Predicting the occurrence of wildfires with the help remote sensing and weather data using machine learning techniques

Dharmendra Lalji Vishwakarma*, Prathamesh Pradeep Patil†,
Dinesh Narayan Gauda‡ and Yogesh Maruti Patil§
School of computing, National College Of Ireland
Advanced Data Mining
*x18108181@student.ncirl.ie, †x18114822@student.ncirl.ie, ‡x17169836@student.ncirl.ie, §x17169828@student.ncirl.ie

***Abstract*—Given the nature of destruction, predicting wildfires is an important research area for environmental prevention that is destroyed due to the large fire initiation. Previous researches have explored the different viewpoints and techniques to solve this problem. The factors that resemble wildfires are summarized and articulated with the help of a literature review. In addition, a new set of attributes are taken into consideration to observe the impact on the predicting models for wildfires classification. The performance of wildfires is observed to be improved with these additional weather attributes in addition to remote sensing data. The experiment was modelled on a set of machine learning algorithms- Random Forest, SVM, ANN and AutoML. The Random forest is found to be the best model with 99% accuracy, 98% sensitivity and specificity 100%. The outcome of this experiment shows an improvement over the state-of-the-art model in the wildfires prediction.**

***Index Terms*—wildfires prediction, machine learning, SVM, Random Forest, ANN, AutoML, NDVI**

I. Introduction

One of the worldwide disasters that has a big impact on earth's environment is wildfires. It is considered to be a catastrophic event in the current world which destroys the thousands of kilometres of natural habitats. This leads to change in the form of living conditions such as climate, pollution and sometimes indirectly affect the economy of the place. A fire can be initiated from various reasons like human factors, natural disasters, wind, lightning etc. In contrary, managing the fire events has another challenging problem to consider.

Due to the ambiguous nature of the wildfires, and their profound impact on the surroundings, it is considered by the many researchers to find solutions to prevent the fire, identify the causing factors and predicting the event. Given the geographical regions of worldwide and the complex meteorological conditions at different places, predicting the event of fires is a challenging computational problem. Moreover, the spread of the fire events is another dimension which adds complexities to the wildfire problem. It requires analysis of several other environmental attributes in the study.

One of the recent study conducted by the authors [1] in which they utilized satellite images to build machine learning models using the remote sensing characteristics of the fire and non-fire region. These attributes include crops state, meteorological conditions like surface temperature and thermal anomalies. Additionally, one of the future directions to improve the fire prediction in the forest is to utilise the weather data of the region. Climatic conditions play an important role in the fire situation. For instance, high wind can help the fire to spread across different regions and make a difficult situation to stop it. Therefore, these factors are considered in this research and presented in the form of the research question.

A. Research Question

> How can the prediction of wildfires be improved with the help of weather attributes such as air temperature, humidity and wind in addition to remote-sensing features using machine learning models?

B. Research Objectives

The objectives of this research are as follows

- To find the independent factors that contribute towards the wildfire prediction.
- To observe the importance of weather information in predicting the wildfire event.
- To derive the mathematical model to classify the event of a wildfire.

The next following sections of this research report are categorised as follows. Section II describes conclusively the previous studies in the field of wildfire prediction. The detailed methodology adapted for carrying out this research is outlined in section III. Section IV illustrates the evaluation techniques and results. Finally, section V provides the conclusion of the research and suggests possible future directions for further exploration in this domain.

II. Related Work

There are many previous researchers who studied the factors and effect of different ML methods in predicting wildfires. The following subsections describe these parameters and techniques used in the literature and summarized in the table I.

TABLE I
SUMMARIZATION OF PREDICTORS AND MODELS USED IN LITERATURE REVIEW

Reference	Predictors Used	Model Used
[2]	Relative humidity, temperature, wind speed, and rain	Random Forests and Support Vector Machines (SVM)
[3]	Wind speed, temperature, rainfall and humidity	Detrended fluctuation analysis [DFA], Detrended cross-correlation analysis [DCCA]
[4]	Wind speed was used along with precipitation and temperature	ANN
[5]	Palmer Drought Severity Index (PDSI), Standard Precipitation Index (SPI)	Generalized additive model
[6]	1st set of predictors: Geographical data of the land with forest 2nd set of predictors: Average temperature 3rd set of predictors: Wind speed direction, evaporation, humidity, transpiration	Random forests, bagging and boosting of decision trees
[7]	Vegetation Cover, human activity, climate, and topography	Random Forest
[8]	Thermal anomalies, NDVI and LST	SVM and neural networks
[9]	Normalized Difference Vegetation Index (NDVI), Normalized Burn Ratio (NBR) and Gross Primary Production (GPP)	Image processing
[10]	Proximity to human settlements, annual rainfall, altitude, proximity to roads	SVM and RF

