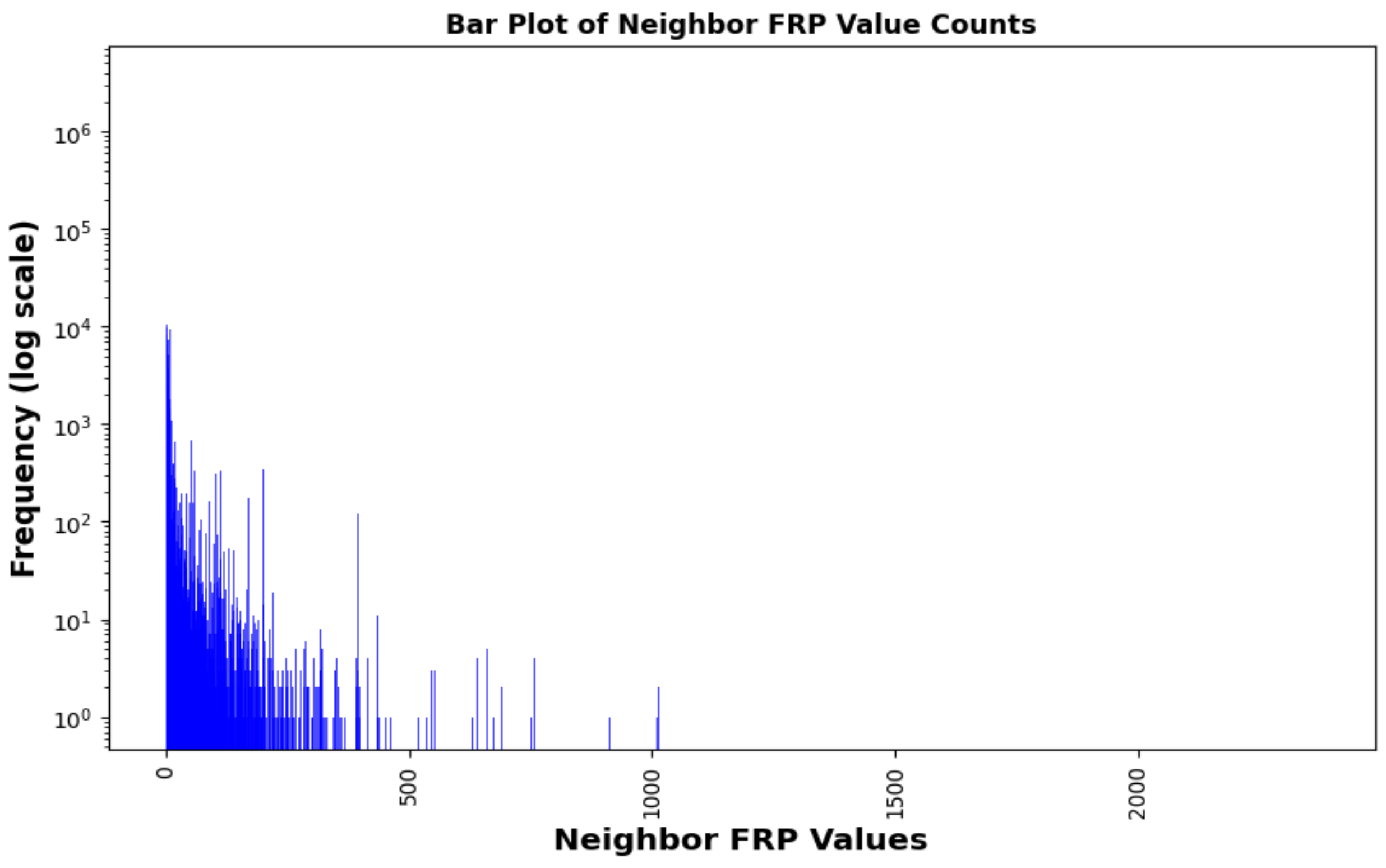

Figure 4: Fire Severity: Distribution of the Class Label

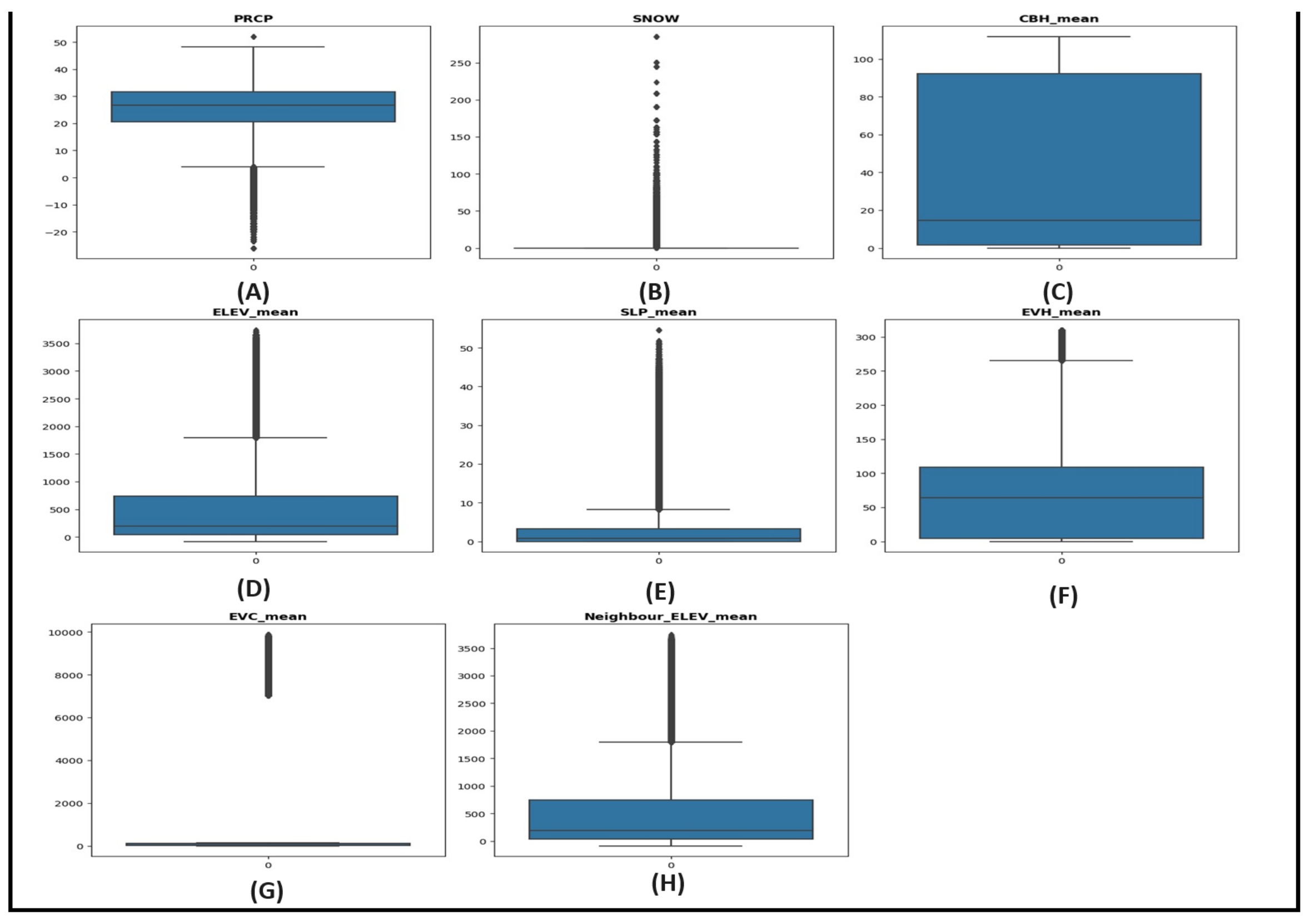

Figure 5:Box Plots: (A). Precipitation. (B). Snow. (C). CBH_mean. (D). ELEV_mean. (E) SLP_mean. (F). EVH_mean. (G). EVC_mean. (H). Neighbor_Elev_Mean

