Project Report: Predicting Wildfire Size in the United States

Colton Hester (colton_hester@berkeley.edu), Leo Lazzarini (leandro.vieira@berkeley.edu),
Nedim Hodzic (nedim.hodzic@berkeley.edu), Shanti Agung (shanti.agung@berkeley.edu)

Abstract

Wildfires pose escalating risks to U.S. communities, with climate change driving increasingly severe fire seasons. We developed machine learning models to predict wildfire size categories (small, medium, large, very large) using the FPA FOD-Attributes dataset containing 2.3 million incidents and 308 environmental, meteorological, and social features. Our primary challenge was extreme class imbalance: 97% of fires remain under 100 acres. We evaluated eight preprocessing pipelines combining imputation strategies, imbalance handling (undersampling, SMOTENC, class weights), and location representations. We compared Random Forest (baseline), XGBoost, Feed-Forward Neural Networks, Logistic Regression, and Entity Embeddings. The FFNN with class weights achieved the best macro F1 score (0.49 validation, 0.47 test), demonstrating balanced performance across all fire size categories. Subgroup analysis revealed consistent performance across socioeconomic groups but regional disparities, with Western fires better predicted than Eastern fires.

Introduction

The risks posed by wildfires to communities, ecosystems, and economies have grown as climate change induces increasingly favorable conditions for severe fires. Events like the 2025 Eaton and Palisades fires in California demonstrate the urgent need for predictive tools to assist in proactive wildfire mitigation. The input to our algorithm is environmental, meteorological, and infrastructure features at ignition time from the Fire Program Analysis Fire-Occurrence Database, and we use Random Forest, XGBoost, FFNN, Logistic Regression, and Entity Embeddings to predict fire size category (small: 0-100 acres, medium: 100-4,999, large: 5,000-29,000, very large: 29,000+).

This four-class binning balances operational relevance with statistical tractability, as the original fire size distribution exhibits extreme right-skewness (mean=78 acres, SD=2,631, max=662,700). The primary technical challenge is severe class imbalance: 97.36% of fires are small, while only 0.05% exceed 29,000 acres. We systematically compare preprocessing strategies and modeling approaches to maximize macro-averaged F1 score—our primary metric for balanced class performance.

Related Work

Prior wildfire prediction research spans multiple methodological approaches. Statistical models using logistic regression established baseline performance for binary large/small fire classification (Preisler et al., 2004). Machine learning approaches including Random Forests and gradient boosting demonstrated improved predictive power by capturing non-linear feature interactions (Rodrigues & de la Riva, 2014; Jakovljević et al., 2018).

Recent work has explored deep learning for wildfire prediction. Radke et al. (2019) applied CNNs to satellite imagery for fire spread prediction, while Hodges & Lattimer (2019) used recurrent architectures for temporal fire behavior. However, most existing work focuses on binary classification rather than multi-class size prediction. The FPA FOD-Attributes dataset (Pourmohamad et al., 2023) provides unprecedented feature richness. Our work extends prior research by framing fire size as a four-class problem, systematically comparing class imbalance strategies, and evaluating model fairness across socioeconomic and regional subgroups.

Data

The FPA FOD-Attributes Dataset contains 2,302,521 U.S. wildfire incidents (1992-2020) with 308 variables covering physical (weather, climate), biological (vegetation indices), social (population density, vulnerability indexes), and administrative (preparedness levels, fire stations) attributes (Pourmohamad et al., 2023).

Data preprocessing: We dropped columns with >80% missing values and selected 46 features including GRIDMET climate variables, NDVI, fire station proximity, and preparedness levels. We used stratified 60/20/20 train/validation/test splits by FIRE_YEAR and FIRE_SIZE_LABEL (1,381,512 / 460,504 / 460,505 examples). Missing values were handled via zero and subgroup mean imputation (STATE as subgroup). To address extreme class imbalance, we compared: no handling, undersampling, SMOTENC oversampling, and class weights during training. Details of all eight preprocessing pipelines are in Appendix A.

