# Group 50 Progress Report:
# Neural Network-Based Wildfire Risk Classification

**Eric Solak, Tyler Yue, Ahmed Sahi**
{solake, yuet5, sahia8}@mcmaster.ca

## 1 Introduction

Wildfires are a growing threat to infrastructure, ecosystems, and human safety. In addition to the immediate danger of fire, wildfire smoke can travel thousands of kilometers, degrading air quality and affecting millions of people **?**. Predicting the likelihood of wildfires in advance can help authorities, first responders, and communities with preparation, resource allocation, and reduce potential damage.

This project aims to classify regions into low, medium, and high wildfire risk using a machine learning model trained on environmental and meteorological data. The resulting risk predictions can support proactive decision making and reduce the overall impact of wildfires on both people and the environment.

## 2 Related Work

Predicting wildfires is a sought after problem in both environmental science and machine learning. Existing approaches can be categorized into deterministic models, statistical models, and machine learning models.

Deterministic models, such as the Canadian Forest Fire Danger Rating System (CFFDRS) rely on empirical formulae derived from field experiments to estimate wildfire risk based on meteorological factors **?**. These models are generally used in operational settings but are limited in their ability to capture non-linear interactions between variables.

Statistical models include regression-based and Bayesian approaches that predict wildfire likelihood using historical fire and weather data **??**. These methods are effective in identifying trends but they are limited in their ability to capture rare or extreme weather conditions, which is critical for high-risk classification.

Machine learning approaches have used tree-ensemble methods such as XGBoost to classify wildfire risk based on tabular environmental data **?**. More recently however, neural networks have been trained on both satellite imagery and meteorological data to predict wildfire spread **?**. These models can learn complex, non-linear relationships but require preprocessing and normalization of inputs.

## 3 Dataset

For this project, we will utilize a Wildfire Prediction Dataset found on Kaggle **?**. This dataset contains 118858 entries and 17 attributes, with no missing values. Further, all values are numerical (float64), simplifying any data manipulation by avoiding issues revolving around categorical data.

### 3.1 Data Sources

There are two main sources used by this dataset: NASA's FIRMS VIIRS SCC system for fire radiative power (FRP) data, and Open-Meteo's meteorological data **??**. As these sources are open and publicly available, the data used is compliant with all terms of service.

### 3.2 Features

The dataset contains the following features:

- **daynight_N**: Indicator for whether the observation was taken during the day or night.
- **lat, lon**: Latitude and longitude coordinates of the location where the observation was recorded.
- **fire_weather_index**: An index representing the overall fire risk based on meteorological conditions.
- **pressure_mean**: Mean atmospheric pressure at the location.
- **wind_direction_mean, wind_direction_std**: Mean and standard deviation of wind direction.

- **solar_radiation_mean**: Average solar radiation received.
- **dewpoint_mean**: Average dew point temperature.
- **cloud_cover_mean**: Average cloud cover percentage.
- **evapotranspiration_total**: Total evapotranspiration over the observation period.
- **humidity_min**: Minimum humidity recorded.
- **temp_mean, temp_range**: Mean temperature and temperature range.
- **wind_speed_max**: Maximum wind speed observed.
- **occured**: Binary flag indicating whether a fire occurred (1) or not (0).
- **frp**: Fire Radiative Power, measuring the intensity of a fire if it occurred.

### 3.3 Preprocessing

The dataset requires minimal preprocessing due to its numerical format. The preprocessing steps used are:

- Clipping negative values: Negative *fire_weather_index* values are set to 0, as negative FWI is meaningless (i.e., negative risk is nonsensical).
- Feature normalization: All numeric features are standardized to facilitate stable training of the neural network. Each feature was scaled to have a mean of 0 and a standard deviation of 1, ensuring that features with larger ranges do not dominate the training process.

### 3.4 Data Split

A standard training/validation/testing split will be used, 70% for training, 15% for validation, 15% for testing.