Models and T-Test

We employed a comprehensive approach by training various regression models, including the XGBoost, Random Forest, Lasso, Ridge, and Linear Regression to predict fire occurrence based on relevant features. Additionally, we constructed a stacked ensemble model that integrates the predictions from these individual models. To further enhance the ensemble's predictive power, we utilized a Lasso Regression model as the meta-model. This multi-step strategy leveraged the strengths of each component model to provide predictions of fire occurrence. We also performed t-tests to assess the differences in canopy cover and canopy height between the years 2012 and 2015. The proposed framework is shown in figure 6. In the rest of the poster, Model 1 denotes Random Forest algorithm. Model 2 represents the Linear regression. Ridge Regressor is denoted by Model 3. XGBoost represents Model 4 and Model 5 represents the stacked ensemble.

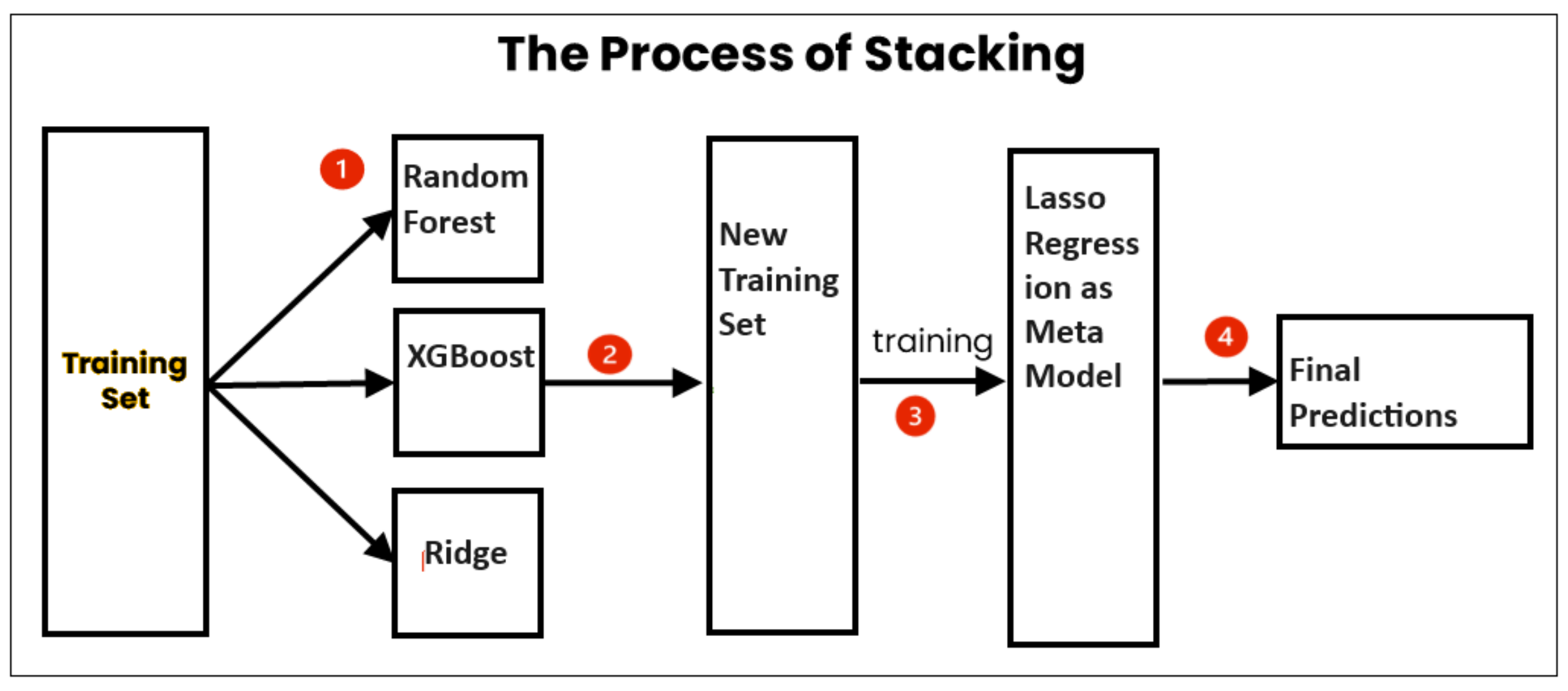

Figure 6: Ensemble Model. The numbers denote the order of the steps.

Results