Fire Size	Number of sample	Percent
Small	2,241,807	97.36
Medium	55,930	2.43
Large	3,682	0.16
Very large	1,102	0.05

Table 1: Class Imbalance Summary

Exploratory data analysis: Geographic analysis revealed that high-impact fires (≥5,000 acres) cluster predominantly west of −100° longitude, especially in California and the Mountain West. The Southeast contains 51% of fires but these are predominantly smaller. Natural (lightning) fires are 3× more likely to become high-impact than human-caused fires (6.0% vs 2.3%), explaining Western fire prominence. Fire station proximity shows protective effects: high-impact rates drop from 7.7% (1-2 stations) to 2.2% (10+ stations). See Appendix A for additional EDA visualizations.

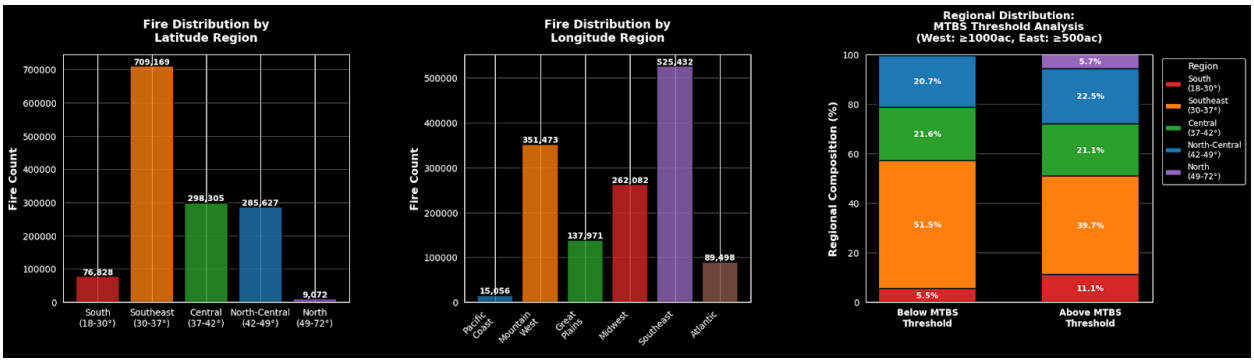

Figure 1: Regional Fire Distribution

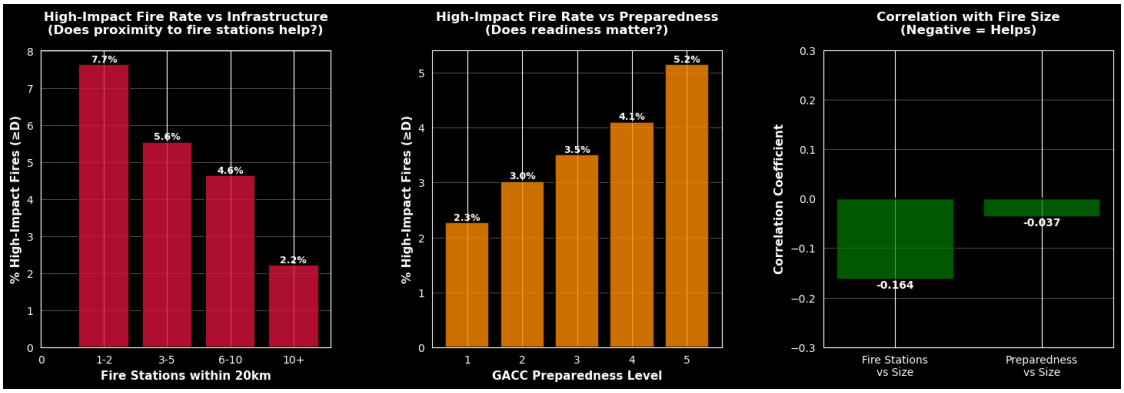

Figure 2: High-Impact Fire Rate vs. Infrastructure & Preparedness

Methodology

We implemented five algorithms: one baseline and four improvements.

