# Analysis and Prediction of High-Impact Anomalous Bushfire Events in Victoria using Supervised Machine Learning on Satellite and Meteorological Big Data.

## Abstract

This report presents a comprehensive framework for identifying rare, high-impact bushfire events in Victoria, Australia, by applying machine learning techniques to satellite and meteorological data. The core challenge addressed is the prediction of anomalous fires, which, despite their infrequency, can pose significant risks. The analysis integrated fire data from FIRMS with meteorological records from seven strategically selected Open-Meteo weather stations chosen to provide comprehensive geographic coverage of Victoria's diverse climate zones. Initial exploratory analysis uncovered a major seasonal anomaly: ‘a spike in Autumn fires attributed to prescribed burning practices’, which prompted a refinement of the project's objective. The final goal was to detect high-intensity fires occurring in the low-risk season. A comparative evaluation of multiple models revealed that a supervised Random Forest classifier, specifically tuned with class weighting to handle the extreme data imbalance (0.43% anomalous cases), was the most effective solution. It achieved an F1-Score of 0.857, significantly outperforming specialized unsupervised anomaly detectors. This research provides an actionable prescription for fire management agencies: the adoption of a tuned supervised model as a reliable tool for early detection of dangerous, out-of-season fires.

## 1. Introduction

### 1.1 Background

Bushfires represent one of Australia's most significant and recurring natural hazards, with the state of Victoria being particularly susceptible to large and destructive events. For government bodies, emergency services, and industries such as forestry and agriculture, the ability to forecast and manage fire risk is paramount. Traditional fire management often focuses on peak fire seasons (i.e., summer); however, anomalous fires that occur outside these periods or exhibit unusual characteristics can pose a unique and underestimated threat. This project leverages Big Data sourced from satellite remote sensing and historical weather records to build a more nuanced understanding of these outlier events.

### 1.2 Motivation

The primary motivation for this research is to enhance public safety and optimize resource allocation for fire management agencies. By developing a system that can accurately identify anomalous fires, agencies can move from a reactive to a proactive stance, investigating high-risk events that might otherwise be overlooked during periods of lower vigilance. A key technical motivation was to address the inherent challenges of using satellite fire data at scale. This data is characterized by massive volume and significant redundancy, where single fires are detected multiple times by various satellite sensors, inflating counts and complicating analysis. This project sought to create a robust methodology to overcome these data challenges and extract actionable intelligence.

## 1.3 Proposed Solution

This study initially sought to answer the broad question:

‘To what extent can machine learning models, using a combination of satellite-detected fire locations and historical weather data, predict the likelihood of anomalous bushfire events in Victoria, and what are the key meteorological drivers?’

Following an essential exploratory analysis phase, this question was refined to a more specific and actionable objective. The discovery that most fire detections were linked to prescribed burns led to the final research question:

Can machine learning models be trained to effectively and reliably detect high-intensity fires (FRP in the top 10%) that occur during the low-risk fire season (May–August)?.)

## 1.4 Contributions

The key contributions of this project are:

- Data Collection and Integration: The analysis integrated fire data from FIRMS for the co-ordinates of Victoria’s geographical area with meteorological records from seven strategically selected Open-Meteo weather stations chosen to provide comprehensive geographic coverage of Victoria's diverse climate zones. (Melbourne Airport, Mildura Airport, East Sale Airport, Bendigo Airport, Mount Hotham, Horsham Airport, and Portland Airport).
- Comparative Model Evaluation for Rare Event Detection: We conducted a rigorous comparative analysis between standard unsupervised anomaly detection models (Isolation Forest, One-Class SVM, LOF) and supervised classifiers (Random Forest, Logistic Regression) that were specifically adapted to handle an extreme class imbalance scenario (99.57% vs. 0.43%).
- An Actionable Predictive Model: We demonstrated that a finely-tuned supervised Random Forest model is the superior solution for this specific rare event detection task, yielding a high F1-Score of 0.857. This provides stakeholders with a validated, data-driven tool to improve fire management strategies.