The results of the t-test are shown in Table 2. The best performance achieved across different seeds are tabulated in table 3. The qqplot of different variables is plotted in figure 7. The figure 8 depicts the performance plots of various models.

Table 2:Results of T-Test

Null hypothesis	Alternate Hypothesis	P-Value
The population mean of Canopy Cover for year 2012 is similar to the population mean of Canopy Cover for year 2015	The population mean of Canopy Cover for year 2012 is different than the population mean of Canopy Cover for year 2015	**6.90E-306** (Reject Null Hypothesis)
The population mean of Canopy Base Height for year 2012 is similar to the population mean of Canopy Base Height for year 2015	The population mean of Canopy Base Height for year 2012 is different than the population mean of Canopy Base Height for year 2015	0 (Reject Null Hypothesis)

Table 3:Model Performance

Model	TRAIN MAE	VAL MAE	TEST MAE	Train MSE	VAL MSE	TEST MSE	TRAIN RMSE	VAL RMSE	TEST RMSE
1	0.83	0.86	1.23	39.93	44.18	79.31	6.32	6.65	8.91
2	1.48	1.48	1.99	48.77	44.89	80.33	6.98	6.70	8.96
3	0.98	0.98	1.24	46.49	45.33	77.27	6.82	6.73	8.79
4	**0.69**	**0.83**	**1.12**	**20.21**	**42.01**	**80.79**	**4.50**	**6.48**	**8.99**
5	0.72	0.97	1.57	12.47	43.05	94.42	3.53	6.56	9.72