1. **Random Forest (Baseline)**: Random Forest handles mixed data types naturally, provides robustness to noise, offers interpretable feature importance, and resists overfitting through bagging. We used 100 trees with default maximum depth and evaluated them using accuracy and confusion matrices.
2. **Feed-Forward Neural Network (Improvement):** FFNNs capture complex, nonlinear relationships in the data. Wildfire size may result from nonlinear interactions among meteorological, fire potential, administrative, and infrastructure factors. Our

architecture uses three dense layers (64, 32, 4 neurons) with ReLU activation, dropout regularization (0.3, 0.2), and softmax output, trained with Adam optimizer and sparse categorical cross-entropy loss.

3. **XGBoost (Improvement):** XGBoost sequentially builds decision trees where each new tree corrects previous errors. This sequential error-correction is well-suited for distinguishing boundaries between size categories. We configured XGBoost with default parameters and logloss evaluation metric.
4. **Entity Embeddings (Improvement):** This model learns dense vector representations for categorical features (NWCG_CAUSE_CLASSIFICATION mapped to 2D, GACC_PL to 4D embeddings). We applied extensive regularization: L2 penalty (λ=0.01), batch normalization, and aggressive dropout (50%, 30%, 25%). The limited benefit (macro F1 of 0.49 vs 0.49 for FFNN) is attributable to only 2 categorical features being available for embedding.
5. **Multiclass Logistic Regression (Improvement):** Appropriate when decision boundaries between classes are linear. Since meteorological features may be linearly related to fire size, logistic regression provides an interpretable comparison point.

Experiments, Results, and Discussion

Experiment Design: We conducted experiments at two levels: pipeline-level (comparing imputation, imbalance handling, and location representation strategies) and model-level (hyperparameter tuning). For FFNN, we tuned hidden layers, neurons, learning rate, batch size, epochs, and dropout rate. Due to class imbalance, we use macro F1 as our primary metric, which balances precision and recall across all classes.

Results: Class weights emerged as the clear winner for handling class imbalance, substantially outperforming SMOTENC across all models. With SMOTENC, validation accuracy ranged from 55.8% to 72.5%, but macro F1 scores remained low (0.33–0.44). Class weights achieved more balanced F1 scores: Random Forest (0.52), XGBoost (0.53), and FFNN (0.51). We selected FFNN with class weights as our final model due to its balanced per-class performance.

Class	P	R	F1
Small	0.58	0.64	0.61
Medium	0.44	0.34	0.38
Large	0.35	0.3	0.32
Very Large	0.5	0.65	0.56
Macro	0.47	0.48	0.47

