|  | 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  30233941  Ashkan Einiaghdam  30270232  Ray Pan  30265201  Socretes Saha  30264159  Chunsheng Xiao  30066914 |

| | *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* |  |  | *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* |  |  | *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* |  |   *\***Contribution total should be 100%.* |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |

Table of Contents

1. **Abstract**Pg.5
2. **Introduction**Pg.5
3. **Problem Statement** Pg.5
4. **Literature Review**Pg.6
5. **Methodology**Pg.6
6. ***Data Collection and Preprocessing***………………………………………………….. Pg.6
7. ***Target Definition and Features***……………………………………………………… Pg.7
8. ***Model Training and Prediction***……………………………………………………… Pg.7
9. ***Evaluation Metrics***…………………………………………………………………… Pg.7
10. **Results & Discussion**………………………………………………………………………… Pg.7
11. **Conclusion and Future Work**……………………………………………………………… Pg.8
12. **References**…………………………………………………………………………………... Pg.8

# 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 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.

**Introduction**

One of the most devastating natural disaster is wildfire which is ruinous hazard to ecosystems, communities, and human life which have been increased in frequency in Alberta, Canada, as a result of a combination of human activity, lightning, and climate change. By knowing that how unpredictable wildfires occur, it is obvious that new technologies are required to predict and control these hazards.

Concentration of this project is on predicting a wildfire model which use machine learning to overcome this growing challenge. By combining satellite images from google earth engine and historical weather data, the model analyzes patterns and predicts wildfire risks. In this project we use Random Forest Classifier method which aims to provide us with accurate insights.

The goal is simple but critical: to give communities a reliable tool to prepare for and respond to wildfires, reducing their impact on people and the environment. By blending cutting-edge technology with practical solutions, this project not only addresses Alberta’s wildfire problem but also offers a scalable approach that could help other wildfire-prone regions around the world.

# Problem Statement

The increasing prevalence and unpredictability of wildfires have heightened the high need for reliable predictive tools in mitigating their impacts. Alberta, Canada, is no exception, with its environmental changes that continued to spur the rise in wildfires that threaten life, ecosystems, and infrastructure. Most classic prediction methods normally fail on many grounds because of deficiencies in data resolution, class imbalance in the occurrence of wildfires, and simplifications of models which fail to represent complex interactions of environmental factors.

While this work tries to tackle most of these challenges through a hybrid machine learning framework that combines Random Forest classifiers, CNNs, and SVR models, there are some limitations. These limitations include but are not limited to imbalanced classes in the training dataset, sensitivities of NDVI features to cloud and snow interference, and a regional focus fit just for Alberta, which can only limit the generalization over other geographical regions. Furthermore, the dependence on static fire intensity thresholds of this model may oversimplify the complex dynamics of wildfire behavior. This gives further reason for possible inaccuracies: gaps in the available data and preprocessing steps.

This solution integrates multi-source data, such as vegetation indices, temperature fluctuations, soil moisture, and wind speed, for the most accurate prediction of wildfire probabilities, despite all the deficiencies. Overcoming broader challenges like regional variability and scalability, this project will deliver a critical tool for assessing the risks of wildfires-highly valued insights into disaster preparedness and response at the local and global scale.

# Literature Review

Prediction of wildfires has been an active area of research in recent years. Many machine learning methods have shown promising performance that contributes to enhancing the accuracy and reliability of wildfire occurrence prediction. Early-stage risk assessment has been performed using conventional methods such as logistic regression, which shows 84.4% success rates using parameters like temperature and humidity (Nikova & Deliyski, 2023). Advanced methods, including Decision Trees and Random Forests, enhance interpretability and robustness, making them suitable for complex, non-linear environmental relationships (Collins et al., 2018). Deep learning models like Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks have emerged as effective tools for spatial and temporal data analysis, achieving wildfire detection accuracies of up to 98.47% (Guo et al., 2022).

Despite these developments, major gaps still exist, particularly in the integration of a variety of data sources and ensuring model adaptability to heterogeneous regions. Most of the studies isolate models that involve small datasets; this makes their solutions less applicable in varied environments such as Alberta. Considering the challenges outlined above, the project zeroes in on the Random Forest model as it has been proven that this can handle complex data sets that include diverse environmental features like temperature, NDVI, soil moisture, and wind speed. Through leveraging data from Sentinel satellite imagery together with historical meteorological records, the Random Forest approach improves the predictive accuracy of these models, hence offering a reliable region-specific tool in wildfire risk assessment and early warning.