Remote sensing has accelerated the amount of climatic data produced which has eventually helped in wildfire predictions. Many studies have been conducted in order to understand the importance of various meteorological measures in predicting wildfires. In a study conducted by [2], four meteorological parameters were used namely relative humidity, temperature, wind speed, and rain in order to predict the wildfires in the north-eastern province of Portugal. Here five different data mining methodologies were used e.g. Random Forests and Support Vector Machines (SVM) and the other three being different feature selection models. Among the models proposed, SVM proved to be the best and the results suggested that the meteorological conditions do affect the model with the most important being the temperature followed by accumulated precipitation. In another study conducted by [10], the same models were used. It stated that topography like elevation, altitude have an indirect effect on the wildfire and climatic factors such as wind, temperature and rain have a direct impact on the wildfire. To predict the occurrence of wildfire SVM and RF were used as they are suitable for classification problems. RF Gini index was used to see the correlation between climatic factors and wildfire. RF Gini index for the wind was 4.8 and for rainfall, it was high 8.85. Gini index gives an estimation of the influence of predictors on the dependent variable. The overall accuracy for RF was 70 per cent and SVM was 83 per cent.

Importance of temperature in wildfire prediction was also mentioned in another study conducted by [3] in Hunan District of southern China. This was carried out to predict wildfires using wind speed, temperature, rainfall and humidity. And it was found that temperature and humidity influence the occurrence of wildfires. Humidity is a measure of water in the air and if that is less with higher temperature then there are more chances of wildfire. The detrended fluctuation analysis (DFA) was used to find correlation analysis between meteorological parameters and wildfire. DFA was used due to the nonlinear nature of time series data. In another study conducted by [4] in South Africa's Cape Town, the wind speed was used along with precipitation and temperature to predict wildfire spread. Emphasis was given on predicting Burning Index (BI) which signifies the severity of the fire. BI was calculated using wind and rain as additional factors to predict wildfire with temperature as the primary predictor for wildfire. Wind and rainfall were used as correction factors to predict BI and the selection process was based on conditions like wind speed greater than 20 km/h and precipitation below 50 per cent. In conclusion, the wind speed and vegetation parameters were found to not affect the wildfire in Cape Town.

In a study conducted by [5], it was found that a low precipitation results into lower soil water content which results in dry vegetation and a greater chance of wildfire. It was also mentioned that Droughts are caused due to prolonged low precipitation in the soil. A long drought period indicates low water content in the soil and dry vegetation. Indicators like the Palmer Drought Severity Index (PDSI), Standard Precipitation Index (SPI) are two major measures of drought. PDSI is calculated using tree rings data and SPI is calculated using precipitation mainly. Here, [5] also stated that during drought periods the burnt area of wildfire was found to be greater when compared to non-drought areas for the same period.

In the research work conducted by [6], prediction of wildfire was performed by utilizing three different data in the region of Slovenia. The first dataset consisted of data from geographical information system (GIS) which basically included the geographical data of the land with forest, urban places, fields, distance from highways, etc. The second dataset comprised of the Multi-temporal MODIS data which consisted of the average temperature of a specific area. The third dataset comprised of the meteorological data namely wind speed direction, evaporation, humidity, transpiration, etc. In terms of accuracy for all the three datasets, bagging of the decision trees gave the best results. On the other hand, in a study conducted by [7], the objective was to identify the relative impact of vegetation, human activity, climate, and topography in order to determine the occurrences of wildfires. Random Forest model was applied to check the effect of the above-mentioned parameters. The results implied that in the north-central regions with less population and the possibility of high elevation, the cause of the fire was lightning. In the regions with high dense population, the cause was human-caused fire. Climatic factors played an important part in predicting the wildfires followed by human activity. The author further suggested incorporating human influences in predicting the wildfires.

The remote sensing data also describes the combined effect of the vegetation that contributes to the factors which are having an influence on the status which is considered as a fuel base for the fires. There are a lot of parameters present that help in predicting the wildfires which include NDVI. It is defined as the distinction among visible and the near-infrared effectiveness in reflecting radiant energy of vegetation cover [11]. Also, this is helpful in estimating the volumetric mass density of the green on a specific area of the earth. Some researchers have used NDVI in order to predict wildfire. In [8], two supervised data mining algorithms namely SVM and neural networks were used for wildfire prediction using three measures namely thermal anomalies, NDVI and LST. The data was obtained from Terra's instrument MODIS, and the data was further processed accordingly. NDVI is determined as the index that shows the state of crop health. LST is determined by the temperature of the radiations of the land cover derived from solar radiation. The results of both the model proved to be good with SVM having 97.48% accuracy and NN having 98.32% accuracy. In future work, the author proposed the use of weather data namely wind, air temperature, and soil moisture in order to check the overall fit of the model. Whereas in another study conducted by [9] in order to review the burnt area over a period of time, three measures viz Normalized Difference Vegetation Index (NDVI), Normalized Burn Ratio (NBR) and Gross Primary Production (GPP) were used for the forest in Gangwon, South Korea. These measures were taken from LANDSAT and MODIS. The measures were examined by making a comparison among burnt and unburnt reference area for a time span of 18 years. Based on the observations made by the use of NDVI it was stated that restoration of the forest would need 9 years but in research, it took more than that so it was further stated that recovery of the model could be improved by adding some more attributes.

