ENGG 680 – Introduction to Digital Engineering

Project Report

Wildfire Assessment and Predictive Modelling

Group 5 – Fall 2024

Vikesh Dharmeshkumar Patel
30255939

Boya Douho
30261119

Kazi Zarin Tasnim Rafa
30233931

Ashkan Einiaghdam
30270232

Ray Pan
30265201

Socretes Saha
30264159

Chunsheng Xiao
30066914

UNIVERSITY OF CALGARY

Title of Project	Wildfire Assessment and Predictive Modelling
Group Number	5

We, the undersigned, certify that this is our own work, which has been done expressly for this course, either without the assistance of any other party or where appropriate we have acknowledged the work of others. Further, we have read and understood the section in the university calendar on plagiarism/cheating/other academic misconduct and we are aware of the implications thereof. We request that the total mark for this assignment be distributed as follows among group members:

Your Name	Vikesh Dharmeshkumar Patel
Student ID	30255939
Contribution (%) and Hours	14.285%, 5
Signature and Date	Vikesh Dharmeshkumar Patel / Dec 8,2024

Your Name	Boya Douho
Student ID	30261119
Contribution (%) and Hours	14.285%, 5
Signature and Date	Boya Douho| Dec 6, 2024

Your Name	Kazi Zarin Tasnim Rafa
Student ID	30233931
Contribution (%) and Hours	14.285%, 5
Signature and Date	Kazi Zarin Tasnim Rafa Dec 7, 2024

Your Name	Ashkan Einiaghdam
Student ID	30270232
Contribution (%) and Hours	14.285%, 5
Signature and Date	Ashkan Einiaghdam Dec 8,2024

Your Name	Ray Pan
Student ID	30265201
Contribution (%) and Hours	14.285%, 5

Signature and Date	Ray Pan | Dec 8, 2024

Your Name	Socretes Saha
Student ID	30264159
Contribution (%) and Hours	14.285%, 5
Signature and Date	Socretes Saha| Dec 8, 2024

Your Name	Chunsheng Xiao
Student ID	30066914
Contribution (%) and Hours	14.285%, 5
Signature and Date	Chunsheng Xiao, Dec 8.2024

** Contribution total should be 100%.*

Table of Contents

1. Abstract

Wildfires can cause great damage and loss due to their effect on the economy, the environment, and society, but specifically in Alberta, this has become more common with time as the climate and mankind have changed. This study proposes a predictive model that has been developed with the aim of estimating and predicting the chances of occurrence of wildfires by machine learning techniques. This method integrates Random Forest Classifier and uses various input including Sentinel satellite images and weather data from Google Earth Engine API and National Weather Service API. The datasets are pre-processed and transformed using the Python libraries Rasterio, Pandas, NumPy, Scikit learn and Seaborn for efficient and effective analysis and visualization.

The model was able to achieve an accuracy of 90% in estimating wildfires when the datasets were balanced. Nevertheless, some drawbacks such as class imbalance within the testing datasets (e.g. overfitting towards "No Fire" category) raise premises for future research. Temperature, Vegetation indices (NDVI), and wind speed were the factors that came up as being the most significant in the fire risk assessment. This gulf of information indicates the applicability of the hybrid model in providing location specific decision-making aids on wildfires and especially on how best to manage and control the wildfires and the emergency response systems while conserving the environment and protecting the infrastructure.

This work clearly highlights the benefit and potential of combining remote sensing with machine learning in addressing difficult natural catastrophes. Future directions include increasing data variety, expanding the model to different areas, and performing practical experiments to confirm its validity.

2. Introduction

Wildfires are devastating natural disasters that threaten lives, communities, and the environment. In Alberta, Canada, they have become increasingly common and severe due to climate change, natural events like lightning, and human activity. Their unpredictability highlights the urgent need for better tools to predict and manage these risks.

This project tackles that challenge by building a wildfire prediction model using machine learning. Combining satellite imagery from Google Earth Engine with historical weather data, the model employs advanced methods like Random Forest Classifier to analyze and predict wildfire risks with high accuracy.