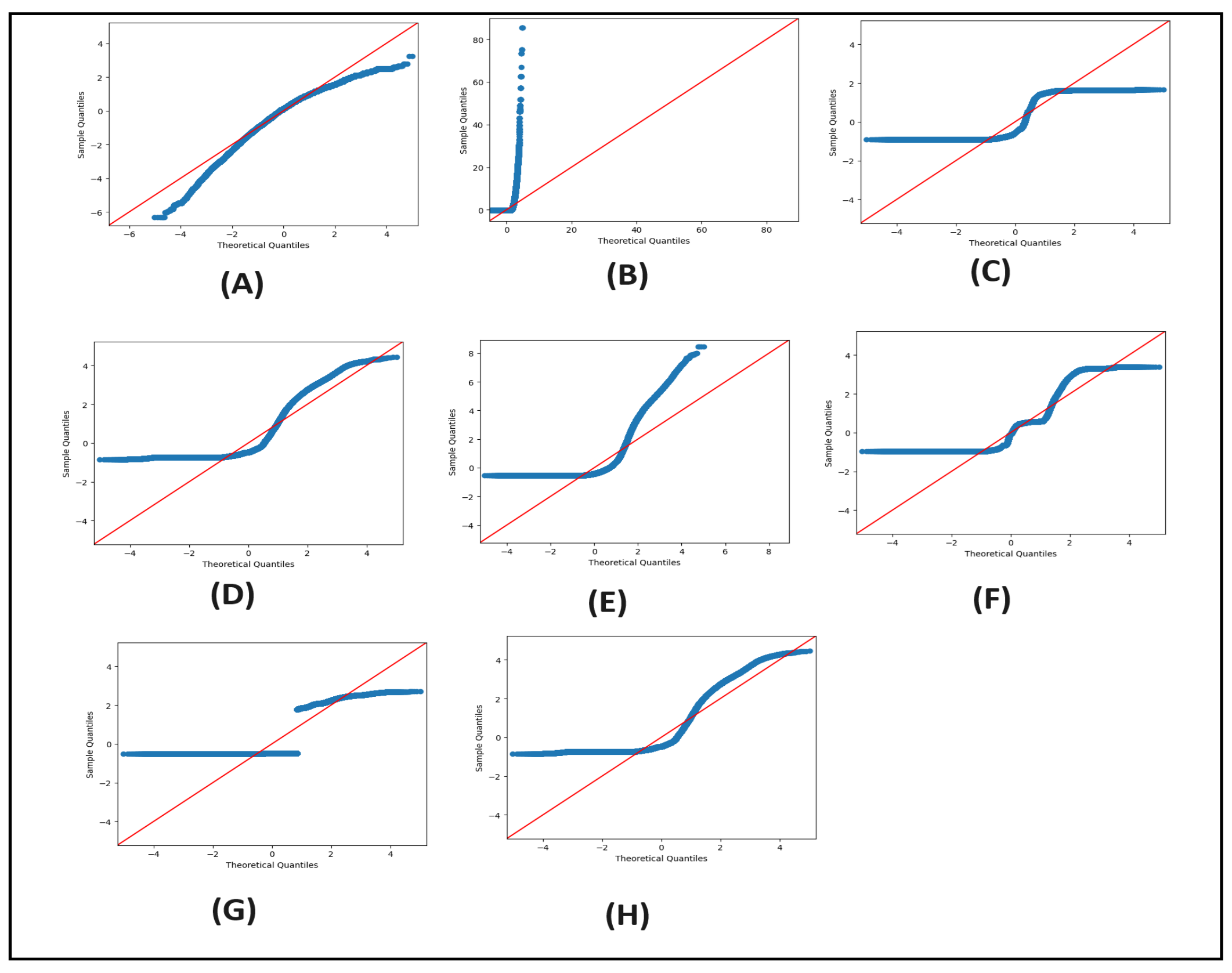

Figure 7: QQPlot (A).Precipitation. (B). Snow . (C). Mean of canopy base height . (D). Mean of elevation. (E). Mean of slope. (F). Mean of existing vegetation height. (G). Mean of existing vegetation cover. (H). Mean of neighbor elevation