III. RESEARCH METHODOLOGY

To implement this research, there was a need to follow a data mining methodology so that each phase of the research could be performed smoothly and successfully. Therefore, CRISP-DM was selected based on its proven success in the past for data mining projects [12]. The business understanding part is explored through the introduction, research question and literature review of this paper. The next step in the research is to understand the data.

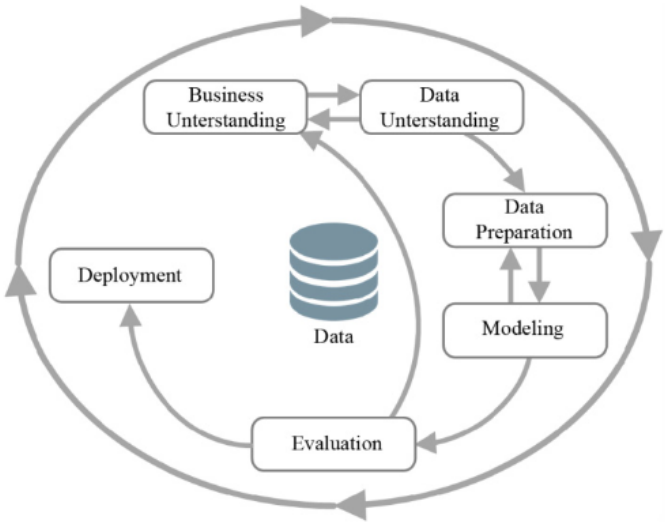

Fig. 1. CRISP-DM Technique

A. Data Understanding

Factors affecting the occurrence of wildfire situations are considered in this study. These factors are NDVI, weather information like humidity, temperature, precipitation, wind and solar. The significance of these parameters is understood with the help of a literature review for the cause and effect of fires.

B. Data Preparation

Due to the large availability of dataset, the Canadian region is chosen for the study. Therefore, the required dataset is gathered from the following public data sources provided by scientific organizations for research perspectives.

1) Wildfires events: All fires data is available from the government of Canada website from natural resources division [13]. It contains all types of fires with geographic coordinates, duration of the fire, date information and other recorded metrics. There is a different variant of fire data available. For this study, large fires with greater than 200 hectares of impacted areas are taken into consideration.

2) Weather information: Meteorological data is captured from the global weather organization [14]. It contains precipitation, relative humidity, wind, temperature and solar metrics at day level with geographical coordinates.

3) NDVI indices: NDVI data which represents the crops conditions, is downloaded from National centres for Environment Information(NCEI) website [15]. It contains worldwide vegetation indices with date, latitude and longitude attributes.

The data preparation step was quite challenging and time-consuming due to the spatial characteristics of raw data. These data are available in different formats and needs special filtering procedures to be able to feed into ML models. Once data is collected, there was a need to transform and integrate these different sets of data. For this purpose, the data pre-processing strategy was applied.

C. Data Preprocessing

The different sources of data were available in a different format.

1) Wildfires: The wildfire data is available in comma separated format with the text file extension. This file is loaded into R programming language and required attributes such as date, wildfire, latitude and longitude are transformed, filtered with R data frame. Firstly, wildfire data for the year 2010 was extracted, but as the occurrence of wildfire was very low in that particular year as compared to the non occurrence of wildfire it resulted in a huge imbalance in dataset and to tackle this issue, only data of wildfire occurrence for the year span ranging from 2011-2014 was added which reduced the class imbalance to some extent. As there were certain coordinates which didn't match with the coordinates present in other data sources, all the values of latitude and longitude were rounded off and later only the unique combinations of columns (Date, Lat, Long) were taken into consideration. The new file with these filtered attributes was stored in a different CSV file.

2) Weather data: The weather data within the specific boundary region is extracted with the help of longitude and longitudes of wildfire data. For each date, this process is repeated and the new dataset is prepared according to date, lat, long. This data didn't require much of cleaning, only the column headers were named properly.

3) NDVI: The NDVI data were taken from [15] which consisted of a large number of .nc files each with a size of approximately 60 MB on a daily basis. These NC files were not feasible to be downloaded manually so the downloading was done programmatically using R. This data downloading process was quite time-consuming as it required two hours of time for each year. The desired NC files to be downloaded were decided on the basis of Date taken from fire_data.csv, and this date was appended into a string in order to form the name of the NC file. Once the files were downloaded further steps involved in NDVI data extraction and processing are mentioned below

- the ncdf4 package was used to open the .nc file.
- In the metadata there were two variables, one was time_bnds which provided the starting date and the end date of the observation along with NDVI consisting of latitude, longitude and time. This NC data couldn't be extracted directly so the file format for each day was converted into TIFF format.
- Now the data needs to be extracted from the TIFF file which required the desired latitude and longitude which were taken from the fire_data.csv for that particular day. Also, the coordinates needed by TIFF file needs to be converted into spatial points and then fed to the raster to extract the NDVI index.
- A new CSV was created with desired date, lat, long and NDVI index which was later merged with the Fire CSV and weather data.