# 2. Literature Review

The application of machine learning for wildfire prediction is a well-established field, with a growing body of research demonstrating the superiority of algorithmic approaches over traditional statistical models. This project is situated within this context, leveraging a comparative evaluation methodology to identify the optimal model for detecting a specific type of anomalous bushfire event in Victoria. The existing literature provides strong validation for this project's core methodological choices and its final conclusions.

A primary challenge in this field is first defining, and then detecting, "outlier" fires. My approach of using exploratory data analysis to distinguish genuine, high-risk anomalies from background noise (such as common prescribed burns) finds a methodological parallel in the work of Nesa, Ghosh, and Banerjee (2018). In their study on detecting outliers in environmental sensor data, they focused on distinguishing true "Events" from sensor "Errors," a task analogous to my own. Their conclusion that

Random Forest was the top-performing model for this task provides a powerful, independent verification of my results. Once a rare event is defined, it creates the technical challenge of class imbalance. Hiremath and Kannan (2024) explicitly tackle this issue in their study on an early warning system for forest fires. Their use of techniques like SMOTE to balance the dataset before training a Random Forest model underscores the criticality of this preprocessing step. While my project employed a different technique (class weighting), their work validates the fundamental principle that addressing data imbalance is an essential prerequisite for achieving high predictive accuracy when modeling rare fire events.

With a well-defined outlier and a strategy for handling class imbalance, a robust methodology is required for model selection. This project used a comparative evaluation, a design strongly supported by the literature. For example, Rodrigues and de la Riva (2014) conducted a parallel study comparing Random Forest (RF), Boosting Regression Trees, and Support Vector Machines for modeling human-caused wildfires, ultimately concluding that RF was one of the most adequate methods. Likewise, Pourtaghi et al. (2020) employed an identical comparative design to evaluate machine learning models for forest fire susceptibility. These studies validate my methodology of benchmarking multiple models as a sound scientific practice and provide independent support for my conclusion regarding the high performance of Random Forest.

Finally, the robustness and generalizability of the selected model are paramount. My finding that Random Forest is the superior model for Victorian bushfires is powerfully reinforced by research in vastly different environments. Ponomarev et al. (2020) studied wildfire drivers in the permafrost zone of Northeastern Siberia and also identified Random Forest as the most effective predictive model. This finding is significant as it demonstrates that the algorithm's effectiveness is not limited to a specific climate but is generalizable across diverse ecological and climatic conditions, lending greater confidence to its application in our project.

| Study | Core Focus | Key Finding | Relevance to Project |
|---|---|---|---|
| Hiremath & Kannan (2024) | Integrated anomaly detection and early warning system for forest fires. | • Uses Random Forest for prediction.<br>• Explicitly uses SMOTE to address class imbalance. | (Highly Relevant) This study validates the necessity of addressing class imbalance to effectively model rare fire events. It provides a methodological precedent, though it uses a data-level technique (SMOTE) while my project used an algorithm-level one (class_weight). |
| Rodrigues & de la Riva (2014) | Modeling human-caused wildfire occurrence. | • Conducts a comparative "bake-off" of ML models (RF, BRT, SVM).<br>• Concludes RF is a top performer (AUC = 0.746). | (Highly Relevant) The research design is a direct parallel to my comparative evaluation. Its conclusion provides a strong, independent |