The goal is to create a practical tool that helps communities prepare for and respond to wildfires, reducing their impact. By leveraging technology like remote sensing and machine learning, this project not only addresses Alberta's wildfire challenges but also offers a scalable solution for other regions facing similar threats. (10)

3. Problem Statement

As both prevalence and unpredictability of wildfires go up, so does the pressing need for reliable predictive tools to mitigate their impacts. No exception is Alberta, Canada, which faces an alarming increase in

wildfires due to ongoing changes in environmental conditions that threaten life, ecosystems, and infrastructure. Many traditional prediction methods have failed due to various limitations: insufficient resolution of data, class imbalance in the occurrences of wildfires, and oversimplified models not capturing the complex interactions among the environmental factors. (2)(6)

The project addresses these challenges through the application of a Random Forest classifier, selected because of its robustness and interpretability with regards to handling diverse environmental datasets. (2). There are, however, some remaining drawbacks. These consist of class imbalances within the training set that may sway the predictions away from wildfires, as well as environmental key parameters, like NDVI, becoming sensitive to noise from clouds and snow interference (3). One of the weaknesses has to do with the regional emphasis on Alberta which affects the applicability of the results to other regions. Reliance on static fire intensity thresholds may further oversimplify the dynamic nature of wildfire behavior, hence introducing inaccuracies.

Nonetheless, this approach utilizes multisource data such as vegetation indices, temperature variations, soil moisture, and wind speed to make precise predictions of wildfire probabilities. This project ought to be considered as an important tool in addressing the regional variabilities and scaling problems associated with wildfires risk assessment and imparting information in forms that can be used for improving activities to deal with disasters on local and global scales.

4. Literature Review

Wildfire prediction has been an active area of research in recent years. Most of the machine learning approaches have shown promising performance, contributing to improving the accuracy and reliability of wildfire occurrence prediction. Early-stage risk assessment has been performed using conventional methods such as logistic regression, showing 84.4% success rates by means of parameters like temperature and humidity (Nikova & Deliyski, 2023). More advanced techniques, such as Decision Trees and Random Forests, have increased interpretability and stability, making them quite appropriate for complex non-linear environmental relationships (Collins et al., 2018). Advanced deep learning models, including Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, have proved to be powerful tools in the analysis of spatial and temporal data; for example, with an accuracy of wildfire detection up to 98.47% (Guo et al., 2022).

While these are major developments, still large gaps exist in the integration of a variety of data sources and ensuring model adaptability to heterogeneous regions. Most of the studies isolate models that deal with a small dataset; this makes their solutions less applicable in varied environments such as Alberta. The Random Forest model, therefore, becomes the center of attention, given the above challenges, since it has already been proved to handle complex datasets with diversified environmental features such as temperature, NDVI, soil moisture, and wind speed. Using data from Sentinel satellite imagery combined with historical meteorological records, the Random Forest approach improves the predictive accuracy of such models and hence offers a reliable region-specific tool in wildfire risk assessment and early warning.

5. Methodology

I. Data Collection and Preprocessing

Data Collection is an integral part of our research since it is what we will be using to train the machine learning model. Therefore, the type of data that is collected is very important and must be picked carefully. In this project all the data that is collected is continuous. We have sourced our data from MODIS and ECMWF. We collected NVDI (vegetation health index) data from MODIS, and the temperature (daily 2m air temperature), wind speed (combined u/v wind components) and soil moisture (volumetric soil water layer) were collected from ECMWF. The data collected was made to be custom to the region of Alberta. We define the region of interest using ee.FeatureCollection("FAO/GAUL/2015/level1") and used the desired region as the administrative boundary. The python libraries that will be used for the preprocessing were Panda and Google earth engine (ee).

Most of the data that was collected was type Json so before we could do any kind of manipulation or integration to the data, we needed to convert it into a CSV file. One that was done could begin preprocessing and filtering the data. We created a class called *FirePredictionModel* and created functions within the class for every data that needed to be processed, filtered, and down sampled. We also added those data to the Google earth engine API map to help with visualization. After all the training data was properly processed, we then exported it to our google drive. Afterwards we downsampled the data using Panda. We created a binary classification so we could train the classifiers by making a column labeled FireOccurred. The column that contained the fire data equaled 1 and the other was equaled to 0.