This way, three files from each data sources were integrated using common date, latitude and longitude attributes. Latitude and longitude values were rounded in order to integrate with multiple regions of data. No fire data set was considered using other remaining data set on which wildfire dates were absent. After integrating, a total of 1846 rows of data generated out of which 401 records with fire class.

D. Modelling

Machine learning models were utilized to predict wildfires events. As mentioned in the literature study, previously, different machine learning algorithms have been applied for this classification problem. For this study, Random forest, SVM and ANN with PCA are used along with Auto Model.

1) Random Forest: In this study, wildfire prediction is a classification problem. Thus, Random Forest was chosen as an ensemble of multiple trees to predict the binary outcome of wildfire in a particular region of Canada. Random Forest is robust to outliers and most suitable for non-linear data hence it was chosen for classification of wildfire [16].

2) SVM: Even though Random forest (RF) is well suited for non-linear and outlier data, RF performs better when there are multiple classes to predict and in this paper the wildfire prediction is binary class. Hence, due to this reason, SVM which performs better for binary classification was used for comparison purpose with RF for wildfire binary classification.

3) ANN with PCA: ANN takes into account the actual physical importance of data in terms of classification which SVM and RF do not. The hidden aspect or weight bias is only captured by ANN. Also, ANN is suitable for non-linear data and binary multi-perceptron classification hence ANN was chosen for the study with PCA for knowing the influential factors.

4) AutoML: AutoML was used for comparison with the results obtained from RF, SVM and ANN. It was just a frame of reference to get a rough idea of what results to be expected from other algorithms.

The next section describes the experimental results obtained from the study.

IV. EVALUATION/RESULTS

To answer the objectives listed in the section I-B, four different methodologies namely RF, SVM, ANN with PCA, and AutoML are used to predict wildfires. The performance of proposed models has been evaluated by comparing values of Accuracy, Kappa, Sensitivity, and Specificity for each methodology.

A. Exploratory Data Analysis

Before applying any machine learning algorithm, Exploratory Data Analysis (EDA) was applied to the data to get a better understanding of data and prepare data for further analysis. EDA started by checking the class balance of the output variable, the result of class balance test showed that there is a class imbalance in the dataset with more fire data in the dataset than no fire data as shown in figure 2, but not as much as we had before data preprocessing. After that, check for missing values was carried out which confirmed there arent any missing or NA values in the dataset.

```
> table(wildfire_data$fire)

  No  Yes
1445  401
```

Fig. 2. Class imbalance in outcome variable

1) Univariate Analysis: Histogram and boxplot were plotted to understand the distribution of NDVI data. The results showed that the mean value is around 0.18 and the distribution is left-skewed meaning most of the NDVI values are closer to 1 than 1 as shown in figure 3 and 4.

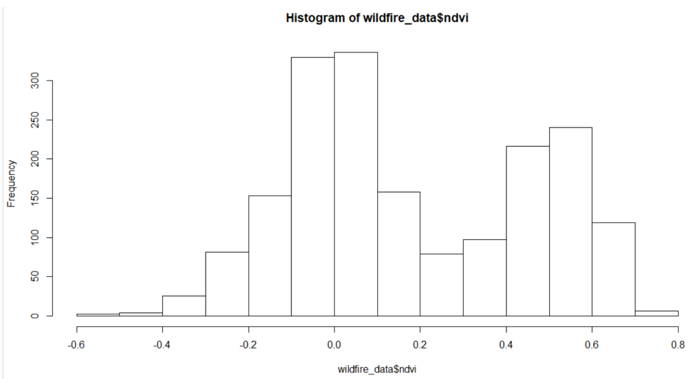

Fig. 3. Histogram of NDVI

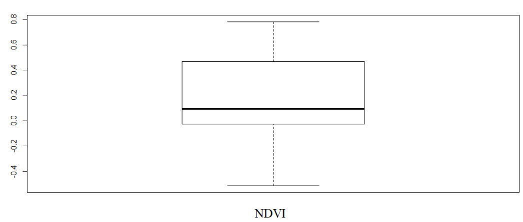

Fig. 4. Boxplot of NDVI

Next boxplot was plotted to find outliers if any, the results as shown in figure 5 showed that precipitation and wind have many outliers. As the data is not scaled, boxplot of few parameters were not clearly visible and so separate box plots were plotted as shown in figure 6.

2) Bivariate Analysis: To check the relationship between NDVI and wildfire, boxplot and snipe plot were plotted as shown in the figures 7 and 8. The results showed that with an increase in NDVI value, chances of occurrences of wildfire increases.

To verify the relationship between max temperature and NDVI, scatter plot was plotted as shown in figure 9. The results showed that the relationship is linear and for higher temperature values, NDVI values are higher.