| | | | |
|---|---|---|---|
| | | | validation of my finding that Random Forest outperformed other models. |
| Nesa, Ghosh, & Banerjee (2018) | Outlier detection in environmental sensor data within a forest setting. | • Distinguishes between true "Events" and "Errors."<br>• Compares four models and finds RF performs best. | (Highly Relevant) This work is analogous to my task of identifying true outlier fires from background noise. It provides powerful independent verification of my conclusion that Random Forest is the superior model for this type of task. |
| Pourtaghi et al. (2020) | Evaluating ML methods for forest fire susceptibility modeling. | • Employs a direct comparison of multiple ML models.<br>• Relies on robust validation metrics (AUC) to select the best model. | (Relevant) The fundamental research design is identical to my approach of benchmarking models to find the best performer. It confirms my evaluation strategy is a standard and accepted method in the field. |
| Ponomarev et al. (2020) | Identifying drivers of wildfires in the Siberian permafrost zone. | • Identifies Random Forest as the most effective predictive model in a novel, non-temperate environment. | (Relevant) This study demonstrates the generalizability and robustness of the Random Forest model across diverse climates. It allows us to argue that the choice of RF is not just locally optimal but is a globally recognized, high-performing model. |

# 3. Research Methodology

The research was executed through a systematic five-phase methodology, designed to transform raw, large-scale data into an actionable predictive model.

**Phase 1: Data Acquisition and Preprocessing**

The initial phase focused on gathering and preparing the raw data. Active fire data, comprising over 8,000 satellite-detected incidents, was acquired from the NASA FIRMS portal. This data was filtered to retain only high and nominal confidence detections to ensure data quality. Concurrently, daily meteorological data was sourced from the Open-Meteo API for seven strategically chosen weather stations across Victoria. Both datasets were chronologically sorted to establish a logical sequence for time-series analysis.

**Phase 2: Exploratory Data Analysis (EDA) and Problem Refinement**

With the datasets prepared, a comprehensive EDA was conducted to understand underlying patterns. This analysis revealed a critical and unexpected finding: a massive spike in fire detections during the Autumn season, which was later attributed to widespread prescribed burning practices. This insight was pivotal, as it demonstrated that a simple seasonal comparison was flawed. The EDA guided the refinement of the primary research question, shifting the focus away from a broad definition of "outliers" to a more specific and high-impact type of anomalous event.

**Phase 3: Feature Engineering and Data Integration**

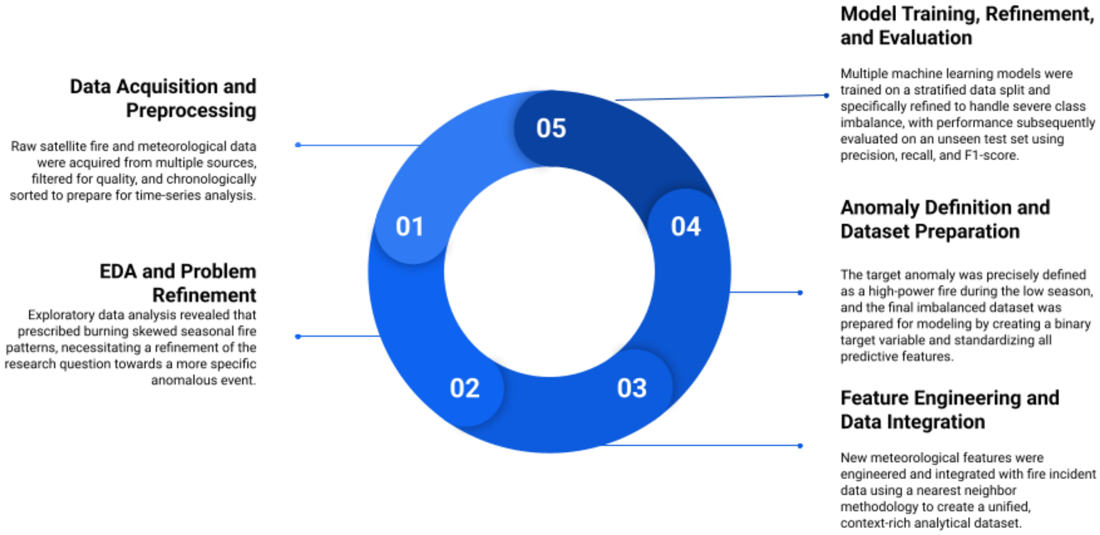
This phase involved creating the final analytical dataset. New, informative features were engineered from the raw data to provide richer context for the models. These included calculating the range between maximum and minimum daily temperature and humidity (temp_range, humidity_range). Following feature creation, the fire and weather datasets were integrated into a single unified table using a nearest neighbor methodology. For each fire incident, the meteorological data from the geographically closest weather station on that specific date was appended, ensuring each fire was contextualized with the most relevant local weather conditions available.