# Methodology (70% training, 30% testing)

***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 Google earth engine, 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 FAO/GAUL/2015/level1 and used the desired region as the administrative boundary.

Most of the data that was collected was type Json so before we could do any kind to 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 that had no fire and created a binary classification so we could train the classifiers.

**Model Training and Prediction**

* **Model Selection:** Random Forest Classifier has been selected for binary classification (Fire/ No Fire classification).
* **Training Process:**
* **Hyperparameter:**
* Number of trees (n\_estimators):
* Minimum samples per split (min\_samples\_split)
* Maximum depth of trees (max\_depth)
* **Class Weights:**

**Predictions**

* Class labels
* Fire Probability
* Confusion Matrix
* Spatial Representation

# Results & Discussion

**Model Evaluation & Visualization**

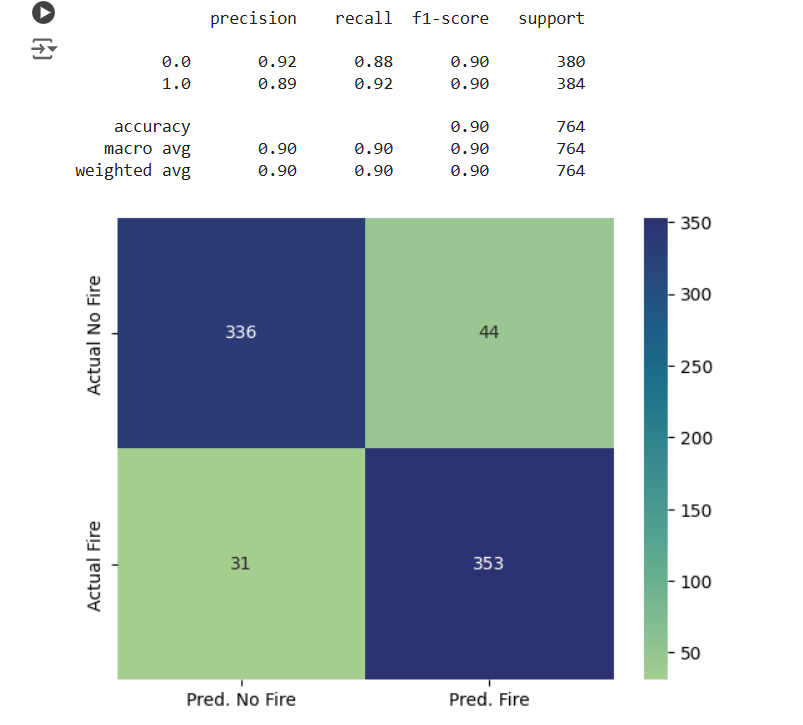
The model evaluation was based on fire risk prediction on

* 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**



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.

1. **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.

1. **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.

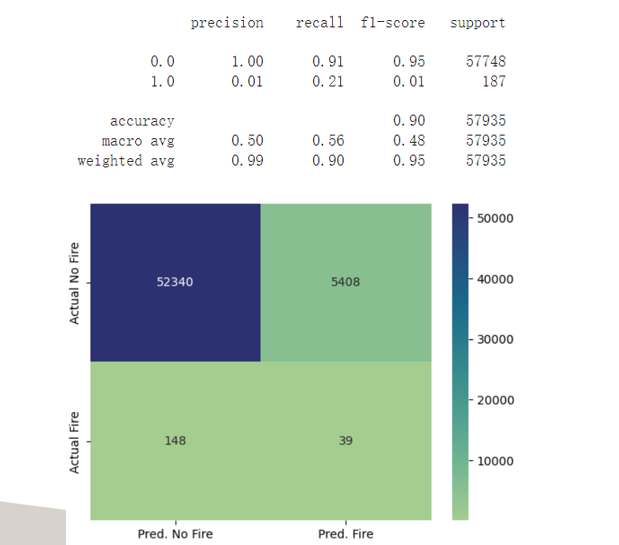
1. **Accuracy:**

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**



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.

1. **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.

1. **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.

1. **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.

**Insights from Feature Analysis**

1. **Temperature:**

Extension of high temperature leads to significant fire risk as it dries out the vegetation and soil.

1. **NDVI:**

Low NDVI regions are highly likely to experience fire whereas high NDVI indicates less susceptibility to fire.