To check multicollinearity between the predictors, correlation matrix was plotted as shown in figure 10. The results

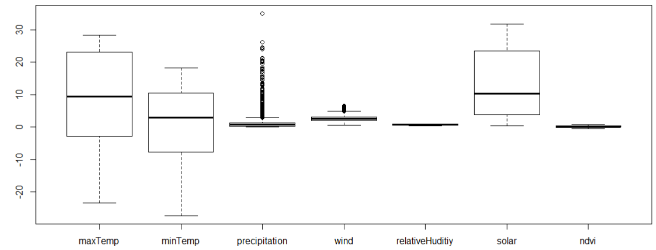

Fig. 5. Outliers in predictors

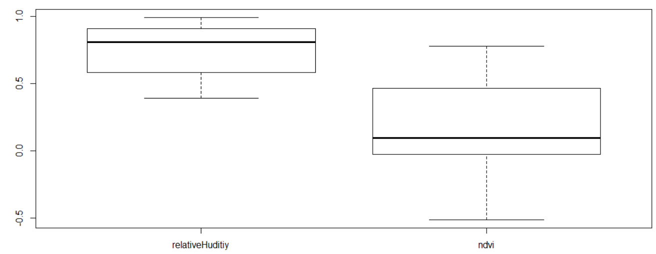

Fig. 6. Outliers in humidity and NDVI

showed that there is a high correlation between the maximum and minimum temperature and to avoid multicollinearity, average temperature is considered for further analysis instead of maximum and minimum temperature.

B. Random Forest

Basic Model: First, a basic model of Random Forest with random ntree was used to predict wildfire. The data used in this model was first scaled and then data was sliced in 75:25 ratio to train and test model. After training the model with training data, MeanDecreaseGini values were plotted using a varImpPlot method which showed the correlation of predictors with the outcome variable as shown in figure 11. As the results showed that all other predictors except wind have a high correlation between with outcome variable, all predictors were considered for future analysis except wind.

The performance of the basic random forest method is used as a bench marking model to compare other tuned random forest models. The performance of the basic random forest model evaluated using confusion matrix was significantly good with the following results (Accuracy: 0.9957, Kappa: 0.9871, Sensitivity: 0.98, Specificity: 1.0000).

1) Tuning of the parameters of RF: To further try and increase the performance of the model, two more new models were created.

Finding the optimal value of a number of trees(tree): In the first model, number of trees in the model were tuned by randomly trying a different number of trees and comparing the performance of each model. The results suggested that

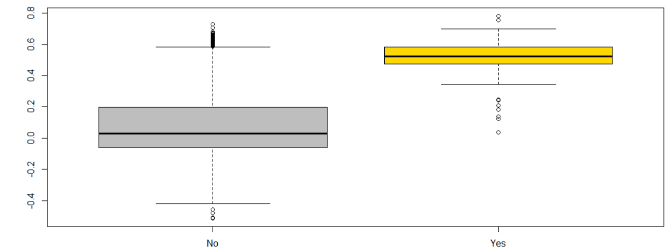

Fig. 7. Boxplot of NDVI vs Wildfire occurrence

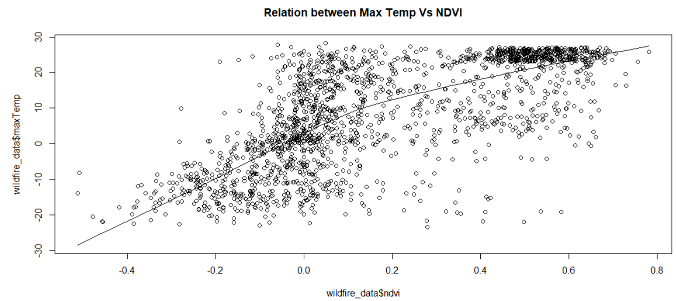

Fig. 9. Scatter plot of maximum temperature vs NDVI

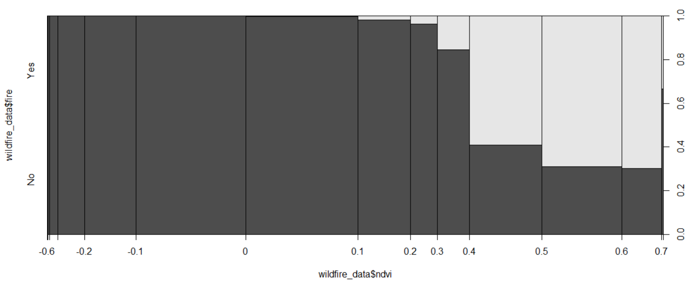

Fig. 8. Spineplot of NDVI vs Wildfire occurrence

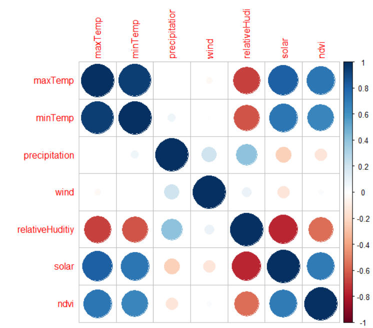

Fig. 10. Correlation matrix

after 2000 number of trees, there isnt much increase in the performance of the model with an increase in a number of trees values as shown in figure 12 and so, 2000 number of trees were selected for future analysis.