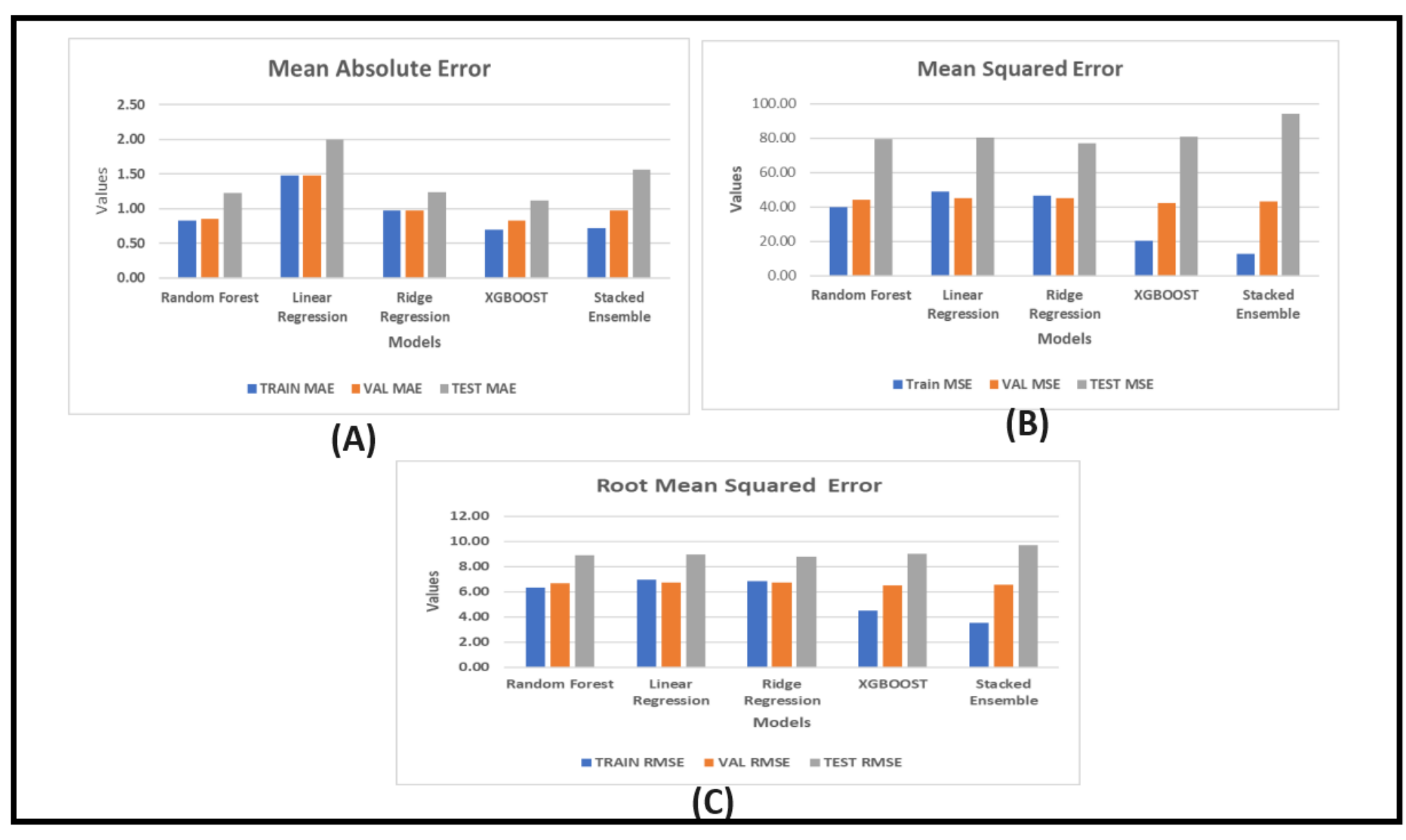

Figure 8: Plots of Model Performance. (A). Mean Absolute Error. (B). Mean Squared Error (C). Root Mean Squared Error

Conclusion

Our study's initial hypothesis was that, even in the lack of topography data, we could still make good forecasts about Fire Radiative Power (FRP) levels in wildfires by examining previous wildfire data in conjunction with information on vegetation and fuel types. When topographic data were removed from the model, the predictive capability of our XGBoost model—which performed the best out of all the machine learning algorithms we tested—was significantly reduced. This disproves our initial theory and emphasizes how important topography is in predicting wildfire behavior. XGBoost regressor model stood out due to its gradient boosting framework and had the capacity to manage complex relationships and non-linear correlations between variables. In the future, we would use a multifaceted feature selection strategy. Furthermore, it could be beneficial to investigate how deep learning methods might be used to capture the spatial correlations in the data.

REFERENCES:

[1]. Diao, T., Singla, S., Mukhopadhyay, A., Eldawy, A., Shachter, R., & Kochenderfer, M. (2020). Uncertainty aware wildfire management. *arXiv preprint arXiv:2010.07915*.

[2] . Singla, S., Diao, T., Mukhopadhyay, A., Eldawy, A., Shachter, R., & Kochenderfer, M. (2020). WildfireDB: A Spatio-Temporal Dataset Combining Wildfire Occurrence with Relevant Covariates. In 34th Conference on Neural Information Processing Systems (NeurIPS 2020).

[3] Artés, T., Oom, D., De Rigo, D., Durrant, T. H., Maianti, P., Libertà, G., & San-Miguel-Ayanz, J. (2019). A global wildfire dataset for the analysis of fire regimes and fire behaviour. Scientific data, 6(1), 296.