II. Model Training and Prediction

- **Model Selection:** Random Forest Classifier has been selected for binary classification (Fire/ No Fire classification).
- **Training Process:** For model training, we considered 70% training data (2015-2023) and 30% (2024) testing data.
- **Hyperparameter:**
  - Number of trees (n_estimators):
  - Minimum samples per split (min_samples_split)
  - Maximum depth of trees (max_depth)

Predictions

- Class labels: Class 0- No Fire; Class 1- Fire
- Fire Probability
- Confusion Matrix: True Positive, True Negative, False Positive, and False Negative
- Spatial Representation: GeoTIFF Raster files
- Evaluation Metrics: Precision, Recall, F1 score, and Accuracy

6. Results & Discussion

a. Model Evaluation

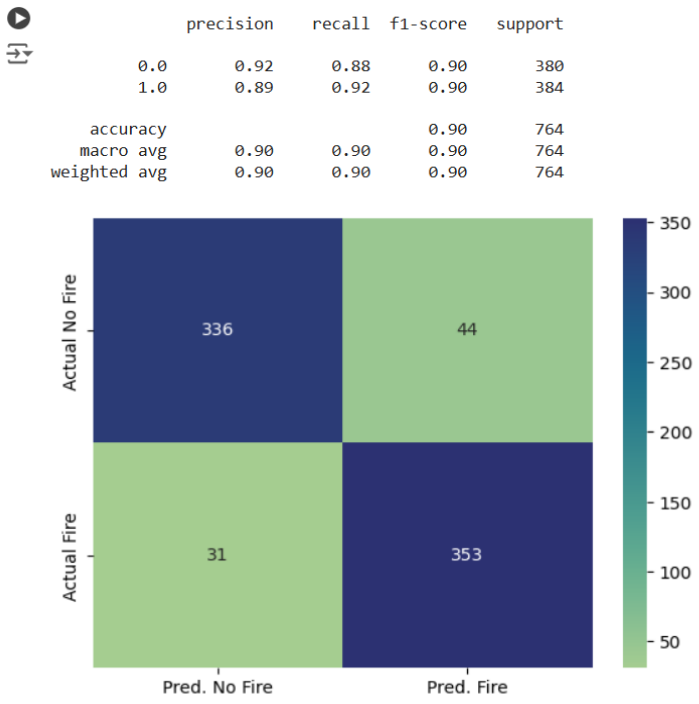
The model evaluation was based on fire risk prediction on

- ❖ The combined 2015–2024 dataset
- ❖ The trained model (2015-2013) to the 2024 testing dataset

The model performance metrics that reflected on the evaluation are:

1. Precision
2. Recall
3. F1 score
4. Accuracy

Case 1: The combined 2015-2024 dataset



	precision	recall	f1-score	support
0.0	0.92	0.88	0.90	380
1.0	0.89	0.92	0.90	384
accuracy			0.90	764
macro avg	0.90	0.90	0.90	764
weighted avg	0.90	0.90	0.90	764

1. **Precision:**
   While Class 0 (No Fire) is 0.92, the Class 1 (Fire) is 0.89, showing 92% "No Fire" incidents are correct whereas 89% of predicted "Fire" instances happen to be accurate.

2. **Recall:**
   Recall for Class 0 (No Fire) is 0.88 and for Class 1 (Fire) is 0.92, indicating 88% accuracy for "No Fire" incidents and 92% accuracy for "Fire" predictions. The recall for class 1 is important because predicting fire incorrectly (False Positives) is more fitting than missing actual fire incidents (False Negatives). Overall, this model performed significantly better for "Fire" cases.

3. **F1 Score:**
   It is the trade-off between precision and accuracy. F1 score is 0.90 that reflects a balanced model performing well for "No Fire" and "Fire" both instances.

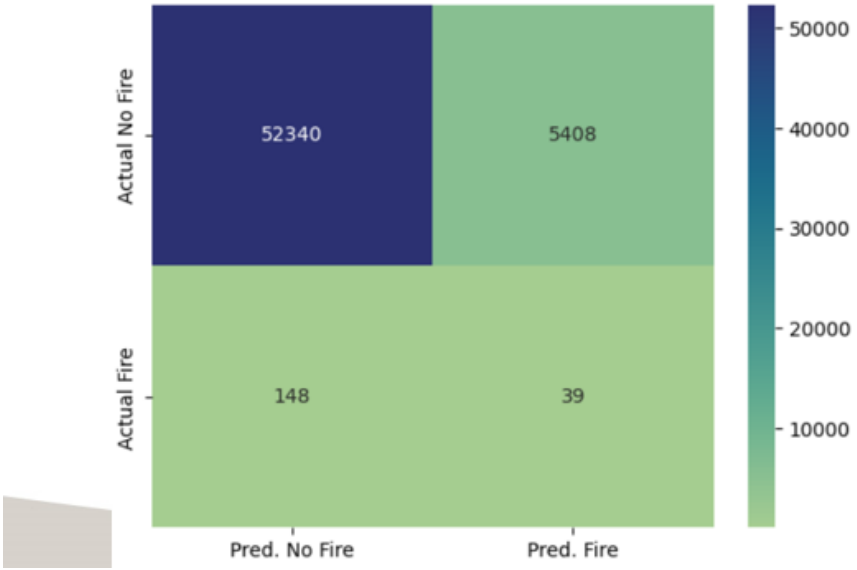
4. **Accuracy:**
   The accuracy of 0.90 interpreting that the model's overall 90% of all predictions (Fire and No Fire) are accurate.