## 4 Features

All 17 numerical features are used as input to the neural network. These features, described in Section 3.2, include positional, meteorological and environmental data. Positional features (*lat*, *lon*) provide the geographical context, important to the input as wildfire risk is often region-specific due to climate and vegetation. Meteorological features, such as *temp_mean*, *temp_range*, *humidity_min*, *pressure_mean*, *cloud_cover_mean*, *dewpoint_mean*, *solar_radiation_mean*, etc, describe the atmospheric conditions that directly influence fire ignition and propagation. Environmental features, including *evapotranspiration_total*, *fire_weather_index*, etc, describe the conditions that influence fire behaviour and ignition potential.

The model's target variable is the wildfire risk category, derived from the *fire_weather_index* and *occurred* variables in the dataset, classified into low, medium, or high risk. By combining multiple features that capture positional, meteorological, and environmental conditions, the neural network can learn complex, non-linear relationships that contribute to wildfire risk.

## 5 Implementation

Describe your model and implementation here. Refer to item 4. This may take around a page.

## 6 Results and Evaluation

Our model performs a three-way data split into training, test, and validation datasets. We used a stratified split to maintain the distribution of our target variable *risk_level* between the sets. The purpose of the splits are as follows:

1. **Training set:** used to fit the model parameters and to update the weights during backpropagation.
2. **Validation set:** used to determine performance during training and used for our early stop gradient descent and dynamic learning rate scheduling.
3. **Test set:** used to evaluate the model performance on unseen data and provide an accurate measurement of our model's ability to generalize and predict on new data.

Our evaluation metrics are based on the classification performance. We computed the precision, recall, and F1-score for each risk class, together with the overall accuracy and a confusion matrix. Accuracy serves as the primary evaluation metric, while the F1-score provides a more balanced measure of performance across all risk categories.

The neural-network classifier (three hidden layers with ReLU activations and Dropout) was trained on GPU for approximately 120 epochs with early stopping. The model achieved a validation accuracy of about 78 before convergence and a test accuracy of 78.5. The F1 scores show relatively good performance:

- **Low-risk (class 0):** F1 = 0.80
- **Medium-risk (class 1):** F1 = 0.74
- **High-risk (class 2):** F1 = 0.80

The confusion matrix shows that most misclassifications are between medium and high-risk wildfires. Precision and recall remain above 0.75 for all classes.

## 7 Feedback and Plans

During the project, our TA provided guidance on narrowing the scope of our wildfire risk classification task and ensuring that our model design aligns with the project requirements. In particular, we were advised to focus on using the numerical environmental and meteorological data effectively, rather than attempting to include complex satellite imagery or external data sources, which could overcomplicate the project given the time constraints. This advice helped us narrow the project scope and prioritize the features most relevant to predicting wildfire risk. The TA also emphasized the importance of properly preprocessing features, specifically the negative and extreme values in *fire_weather_index* to ensure stable neural network training.

Based on this feedback and reflections on the project's progress so far, our plan for the remaining weeks is to finalize the preprocessing pipeline to ensure all unrealistic values are handled appropriately (extreme *fire_weather_index* values). We also plan on refining the presentation of the results and conduct exploratory analysis. This includes generating detailed visualizations of feature importance using partial dependence plots to explain the model's predictions for specific regions. Including these graphs in the final report will help communicate both the strengths and limitations of the model.

## 8 Template Notes

You can remove this section or comment it out, as it only contains instructions for how to use this template. You may use subsections in your document as you find appropriate.

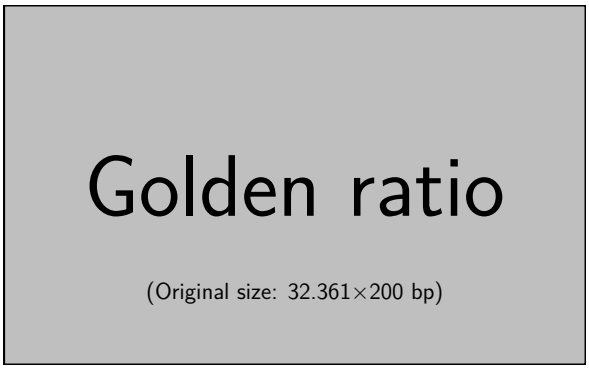


Figure 1: A figure with a caption that runs for more than one line. Example image is usually available through the `mwe` package without even mentioning it in the preamble.