Accuracy: 48.1%

Table 2: FFNN with Class Weights Results

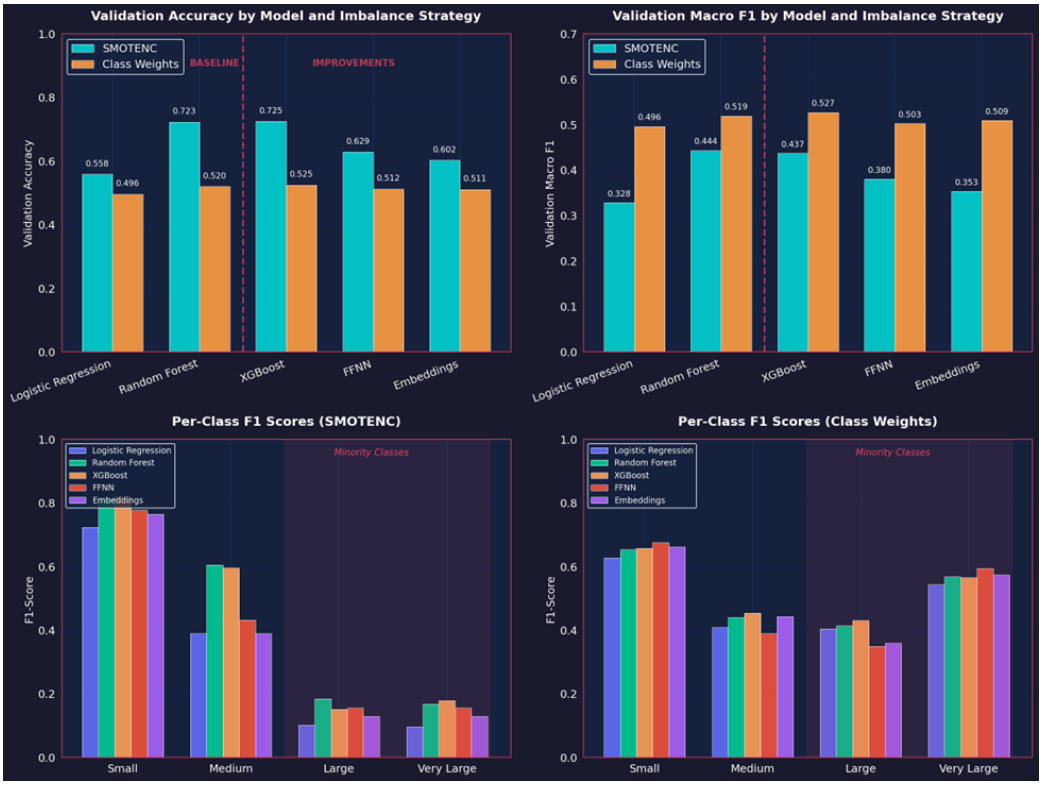

Figure 3: Model comparison across imbalance strategies. Class weights (orange) achieves more balanced per-class F1 scores than SMOTENC (cyan), particularly for minority classes (Large, Very Large).

Overfitting Mitigation: The FFNN exhibited minimal overfitting, as training and validation loss curves remained closely aligned, converging to nearly identical values (training ≈ 1.09, validation ≈ 1.08). We employed dropout regularization, class weights (avoiding synthetic noise from SMOTENC), early stopping, feature standardization, and stratified splits by year and fire size.

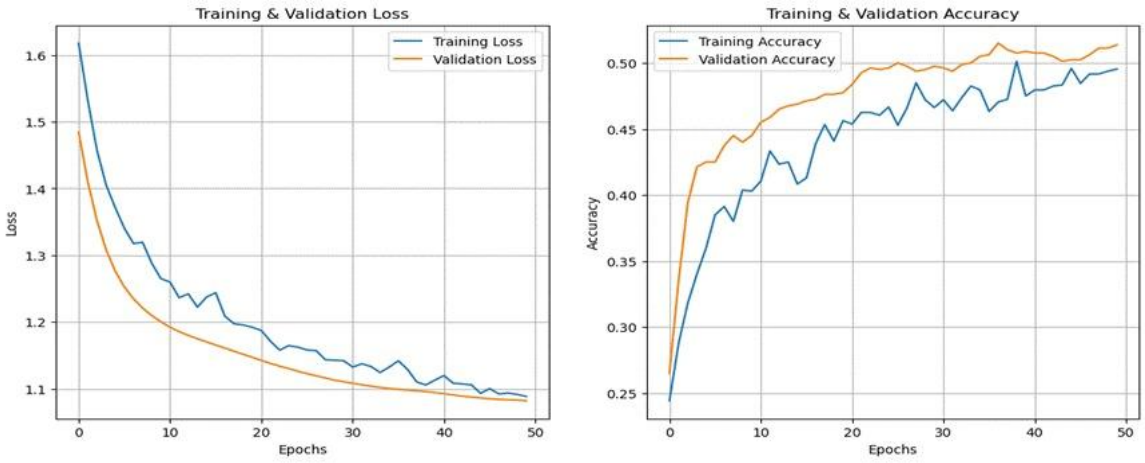

Figure 4: Training & Validation Loss / Accuracy Curves

Subgroup Analysis: Across socioeconomic groups (income status, population quartile, pollution burden), we found no meaningful differences in overall accuracy and macro F1. The model performs better at predicting very large fires in low-income counties (F1=0.64) than standard counties (F1=0.54). However, regional analysis revealed disparities: the model performs well in the Gulf States and Central U.S. but poorly in the North. The Pacific Coast achieves F1=0.67 for very large fires, while Midwest, Southeast, and Atlantic regions achieve F1=0. See Appendix D for detailed subgroup results.

