**BMEN619 Wildfire Detection Evaluation Criteria**

*Veronica Obodozie*

7 March 2025

**1.** Introduction

Wildfires are a major cause of concern as they increase methane emissions and cause issues. Due to the decreased predictability of wildfires, damage has increased. Efforts to improve the wildfire detection system have been made by various emergency services as it informs how to reduce the damage through early intervention [1]. This includes sensor networks, aerial monitoring, on-ground cameras, and satellite images. Recent efforts to combine the methods with Artificial Intelligence Models have been made [2], [3].

This assignment describes a proposed experimental design and an evaluation criterion for a wildfire detection model developed using the Sen2Fire dataset. This dataset consists of freely accessible raster data from the Sentinel 2 satellite which monitors **what?** and Sentinel 5 Precursor which has 2 bands of **what**. The ground truth of the dataset will be the Moderate Resolution Imaging Spectroradiometer (MODIS) Terra which monitors fires and hotspots [4].

The goal is to train a transparent and efficient binary classification model using supervised learning to detect if wildfire was detected in an area or not.

**2. Data Loading and preprocessing**

The dataset contains 13 bands which have 1 coastal aerosol, 5 colours (red, green blue, near-infrared), 4 vegetation red edge, 1 water vapour, 2 short-wave infrared, and 1 Ultraviolet Aerosol index [5] acquired with different wavelengths. This contains the image data, aerosol data, and labels, all in a NumPy zipped file. This data was resampled by the curators to ensure consistent spatial resolution between merged satellite data and no data is missing. Resampling methods and other pre-processing performed were not outlined.

Based on data exploration, the dataset is extremely imbalanced with 96.3% of the data labels being non-fire and 3.9% of the labels being fire-detected. **How to handle imbalanced data**

To avoid redundancy, and decrease computation time, specific bands within the dataset will be selected for model building. These are large images and will be resized to adapt to the model. A detailed preprocessing subsection will explain how these issues are handled.

**2.1. Data Preprocessing**

Data preprocessing is an important step in ensuring models are generalizable and reducing bias [6]. Random sampling can be implemented to mitigate the effects of dataset imbalance,

Exhaustive feature analysis is used to identify the relevant bands in the dataset and this will be extracted and used as the primary input data [7].

Handling

**3. Experimental Design**

This section highlights the tools and considerations for reproducibility. It also gives a breakdown of the proposed experimental design and the explainability of the model.

**3.1. Tools**

Python will be used to develop this detection algorithm and the PyTorch framework. The University Research cluster is the primary processing equipment, and resource availability and power are a major consideration, CPUs are generally.

The model will be obtained from a deeper dive into learning Python packages built using PyTorch.

Captuum is an open-source Python library built on Pytorch to increase interpretability among researchers. It will primarily be used to gain insights from explainability algorithms and models [8].

**3.2. Reproducibility**

Considerations for reproducibility

The dataset used is open source and is per-split for ease of.

and all tools and packages are free and in the public domain.

**3.3. Experimental Design**

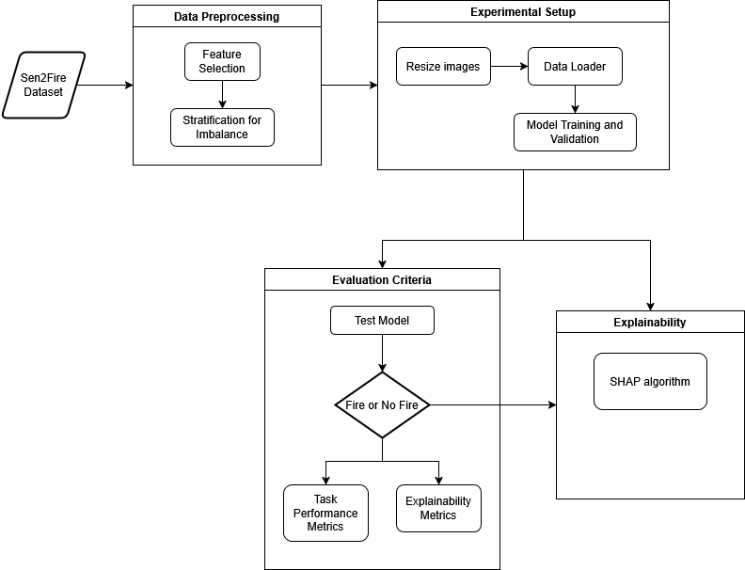
How was data split, what’s the ration

The dataset came pre-split into training, validation, and test sets. During curation, data was extracted from 4 distinct areas in Australia.

These were subsequently broken into patches, and to prevent data leakage the areas within the patch were used for specific development sections. So 50% of the dataset, areas 1 and 2, were used for training with an accompanied text file. 25% of the dataset, all from area 3, was used for validation. And 25% of the dataset Scene 4 was used for testing [5]. No rationale was specified for the split information, or how each scene was chosen for a specific test.

EXPERIMENTAL DESIGN: Data Augmentation, data loader

Transfer learning will be the method used to train the model, and based on the literature, Densely Connected Networks DenseNet [9]) is the model used for space-based applications. This is a deeper network than the ResNet-50 so the image size will be reduced to 96.