Finding the optimal value of mtry:
To find the optimal value of mtry, train method of caret package was used, wherein initial mtry value using the square root of the number of predictors was used and then with the help of caret method, best value of mtry was calculated. The results suggested that with mtry value of 4 the accuracy is maximum while with an increase in mtry the accuracy also increases but after 4 it decreases as shown in figure 13 hence, mtry value of 4 was used for the future analysis.

Final optimized RF model:
Now with the tuned values of random forest parameters, a new model of the random forest was used to predict wildfire. 75% of data was used to train the model and rest 25% to test the model. The results of the new model evaluated using confusion matrix (Accuracy: 0.9935, Kappa: 0.98, Sensitivity: 0.98, Specificity: 0.9972) showed that there isnt much improvement in the performance as compared to the performance of the basic model, rather in few aspects, the performance degrades.

C. SVM

Two models of SVM was created to try and find the best model for classification of wildfires. First, an SVM model using kernel as linear was created and then using a radial kernel. Here also training and testing sets of data created similar to previous model.

1) SVM model using a linear kernel: A basic model of SVM was first created using a linear kernel, cost as 10, and gamma value of around 0.167. Now to tune model so as to improve the accuracy, multiple values of cost were tried and tune method was used to get the best model. After that prediction was done using the best model created and testing data. The results created using confusion matrix (Accuracy: 0.9761, Kappa: 0.931, Sensitivity: 0.9700, Specificity: 0.9778) suggested that the performance of the SVM model with the linear kernel is not better than Random Forest. Now to try and improve the performance SVM model using radial kernel was created.

2) SVM model using a radial kernel: A similar approach to SVM with linear kernel was followed here as well, starting with a basic model and then tuning the parameters of the model to get the best model. The best model created was then used for prediction and its performance was evaluated using the confusion matrix (Accuracy: 0.9848, Kappa: 0.9558, Sensitivity: 0.98, Specificity: 0.9861). The results suggested that the performance of the SVM model with radial kernel was better than SVM using linear kernel but, the performance of RF was still better than both of these SVM models.

D. PCA

To perform PCA on the dataset, 80 per cent of data is used and two components were created wherein component one accounts for around 53.5% of the variance in outcome variable while component two accounts for around 21% of variance as shown in figure 14. To cover the maximum amount of variance,

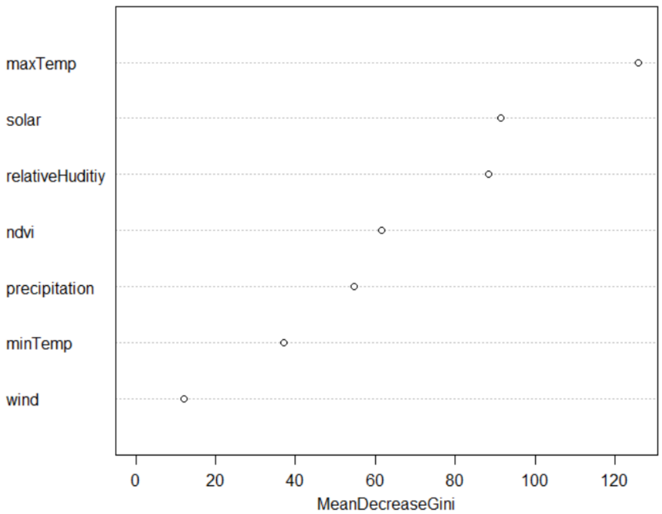

Fig. 11. Gini values of predictors

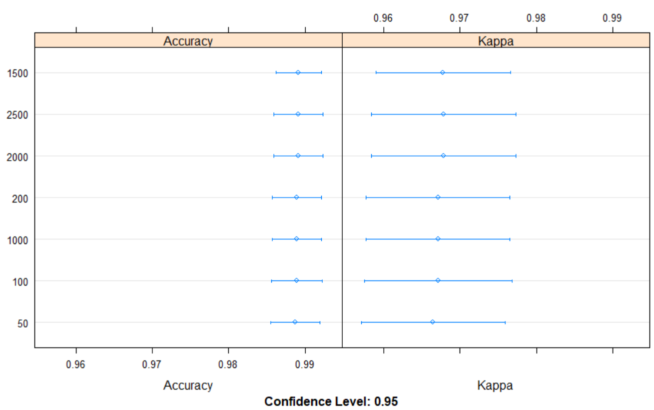

Fig. 12. Number of tress vs Accuracy and Kappa

different number of components were tried and a plot was plotted as shown in figure 15 which showed four components account for maximum variance in outcome variable and hence PCA was done using 4 components. Training set and testing set of data were created using PCA data by splitting the data in 75:25 ratio which is future applied to ANN for prediction of wildfires.