### 8.1 Tables and figures

See Table 1 for an example of a table and its caption. See Figure 1 for an example of a figure and its caption.

### 8.2 Citations

Table 1 shows the syntax supported by the style files. We encourage you to use the natbib styles. You can use the command `\citet` (cite in text) to get "author (year)" citations, like this citation to a paper by **?**. You can use the command `\citep` (cite in parentheses) to get "(author, year)" citations (**?**). You can use the command `\citealp` (alternative cite without parentheses) to get "author, year" citations, which is useful for using citations within parentheses (e.g. **?**).

### 8.3 References

Many websites where you can find academic papers also allow you to export a bib file for citation or bib formatted entry. Copy this into the `custom.bib` and you will be able to cite the paper in the LaTeX. You can remove the example entries.

### 8.4 Equations

An example equation is shown below:

$$A = \pi r^2 \tag{1}$$

Labels for equation numbers, sections, subsections, figures and tables are all defined with the `\label{label}` command and cross references to them are made with the `\ref{label}` command. This an example cross-reference to Equation 1. You can also write equations inline, like this: $A = \pi r^2$.

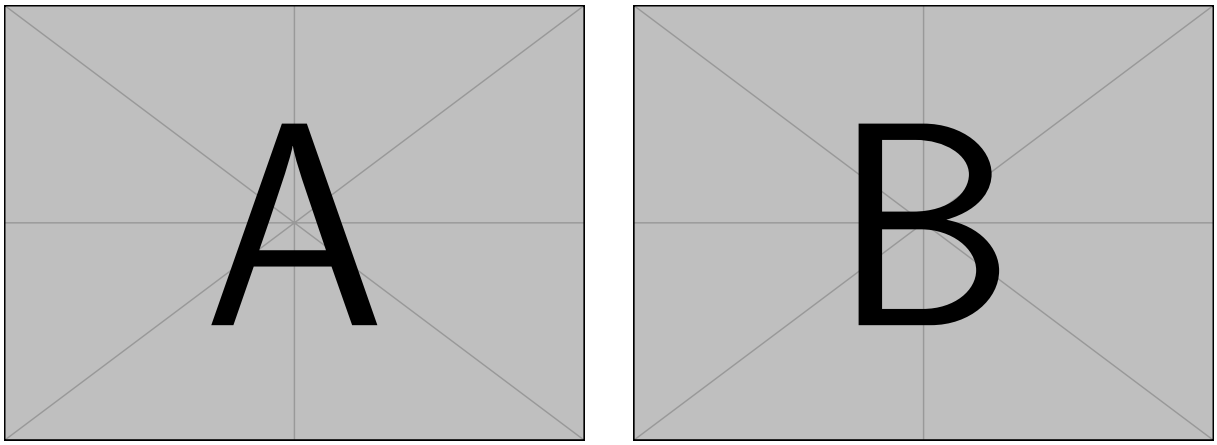


Figure 2: A minimal working example to demonstrate how to place two images side-by-side.

| Output | natbib command | ACL only command |
|---|---|---|
| (**?**) | `\citep` | |
| **?** | `\citealp` | |
| **?** | `\citet` | |
| (**?**) | `\citeyearpar` | |
| **?**'s (**?**) | | `\citeposs` |

Table 1: Citation commands supported by the style file.

## Team Contributions

- **Eric Solak:** Created Team Contract. Authored and completed Project Proposal sections *Overview*, and *Task Definition*. Set up the GitHub repository, including modularization, project structuring, separation of concerns. Authored and completed Project Progress Report sections 1 *Introduction*, 2 *Related Work*, 3 *Dataset*, 4 *Features*, and 7 *Feedback and Plans*.
- **Tyler Yue:**
- **Ahmed Sahi:**

## References