Conclusions:

Our best-performing model for predicting U.S. wildfire size is the FFNN with class weights, achieving a test macro F1 of 0.47. The model performs best on small (F1=0.61) and very large fires (F1=0.56) but struggles with medium and large fires. Regarding fairness, the model shows no bias toward vulnerable socioeconomic groups but indicates regional disparities.

Limitations: The model does not account for time-series aspects of the data, and target variable binning does not reflect different MTBS thresholds for Eastern vs. Western U.S.

Future improvements: Include log transformation before binning, RNN models for temporal patterns, additional climate variables, and training regional models separately for Eastern and Western U.S.

Contributions

- Colton Hester: Data ingestion, undersampling pipelines, modeling experiments. Report sections: Abstract, Introduction, Related Work; revised report from 12 to 5 pages. Styled slide deck; presented introductory slides.
- Leo Lazzarini: Developed FFNN with class weights (final model). Designed model comparison framework. Created preprocessing pipeline (`src/preprocessing_pipeline.py`). Regional subgroup evaluation. Contributed to Results, Discussion, Methodology.
- Nedim Hodzic: Early binning strategy. Three model configurations with RF baseline, FFNN, XGBoost. East/West binary location variable. EDA visualizations. Presented on EDA, Conclusion, Improvements.
- Shanti Agung: Feature selection, preprocessing pipelines (subgroup mean imputation, SMOTENC). Built RF, FFNN, Logistic models. Result-consolidation templates. Contributed to preprocessing, methodology, experiments, discussion sections.

GitHub: https://github.com/ColtonHester/mids-w207-section1-team1-finalproject
Medium: https://medium.com/@nedimhodzic0111/predicting-wildfire-size-in-the-united-states-699de63de24c

References

Ham, Y. G., Nam, S. H., Kang, G. H., & Kim, J. S. (2024). Regionally optimized fire parameterizations using feed-forward neural networks. *Environmental Research Letters*, *20*(1). https://iopscience.iop.org/article/10.1088/1748-9326/ad984a/pdf

Jakovljević, G., Gigović, L., Sekulovíc, D., & Regodić, M. (2018). GIS Multi-Criteria Analysis for Identifying and Mapping Forest Fire Hazard: Nevesinje, Bosnia and Herzegovina. *Tehnicki Vjesnik*, *25*. https://doi.org/10.17559/TV-20151230211722

Joshi, J., & Sukumar, R. (2021). Improving prediction and assessment of global fires using multilayer neural networks. *Scientific Reports*, *11*(1), 3295. https://www.nature.com/articles/s41598-021-81233-4.pdf

Li, F., Zhu, Q., Yuan, K., Ji, F., Paul, A., Lee, P., Radeloff, V., & Chen, M. (2024). Projecting large fires in the western US with an interpretable and accurate hybrid machine learning method. *Earth's Future*, *12*(10). https://doi.org/10.1029/2024EF004588

Pourmohamad, Y., Abatzoglou, J. T., Belval, E. J., Fleishman, E., Short, K., Reeves, M. C., Nausea, N., Higuera, P., Henderson, E., Ball, S., AghaKouchak, A., Prestemon, J., Olszewski, J., & Sadegh, M. (2023). Physical, social, and biological attributes for improved understanding and prediction of wildfires: FPA FOD-attributes dataset. *Earth System Science Data Discussions*,*16*(6), 1-29. https://doi.org/10.5194/essd-16-3045-2024

Preisler, H. K., & Benoit, J. W. (2004). *A State Space Model for predicting Wildland Fire Risk*. USDA Forest Service, Pacific Southwest Research Station, Albany and Riverside, California. https://www.fs.usda.gov/psw/publications/preisler/psw_2004_preisler001_asa.pdf

Radke, D., Hessler, A., & Ellsworth, D. (2019). FireCast: Leveraging Deep Learning to Predict Wildfire Spread. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 4575-4581. https://doi.org/10.24963/ijcai.2019/636

Rodrigues, M., & de la Riva, J. (2014). An insight into machine-learning algorithms to model human-caused wildfire occurrence. *Environmental Modelling & Software*, *57*, 192-201. https://www.sciencedirect.com/science/article/pii/S1364815214000814