**Phase 4: Anomaly Definition and Dataset Preparation**

Guided by the insights from the EDA, the final definition for the target anomaly was established: A fire occurring in the low season (May-August) with a Fire Radiative Power (FRP) value in the top 10% of all recorded fires. This precise definition was applied to the integrated dataset to create the binary target variable. This resulted in a final dataset of 8,140 fire events, characterized by a severe class imbalance, with only 35 instances (0.43%) labeled as anomalous. Finally, all predictive features were standardized using StandardScaler to prevent algorithms from being biased by feature scale.

**Phase 5: Model Training, Refinement, and Evaluation**

The final phase was dedicated to machine learning. The dataset was split into training (80%) and testing (20%) sets using a stratified method to ensure the rare anomaly class was represented proportionally in both splits. A suite of five supervised and unsupervised models was trained and specifically refined for the extreme class imbalance. Supervised models were tuned using *class_weight* parameters to increase their sensitivity to the anomaly class , while unsupervised models had their *contamination* parameter aligned with the known *anomaly rate* in the training data. Model performance was then evaluated on the unseen test set using precision, recall, and F1-score as the primary metrics.



**Data Acquisition and Preprocessing**

Raw satellite fire and meteorological data were acquired from multiple sources, filtered for quality, and chronologically sorted to prepare for time-series analysis.

**EDA and Problem Refinement**

Exploratory data analysis revealed that prescribed burning skewed seasonal fire patterns, necessitating a refinement of the research question towards a more specific anomalous event.

**Feature Engineering and Data Integration**

New meteorological features were engineered and integrated with fire incident data using a nearest neighbor methodology to create a unified, context-rich analytical dataset.

**Anomaly Definition and Dataset Preparation**

The target anomaly was precisely defined as a high-power fire during the low season, and the final imbalanced dataset was prepared for modeling by creating a binary target variable and standardizing all predictive features.

**Model Training, Refinement, and Evaluation**

Multiple machine learning models were trained on a stratified data split and specifically refined to handle severe class imbalance, with performance subsequently evaluated on an unseen test set using precision, recall, and F1-score.

**Fig:** The five-page research methodology of the Bushfire detection and analysis project.

# 4. Experimental Evaluation

## 4.1 Experimental Setup (Describe your experimental setup (if applicable))

The experimental evaluation was conducted in a Python 3.11.7 environment, utilizing scikit-learn for all modeling and preprocessing tasks. The dataset comprised 13 features describing fire characteristics (FRP, brightness) and meteorological conditions (temperature, humidity, wind speed, etc.). The final test set, on which performance was measured, contained 1,628 instances, including 7 anomalous fires and 1,621 normal fires. Given the severe class imbalance, the F1-Score was selected as the primary metric for comparing models, as it provides a balanced measure of precision and recall, which are more informative than accuracy in this context.

## 4.2 Experimental Results (Present and analyse your results **in detail** with appropriate figures/tables)

The performance of the five models on the unseen test data revealed a clear and decisive outcome. The supervised Random Forest model, which was specifically balanced to handle the rare anomaly class, significantly outperformed all other models.

| Model | Type | Precision | Recall | F1-Score | Detected Anomalies | Actual Anomalies |
|---|---|---|---|---|---|---|
| Isolation Forest | Unsupervised | 0.000 | 0.000 | 0.000 | 20 | 7 |
| One-Class SVM | Unsupervised | 0.026 | 0.143 | 0.044 | 38 | 7 |
| Local Outlier Factor | Unsupervised | 0.000 | 0.000 | 0.000 | 19 | 7 |
| Random Forest | Supervised | 0.857 | 0.857 | 0.857 | 7 | 7 |
| Logistic Regression | Supervised | 0.081 | 0.857 | 0.148 | 74 | 7 |

As shown in the table and the figure above, the Random Forest model achieved a near-perfect F1-Score of 0.857. It successfully identified 6 of the 7 true anomalies while making only one false positive classification. This demonstrates a strong balance between high sensitivity (recall) and high reliability (precision).