E. ANN

A basic model of ANN applied to the PCA data using the h2o package of R which uses the distributed environment in the background for the analysis. Activation function Tanh was used along with 160 neurons and three hidden layers. The epochs were selected as 20. The performance of model evaluated using confusion matrix (Accuracy: 0.9826, Kappa: 0.9489, Sensitivity: 0.9889, Specificity: 0.96) confirmed that performance of ANN is better than SVM with linear kernel but worst than RF and SVM with the radial kernel.

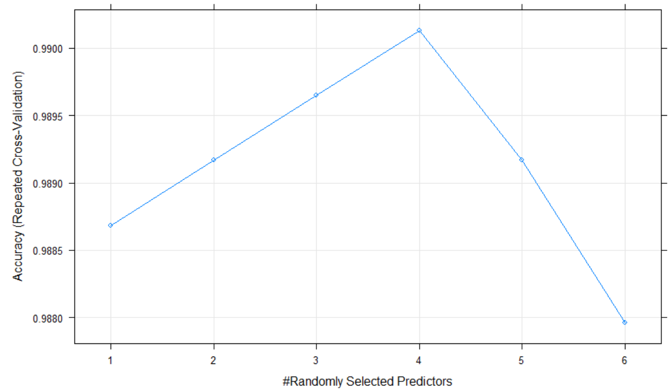

Fig. 13. Mtry vs Accuracy

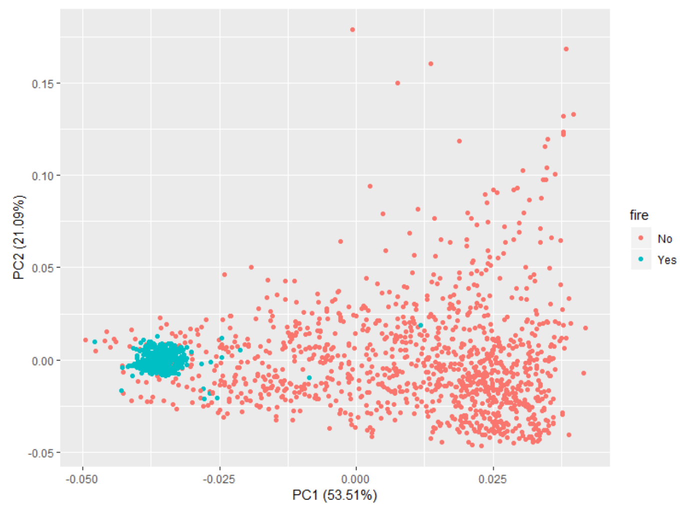

Fig. 14. PCA

1) Optimization of parameters of ANN: To improve the performance, tuning of parameters of ANN was done. A validation set of data was also created using one-third of the data. Different combinations of a number of neurons and hidden layers in the model were used along with two different values of L1 & L2 regularization. Three different activation functions were also used. After that, the best model with the highest accuracy is chosen for predicting wildfires. The performance evaluated using confusion matrix (Accuracy: 0.9805, Kappa: 0.9415, Sensitivity: 0.9945, Specificity: 0.963) confirmed that performance of ANN with optimized parameters has no improvement over base ANN model.

F. AutoML

AutoML is applied to PCA data using the h2o package of R. In AutoML, different machine learning algorithms are applied and the prediction is done using the best model. The performance evaluated using confusion matrix (Accuracy: 0.9706, Kappa: 0.9145, Sensitivity: 0.9832, Specificity: 0.9265) meant that performance of AutoML is even worst

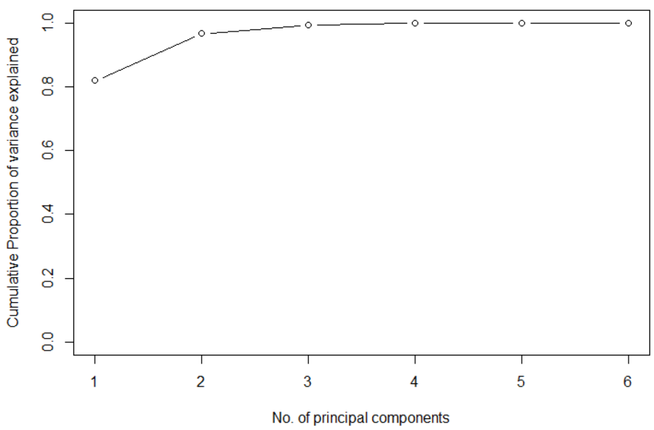

Fig. 15. Number of components vs Cumulative proportion of variance explained

TABLE II
COMPARISON OF MODELS

Methodology	Accuracy	Sensitivity	Specificity
RF-basic	0.9957	0.98	1
RF-optimized	0.9935	0.98	0.9972
SVM-linear	0.9761	0.9700	0.9778
SVM-radial	0.9848	0.98	0.9861
ANN	0.9805	0.9945	0.963
AutoML	0.9706	0.9832	0.9265
Base Model	0.9832	0.98	0.9797

than SVM with linear model concluding that performance of AutoML is lowest among all compared models whereas the performance of random forest is the best model as shown in table II.