**Fig. 1.** Proposed Model Framework for Wildfire Detection Classification

**3.4. Explainability**

SHAP to find the minimum perturbation in the image

Test robustness of the system is an important part to evaluate

Explainability can be evaluated using specific methods to analyze robustness, interpretability, and contributing layers. A weighted histogram of the layers can be produced, and algorithms like Shapely Additive exPlanations (SHAP) are a post hoc interpretability method to describe the relationship between input and output [6]. Although SHAP has good results, it does not explain WHY the model makes its decision.

Gradient SHAP gives feature importance to input features.

**4. Evaluation Criteria**

This section focuses on the metrics used to evaluate the performance and explainability of the model. Based on the nature of the task, the metrics of evaluation used to measure performance that was used in similar wildfire detection classification models from literature shall consider the [10], [11]:

Accuracy measures the performance of the model, it is the ratio of the correct model predictions and the total labels.

|  |  |
| --- | --- |
|  | **(1)** |

Recall value should

|  |  |
| --- | --- |
|  | **(2)** |

Precision describes the reliability of the model.

|  |  |
| --- | --- |
|  | **(3)** |

The F1 score is a weighted average of precision and recall values

|  |  |
| --- | --- |
|  | **(4)** |

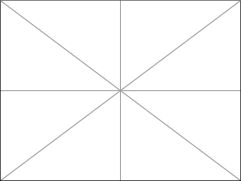
The confusion matrix for the True Positives and False Negatives of the predicted and target values will be created.

Task-appropriate metrics of evaluation

Data characteristics and imbalance

Selection of Metrics (task, imbalance, characteristics).

Evaluating the model’s explainability will require the use of infidelity and sensitivity metrics.



**Fig. 1.** Example of placing a figure with experimental results.

**5. References**

[1] A. Mohapatra and T. Trinh, “Early Wildfire Detection Technologies in Practice—A Review,” *Sustainability*, vol. 14, no. 19, p. 12270, Sep. 2022, doi: 10.3390/su141912270.

[2] A. M. Gonçalves, T. Brandão, and J. C. Ferreira, “Wildfire Detection With Deep Learning—A Case Study for the CICLOPE Project,” *IEEE Access*, vol. 12, pp. 82095–82110, 2024, doi: 10.1109/ACCESS.2024.3406215.

[3] K. Thangavel *et al.*, “Autonomous Satellite Wildfire Detection Using Hyperspectral Imagery and Neural Networks: A Case Study on Australian Wildfire,” *Remote Sensing*, vol. 15, no. 3, Art. no. 3, Jan. 2023, doi: 10.3390/rs15030720.

[4] L. Giglio and C. Justice, “MODIS/Terra Thermal Anomalies/Fire Daily L3 Global 1km SIN Grid V061.” NASA EOSDIS Land Processes Distributed Active Archive Center, 2021. doi: 10.5067/MODIS/MOD14A1.061.

[5] Y. Xu, A. Berg, and L. Haglund, “SEN2FIRE: A Challenging Benchmark Dataset for Wildfire Detection Using Sentinel Data,” in *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*, Jul. 2024, pp. 239–243. doi: 10.1109/IGARSS53475.2024.10641441.

[6] M. Arashpour, “AI explainability framework for environmental management research,” *Journal of Environmental Management*, vol. 342, p. 118149, Sep. 2023, doi: 10.1016/j.jenvman.2023.118149.

[7] Y. Dubey, A. Bhongade, P. Palsodkar, and P. Fulzele, “Efficient Explainable Models for Alzheimer’s Disease Classification with Feature Selection and Data Balancing Approach Using Ensemble Learning,” *Diagnostics (Basel)*, vol. 14, no. 24, p. 2770, Dec. 2024, doi: 10.3390/diagnostics14242770.

[8] *7. Explainability for Deep Learning Models*. Accessed: Mar. 02, 2025. [Online]. Available: https://learning.oreilly.com/library/view/explainable-ai-recipes/9781484290293/html/540435\_1\_En\_7\_Chapter.xhtml

[9] “8.7. Densely Connected Networks (DenseNet) — Dive into Deep Learning 1.0.3 documentation.” Accessed: Mar. 02, 2025. [Online]. Available: https://d2l.ai/chapter\_convolutional-modern/densenet.html

[10] M. Prakash, S. Neelakandan, M. Tamilselvi, S. Velmurugan, S. Baghavathi Priya, and E. Ofori Martinson, “Deep Learning-Based Wildfire Image Detection and Classification Systems for Controlling Biomass,” *International Journal of Intelligent Systems*, vol. 2023, no. 1, p. 7939516, 2023, doi: 10.1155/2023/7939516.

[11] R. Manoranjitham *et al.*, “Wildfire CNN: An Enhanced Wildfire Detection Model Leveraging CNN and VIIRS in Indian Context,” *TOCIEJ*, vol. 18, no. 1, p. e18741495324737, Oct. 2024, doi: 10.2174/0118741495324737240722111958.