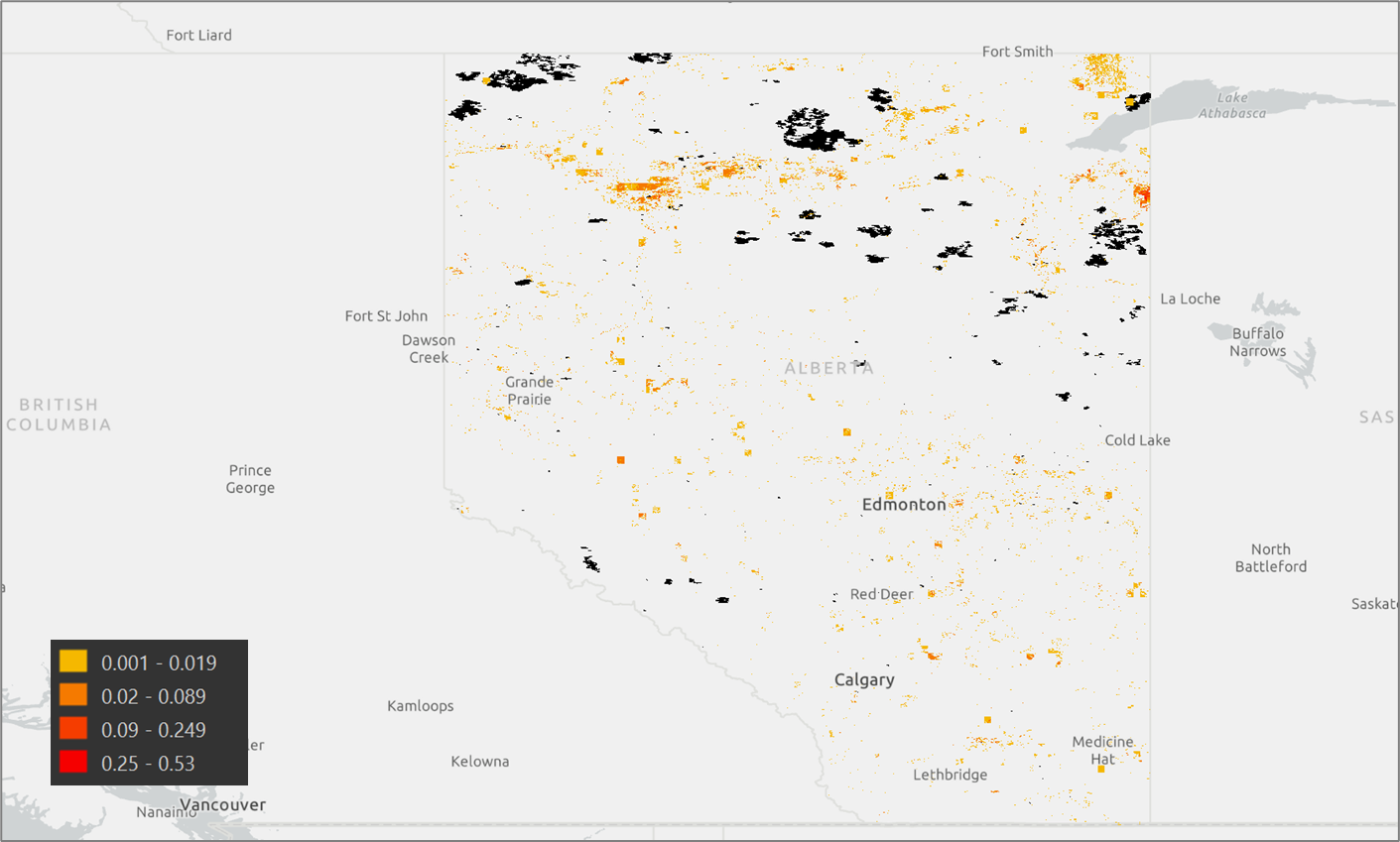
1. **Relative Humidity and Soil Moisture:**

Low soil moisture indicates dry conditions that increase the risk of fire.

1. **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.

**Visualization:**



**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 losses many details. This means that the dataset does not have a 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.

**Future Directions**

The algorithm we created functioned as intended 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 be to improve the region of interest and increase the range of detection. We also believe 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 believe that we should perform real-world tests using this algorithm to better understand its reliability.

**Conclusion**

# 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>