V. CONCLUSION & FUTURE WORK

In this research study, Wildfire problems were introduced and factors that can influence the behaviours of fires is also outlined. With the help of this study, it can be concluded that meteorological data can also be useful predictors along with remote sensing data for wildfire prediction. Moreover, this research considers the methodology in which data gathering, processing, modelling and evaluations are performed using the CRISP-DM process. The required data is extracted, transformed, cleaned and scaled before applying ML algorithms. These models are evaluated and fine-tuned using several optimization techniques to build the best models. Furthermore, trained models are evaluated, reiterated and best model is prepared. Finally, The random forest was found to be best performing with an accuracy of 99%. The results showed significant improvement over the base model proposed by previous authors in the field of wildfire prediction.

Some of the future drifts in wildfire could be predicting the spread of fire. This would help in evacuation plans & prevention policies for firefighter team. Another future work could consider human causes of fire as fire could be triggered due to human activities. This can be achieved through surveying historic fire regions which involve human participants. Additionally, the proximity of dry vegetation area to cities is another research scope which would help in identifying fires causes. A solution to these problems will help firefighters to control the event in the near future more prominently.

REFERENCES

[1] Y. O. Sayad, H. Mousannif, and H. A. Moatassime, "Predictive modeling of wildfires: A new dataset and machine learning approach," *Fire Safety Journal*, vol. 104, pp. 130 – 146, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0379711218303941

[2] P. Cortez and A. d. J. R. Morais, "A data mining approach to predict forest fires using meteorological data," vol. 2014, no. November, 2007.

[3] J. Lu, T. Zhou, B. Li, and C. Wu, "Scale Analysis and Correlation Study of Wildfire and the Meteorological Factors That Influence It," *Mathematical Problems in Engineering*, vol. 2018, pp. 1–10, 2018.

[4] S. Lall and B. Mathibela, "The application of artificial neural networks for wildfire risk prediction," *International Conference on Robotics and Automation for Humanitarian Applications, RAHA 2016 - Conference Proceedings*, 2017.

[5] J. L. Crockett and A. Leroy Westerling, "Greater temperature and precipitation extremes intensify Western U.S. droughts, wildfire severity, and sierra Nevada tree mortality," *Journal of Climate*, vol. 31, no. 1, pp. 341–354, 2018.

[6] D. Stojanova, A. Kobler, S. Džeroski, and K. Taškova, "Learning To Predict Forest Fires," *In 9th International multiconference Information Society*, no. May 2014, pp. 3–6, 2006.

[7] Z. Wu, H. S. He, J. Yang, Z. Liu, and Y. Liang, "Relative effects of climatic and local factors on fire occurrence in boreal forest landscapes of northeastern China," *Science of the Total Environment*, vol. 493, pp. 472–480, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.scitotenv.2014.06.011

[8] Y. O. Sayad, H. Mousannif, and H. Al Moatassime, "Predictive modeling of wildfires: A new dataset and machine learning approach," *Fire Safety Journal*, vol. 104, pp. 130–146, 2019. [Online]. Available: https://doi.org/10.1016/j.firesaf.2019.01.006

[9] J. H. Ryu, K. S. Han, S. Hong, N. W. Park, Y. W. Lee, and J. Cho, "Satellite-Based Evaluation of the Post-Fire Recovery Process from the Worst Forest Fire Case in South Korea," *Remote Sensing*, vol. 10, no. 6, 2018.

[10] A. Jaafari and H. R. Pourghasemi, *Factors Influencing Regional-Scale Wildfire Probability in Iran*. Elsevier Inc., 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/B9780128152263000284

[11] "Earthobservatory." [Online]. Available: https://earthobservatory.nasa.gov/features/MeasuringVegetation

[12] R. Wirth, "CRISP-DM : Towards a Standard Process Model for Data Mining," no. 24959, 2000, pp. 29—-39.

[13] "Canadian Wildland Fire Information System — Download data." [Online]. Available: http://cwfis.cfs.nrcan.gc.ca/datamart/download/nfdbpnt?token=b30e57bb32a7dff7334e2515a73c11d9

[14] "Global Weather Data for SWAT." [Online]. Available: https://globalweather.tamu.edu/

[15] "Index of /data/avhrr-land-normalized-difference-vegetation-index/access." [Online]. Available: https://www.ncei.noaa.gov/data/avhrr-land-normalized-difference-vegetation-index/access/

[16] T. Shi and S. Horvath, "Unsupervised learning with random forest predictors," *Journal of Computational and Graphical Statistics*, vol. 15, no. 1, pp. 118–138, 2006. [Online]. Available: https://doi.org/10.1198/106186006X94072