Overall observation for Case 1:

- In case of bias, the model shows a slight overweight to the No Fire incidents since the True Negative (336) is relatively higher than the False Positive (44).
- For Class balance and Real-time implementation, the support values (380 NF vs 384 F) indicate a rather near-balanced dataset. That way, the metrics become accurate representations of real-time performance.

Case 2: The trained model (2015-2013) to the 2024 testing dataset

	precision	recall	f1-score	support
0.0	1.00	0.91	0.95	57748
1.0	0.01	0.21	0.01	187
accuracy			0.90	57935
macro avg	0.50	0.56	0.48	57935
weighted avg	0.99	0.90	0.95	57935



1. **Precision:**
   While Class 0 (No Fire) is 1.00, Class 1 (Fire) is only 0.01, showing 100% of "No Fire" incidents are correct whereas only 1% of predicted "Fire" instances happen to be accurate. This emphasizes the severe overprediction/class imbalance of Fire instances.
2. **Recall:**
   Recall for Class 0 (No Fire) is 0.91 and for Class 1 (Fire) is 0.21, indicating 91% accuracy for "No Fire" incidents and only 21% of actual fire events were identified but missed the remaining 79% of fire occurrences.
3. **F1 Score:**
   It is the trade-off between precision and accuracy. The F1 score for Class 1 is only 0.01 that reflects the poor model performance.
4. **Accuracy:**
   The accuracy of 90% is misleading as the model is highly biased towards the majority class (No Fire) events.

Overall observation for Case 2:

- The dataset significantly represents class imbalance. While the No Fire data is 57748, the actual fire data is only 187. The model shows significant bias towards No Fire events, making it the majority class.
- The fire precision for Class 1 is only 1%. It means that during the model training, the model encountered extremely few Fire examples which is why the model highly overpredicted the False Positives. The model cannot effectively differentiate between No Fire and Fire incidents.
- Due to lack of distinction between Fire and No Fire, the feature ambiguity (Soil Moisture, NDVI, etc.) arises, leading to confusion.