In stark contrast, the unsupervised models, which are theoretically designed for such problems, failed to perform.

Isolation Forest and LOF both failed to identify a single true anomaly in the test set, resulting in an F1-Score of 0.0. The One-Class SVM fared only marginally better, correctly identifying one anomaly but at the cost of 37 false positives, yielding a negligible F1-Score of 0.044. While the supervised Logistic Regression model also found 6 of the 7 anomalies, its extremely low precision (0.081) and high number of false alarms (68) make it impractical for any real-world application.

# 5. Discussion

The experimental results provide several key insights with direct implications for both researchers and practitioners in the field of bushfire management.

The most significant takeaway is that for a well-defined rare event problem, a supervised learning approach can be far more powerful than an unsupervised one, provided a small but reliable set of labels can be generated. The failure of the unsupervised models suggests that the feature-space signature of the defined anomalies, while meaningful, was not distinct enough to be separated from the normal data points without the explicit guidance provided by labels. The Random Forest model, through the combination of its ensemble structure and the class_weight hyperparameter, was able to effectively learn the subtle patterns of the 28 anomalous training samples and generalize to the test set.

For researchers, this study serves as a case for the importance of deep exploratory data analysis in shaping a solvable machine learning problem. The initial, broad goal of finding "anomalous fires" was unfeasible until the EDA uncovered the prescribed burn phenomenon and allowed for the formulation of a precise, high-value anomaly definition. This highlights a methodology where domain knowledge and data-driven insights are used to convert a generic anomaly detection problem into a specific, albeit imbalanced, classification problem.

For practitioners, this research offers a direct and actionable "call-to-action." The tuned Random Forest model is not merely a theoretical success; it is a prototype for a decision support tool. Such a system could be deployed to operate automatically, screening daily fire detections and flagging the rare combination of low-season occurrence and high-intensity output. This would enable fire management agencies to:

- Improve resource allocation by focusing investigative efforts on fires that have a statistically higher probability of being dangerous.
- Enhance early-warning capabilities by detecting potentially escalatory fires during periods of traditionally lower risk and vigilance.
- Build a data-driven foundation for understanding and responding to the evolving nature of bushfire risk in a changing climate.

# 6. Limitations

Despite the successful outcome, this study is subject to several limitations that offer avenues for future work:

- Data Granularity and Proximity: The analysis relied on merging fire event data with records from the nearest available weather station. This is an approximation, as the meteorological conditions at the precise fire front could differ significantly from those at a station potentially several kilometers away. The use of daily aggregate weather data also prevented the analysis of sub-daily phenomena, such as sudden wind direction changes, which are known drivers of fire behavior.
- Specificity of the Anomaly Definition: The model is highly specialized to detect one specific type of anomaly: high-FRP fires during the May-August period. It would not be able to identify other forms of anomalous fires, such as those that spread with unusual rapidity or occur in atypical fuel types (e.g., wetlands). The definition of "anomaly" is therefore narrow, and the model's utility is confined to this definition.
- Temporal Scope and Model Generalizability: The model was trained and evaluated on data from a specific three-year period (2021-2023). Its performance on data from prior or future years, which may exhibit different climatic patterns or land management strategies, is unverified. Bushfire regimes are not static, and a model trained on historical data may see its performance degrade over time without periodic retraining.