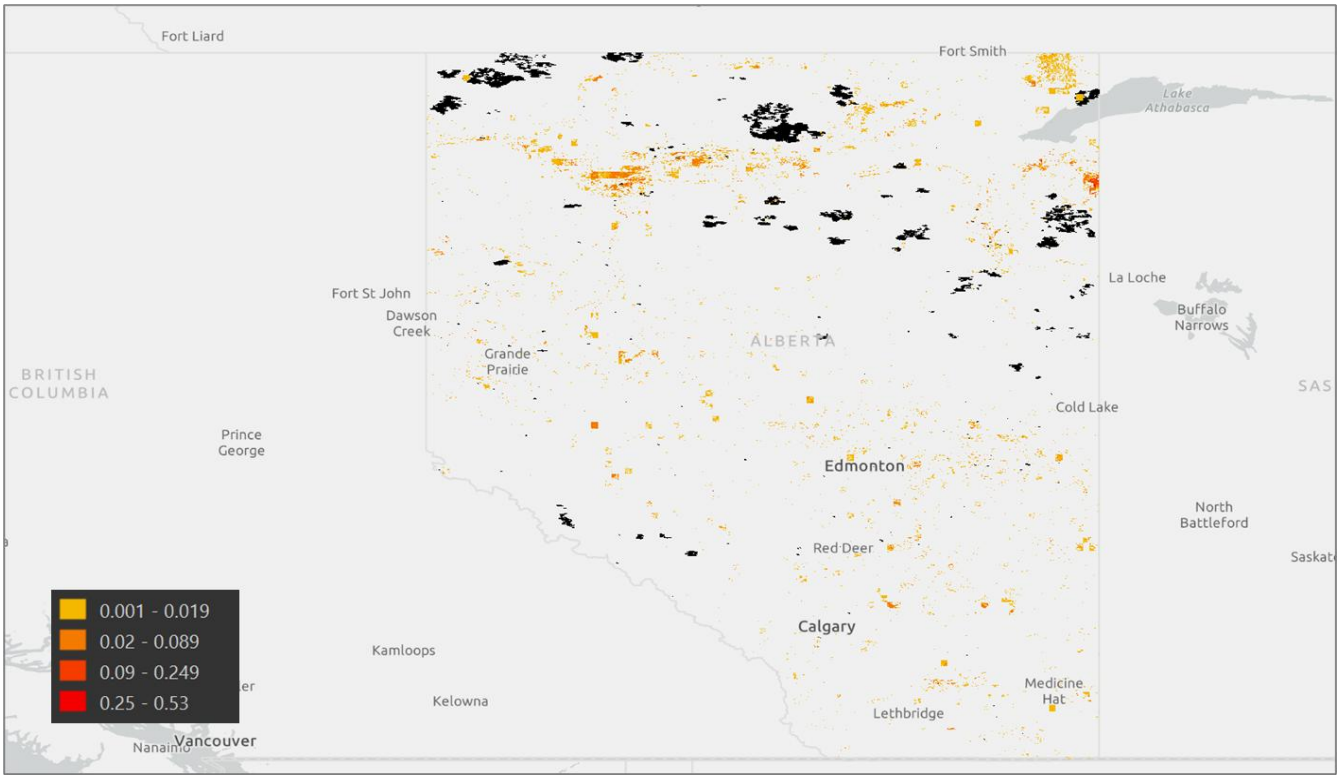
Remarks (based on Case 1 and Case 2):

Case 1 is more practically applicable due to its practical relevance to the real-world implications.

b. Insights from Feature Analysis

1. **Temperature:**
   Extension of high temperature leads to significant fire risk as it dries out the vegetation and soil.
2. **NDVI:**
   Low NDVI regions are highly likely to experience fire whereas high NDVI indicates less susceptibility to fire.
3. **Relative Humidity and Soil Moisture:**
   Low soil moisture indicates dry conditions that increase the risk of fire.
4. **Wind Speed:**
   Wind speed can intensify the fire spread. High wind speed regions are more prone to fire risks even if the other features are balanced.

c. Visualization:



The map depicts the spatial distribution of wildfire probabilities across Alberta, Canada. The color-coded legends indicate the fire probability at various levels from the model's prediction. And the black spots indicate the real wildfire data.

d. Key observations:

Zone	Color	Risk Probability	Area Details and Contributors
High-risk	Red and Black	0.25-0.53	Northern Alberta Dense vegetation and extended high temperature can be observed.
Moderate risk	Orange	0.09-0.249	Regions close to Edmonton and Red Deer, scattered across Central Alberta Irregular rainfall pattern
Low risk	Yellow	0.001-0.089	Dispersed zones in Southern Alberta (near Calgary) Urban prevalence
Uncolored	White	Negligible or no fire predictions	No definite details

7. Limitations

- **Class Imbalance:** The dataset has much less fire data than non-fire data. This is due to the nature of the fire dataset. We tried multiple solutions to mitigate the effect. Two examples are finding as much fire data as possible and applying a class weight to the random forest classifier.
- **Data Gaps:** Some data and features are missing from our dataset. Jain et al. (2020) found that wildfire is greatly correlated with human activity. The more human activities in the region, the more likely wildfire will occur. Lightening is another factor of wildfire. These factors are not included in our dataset and since then the model cannot give a good performance.
- **Data Granularity:** Granularity refers to the level of details stored in the dataset. In our project, the dataset only consists of monthly mean data which loses many details. This means that the dataset does not have good data granularity. However, wildfire might relate to some extreme but ephemeral weather data.
- **NDVI Accuracy:** NDVI is sensitive to cloud and snow interference. Since then, the accuracy of NDVI values might change in cloudy conditions.
- **Regional Focus:** Model is trained only with Alberta data. This means the model is tailored to Alberta and it requires further training for other places.
- **Static Thresholds:** Fixed fire intensity thresholds may oversimplify dynamics.