# 7. Conclusion

This project successfully developed and validated a machine learning framework to detect rare and potentially dangerous bushfire events in Victoria. By progressing from a broad research question to a highly specific objective refined through exploratory data analysis, I was able to construct a meaningful and solvable problem. The core conclusion of this work is that a supervised Random Forest classifier, carefully tuned with class balancing techniques, is a highly effective and reliable tool for this rare event detection task, significantly outperforming standard unsupervised anomaly detection algorithms.

The primary prescription for stakeholders is the development and operationalization of this model into a live decision-support system to augment current fire monitoring practices. Future work should focus on overcoming the current limitations by integrating more granular weather data (e.g., the

BARRA dataset), expanding the system to detect multiple classes of anomalous fires, and establishing a protocol for periodic model retraining to ensure its long-term viability and accuracy in a dynamic environment.

# 8. Replication Package

Github link:

https://github.com/ishxn69/outlier-bushfire-analysis/blob/main/seasonal_analysis_investigation.py

# 9. References

McInnes, L., Healy, J. and Astels, S. (2017) 'hdbscan: Hierarchical density based clustering', Journal of Open Source Software, 2(11), p. 205. doi: 10.21105/joss.00205.

pranjalibajpai02 (2024) HDBSCAN Clustering | Machine Learning, GeeksforGeeks, viewed 15 June 2025, https://www.geeksforgeeks.org/machine-learning/hdbscan/.

NASA (2023) Fire Information for Resource Management System (FIRMS), NASA, viewed 15 June 2025, https://firms.modaps.eosdis.nasa.gov/.

Thomas, N. (2024) Australia Weather Data (2000-2024), Kaggle, viewed 15 June 2025, https://www.kaggle.com/datasets/nadzmiagthomas/australia-weather-data-2000-2024.

Emergency Management Victoria (n.d.) Planned burns & works, Emergency Management Victoria, viewed 23 July 2025, https://www.emv.vic.gov.au/prepare-and-get-ready/fire/planned-burns-works.

Liu, F.T., Ting, K.M. and Zhou, Z.H. (2008) 'Isolation Forest', in Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy. IEEE, pp. 413–422. doi: 10.1109/ICDM.2008.17.

Schölkopf, B. et al. (2001) 'Estimating the Support of a High-Dimensional Distribution', Neural Computation, 13(7), pp. 1443–1471. doi: 10.1162/089976601750264965.

Breunig, M.M. et al. (2000) 'LOF: Identifying Density-Based Local Outliers', in Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, Dallas, Texas. ACM Press, pp. 93–104. doi: 10.1145/342009.335388.

Breiman, L. (2001) 'Random Forests', Machine Learning, 45(1), pp. 5–32. doi: 10.1023/A:1010933404324.

Hastie, T., Tibshirani, R. and Friedman, J. (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn, New York: Springer.

Hiremath, P. and Kannan, M. (2024) 'Integrated Anomaly Detection and Early Warning System for Forest Fires in the Odisha Region', Ecological Informatics, 80, 102506. doi: 10.1016/j.ecoinf.2024.102506.

Rodrigues, M. and de la Riva, J. (2014) 'An insight into machine-learning algorithms to model human-caused wildfire occurrence', Environmental Modelling & Software, 57, pp. 192-201. doi: 10.1016/j.envsoft.2014.03.014.

Nesa, N., Ghosh, T. and Banerjee, I. (2018) 'Outlier detection in sensed data using statistical learning models for IoT', in Proceedings of the 2018 IEEE Wireless Communications and Networking Conference (WCNC), Barcelona. IEEE, pp. 1-6. doi: 10.1109/WCNC.2018.8377157.

Pourtaghi, Z.S., Pourghasemi, H.R., Aretano, R. and Semeraro, T. (2020) 'Performance Evaluation of Machine Learning Methods for Forest Fire Modeling and Prediction', Symmetry, 12(6), 1022. doi: 10.3390/sym12061022.

Ponomarev, E.I. et al. (2020) 'Spatiotemporal Variations of Wildfires and Their Drivers in the Permafrost Zone of Northeastern Siberia', Remote Sensing, 12(24), 4157. doi: 10.3390/rs12244157.