8. Future Directions

The algorithm we created functioned as intended for the most part, but we can make it better. The goal of this project is to create an accurate and reliable wildfire early prevention predictor to better assist first responders. Focusing on this goal we believe that the future direction of this project should include but not be limited to improving the region of interest, increasing the range of detection, incorporating cross-disciplinary collaboration, vegetation dynamics, etc. Our key observation is that we should gather an even richer diverse set of data so that our model can perform at an even better level and resolve our current data limitations. We also suggest that we should perform real-world implementations using this algorithm to better understand its reliability.

9. Conclusion

The project reflects on the significance of machine learning to be a prominent mechanism for identifying and mitigating the risk prediction for wildfires since it provides an analytical approach for optimizing resource allocation and implementing preventive measures. These insights help policymakers and emergency response teams to effectively recognize and handle susceptible fire zones. That way, the communities can develop targeted strategies for wildfire events.

Our project aimed to address a few significant existing gaps in wildfire assessment and predictive modeling and tried to determine the complexity to help approach future directions. Through our modeling in the Random Forest Classifier, we observed that there lies a complex relationship between wildfire prediction and key features with the imbalanced dataset. Missing features can initiate that. Overcoming the class imbalance, and low precision can become quite a challenge which needs to be

highly emphasized for further research.

The Random Forest model performed the best under a balanced dataset achieving 90% accuracy. It also provided critical spatial insights regarding the high fire-risk areas, especially in Northern Alberta. This shows that if the class imbalance can be overcome, Random Forest can demonstrate its highest effectiveness at handling complex larger data sets assuring accuracy. For future endeavors, it is quite important to address the limitations by incorporating advanced dataset balancing methods, more granular spatial data, temporal feature inclusion, etc. to enhance the model's robustness and reliability. This approach provides an adaptable outline for predicting wildfires with more precision in other regions facing equivalent environmental challenges as well. Real-time monitoring can further enhance the capabilities for more accurate prediction.

References

1. Abid, F., & Izeboudjen, N. (2020). Decision tree-based system on chip for forest fires prediction. *2020 International Conference on Electrical Engineering (ICEE)*, Istanbul, Turkey, 1-4. https://doi.org/10.1109/ICEE49691.2020.9249954
2. Collins, L., Griffioen, P., Newell, G., & Mellor, A. (2018). The utility of random forests for wildfire severity mapping. *Remote Sensing of Environment*, *216*, 374–384. https://doi.org/10.1016/j.rse.2018.07.005
3. Guo, Y., Chen, G., Wang, Y., Zha, X., & Xu, Z. (2022). Wildfire identification based on an improved two-channel convolutional neural network. *Forests*, *13*(8), 1302. https://doi.org/10.3390/f13081302
4. Nikova, H., & Deliyski, R. (2023). Binary regression model for automated wildfire early prediction and prevention. *2023 International Scientific Conference on Computer Science (COMSCI)*, Sozopol, Bulgaria, 1-5. https://doi.org/10.1109/COMSCI59259.2023.10315856
5. Singh, K. R., Neethu, K., Madhurekaa, K., Harita, A., & Mohan, P. (2021). Parallel SVM model for forest fire prediction. *Soft Computing Letters*, *3*, 100014. https://doi.org/10.1016/j.socl.2021.100014
6. Sayad, Y. O., Mousannif, H., & Al Moatassime, H. (2019). Predictive modeling of wildfires: A new dataset and machine learning approach. *Fire Safety Journal*, *104*, 130–146. https://doi.org/10.1016/j.firesaf.2019.01.006
7. Pérez-Sánchez, N., Jimeno-Sáez, N., Senent-Aparicio, N., Díaz-Palmero, N., & De Dios Cabezas-Cerezo, N. (2019). Evolution of burned area in forest fires under climate change conditions in Southern Spain using ANN. *Applied Sciences*, *9*(19), 4155. https://doi.org/10.3390/app9194155
8. Google Earth Engine API. (n.d.). *Landsat Surface Reflectance and Meteorological Data for Alberta*.
9. Jain, P., Coogan, S. C. P., Subramanian, S. G., Crowley, M., Taylor, S., & Flannigan, M. D. (2020). A review of machine learning applications in wildfire science and management. Environmental Reviews, 28(4), 478–505. https://doi.org/10.1139/er-2020-0019
10. Jain, P., Coogan, S. C. P., Subramanian, S. G., Crowley, M., Taylor, S., & Flannigan, M. D. (2020). A review of machine learning applications in wildfire science and management. Environmental Reviews, 28(4), 478–505. https://doi.org/10.1139/er-2020-0019