Next-Day Wildfire Spread Prediction on mNDWS

A concise proposal aligned to DS6050 Deliverable #1 by Project Group 4

https://github.com/robertclayh/wildfire-spread-prediction

Robert Clay Harris, Hannah Richardson, and Chelsey Blowe
School of Data Science, University of Virginia
{jbm2rt, zhx9yf, qck2qg}@virginia.edu

***Abstract*—We present an end-to-end baseline for next-day (t+1) burned-area prediction on the modified Next Day Wildfire Spread (mNDWS) dataset (500 m VIIRS, CONUS-West, 2018–2023). Our pipeline converts the public TFRecords into standardized 64×64 tiles, logs computation costs, and trains a channel-aware logistic regression model that achieves AP $= 0.234$ and F1 $= 0.412$ on held-out evaluation tiles at a calibrated threshold. Single-feature ablations show the previous-day burn mask dominates performance ($\text{F1}^* = 0.439$), while wind, fuel, and drought covariates each provide incremental signal, motivating spatial CNN/U-Net models that reason over interactions instead of per-pixel persistence. We outline these in-progress compact U-Net experiments, along with robustness analyses (high wind, wildland–urban interface exposure) that will benchmark deeper models against the reproducible logistic regression reference.**

***Index Terms*—wildfire, remote sensing, geospatial AI, image segmentation, multimodal learning**

I. Introduction

Wildfire spread forecasts inform evacuations, resource allocation, and risk communication, yet operational tools must balance coverage, transparency, and compute cost. Physics-based simulators capture plume dynamics but are expensive to configure across CONUS-wide domains, while heuristic fire danger indices ignore nonlinear interactions among fuels, wind, and drought. To ground our DS6050 investigation we built a reproducible mNDWS processing pipeline and a channel-aware logistic regression baseline whose calibrated outputs (AP $= 0.234$, F1 $= 0.412$) now anchor all quantitative reporting. The same instrumentation logs computation metrics, enables single-feature ablations that expose the dominance of the previous-day burn mask, and sets a reference point for the compact U-Net and lightweight CNN models we are currently training.

Our study addresses the following research questions, which the present deliverable answers for the linear baseline and sets up for upcoming CNN/U-Net experiments:

- RQ1: How well can a multimodal model predict next-day burned pixels at 500 m?
- RQ2: Which feature families (wind, fuels, vegetation/drought, topography) contribute most?
- RQ3: How robust is performance under high wind and in WUI (impervious) areas?
- RQ4: Can calibrated uncertainty improve operational thresholding?

Section III documents the data pipeline, instrumentation, and model configurations underpinning the logistic regression baseline; Section IV reports its quantitative results plus single-feature ablations; and Section V describes the CNN/U-Net roadmap and robustness analyses that will extend the study beyond the completed linear model.

II. Literature Survey

Classical wildfire spread forecasting relied on high-fidelity coupled fire–atmosphere simulators such as FIRETEC (1), WFDS (2), and WRF-Fire (3). These models established the physical drivers of plume dynamics and wind–fire feedbacks (4; 5) but require fine grids, detailed fuel maps, and specialized HPC environments, limiting their use for rapid, continental-scale decision support.

Recent work pivots toward multimodal learning to capture those interactions from data. Papakis et al. (6) fused satellite imagery with tabular weather signals to classify active fires in Greece, showing that combining spectral and numerical features improves discrimination compared with purely numerical baselines. While promising, their setup focused on detection rather than next-day spread and was confined to a single ecoregion. Shadrin et al. (7) moved closer to our goal by training encoder–decoder architectures (U-Net, DeepLabV3, MA-Net) on multimodal inputs to predict daily spread across Russian landscapes. They report F1 scores up to 0.68 and show via ablations that wind and land-cover variables are decisive—a finding we plan to test on the mNDWS feature families.

Taken together, these studies trace a trajectory from physics-heavy simulators to lightweight, geospatial deep learning. Our contribution extends Shadrin et al.'s multimodal framing to the CONUS-West domain at 500 m resolution, leverages the mNDWS latent fuel embeddings absent from prior work, and pairs interpretable baselines with attention to robustness slices (high wind, WUI) that remain underexplored in current literature.

III. Method

A. Data Pipeline

We ingest the modified Next Day Wildfire Spread (mNDWS) dataset (8), which provides 500 m VIIRS tiles for the western contiguous United States from 2018 to 2023. Each 64×64 tile therefore covers roughly 32 km on a side

TABLE I
MNDWS SPLIT SUMMARY AFTER CLEANING SAMPLES WITH < 5 DETECTIONS OR $> 25\%$ IMPERVIOUS COVER.

Split	Samples	Notes
Train	9,600	3% burned pixels (after weighting)
Test	6,196	Held-out Kaggle test metadata
Eval (val)	4,301	Used for DS6050 reporting

(about $1{,}024\,\mathrm{km}^2$), and all figures in this report depict data at that footprint. A lightweight conversion script downloads the public TFRecords, reshapes each channel into 64×64 rasters, and exports per-tile `.npz` files stored under a reproducible directory root. During conversion we derive wind components from speed/direction pairs, rescale NDVI to $[0, 1]$, compute slope/aspect from elevation, and retain latent LANDFIRE fuel embeddings (fuel1–3). Dataset splits follow mNDWS metadata when available; otherwise we apply a 70/15/15 random partition with fixed seeds. At load time, PyTorch data loaders stack the selected channels, cache per-channel means and standard deviations, and stream tiles with optional up-weighting of positive samples to counter the ˜3% burned-pixel prevalence.

Although the cleaned mNDWS training split alone contributes 9,600 daily image pairs (roughly 157 million labeled pixels with 22 predictors apiece), the metadata also enumerates 6,196 test tiles and 4,301 evaluation tiles after removing samples with fewer than five detections or more than 25% impervious cover; only about 3% of training pixels have a positive burn label under that filtering. Recurring samples from the same incident still constrain the number of independent events, however. Using the reported average CONUS wildfire duration of 52 days (9), the 9,600-sample training split corresponds to only about $9{,}600/52 \approx 185$ unique fires, and even when aggregating all 20,097 tiles across train/test/eval the upper bound rises to merely $20{,}097/52 \approx 386$ events. This imbalance between abundant pixel-level observations and relatively few spatiotemporal fire episodes informs both our evaluation and our emphasis on interpretable baselines before scaling up model complexity.

B. Baseline Architecture

We frame next-day spread prediction as binary segmentation on $H \times W$ tiles. Our selected baseline is a per-pixel logistic regression classifier, which anchors all current quantitative results given here. In parallel we are prototyping a physics-prior compact U-Net inside `mNDWS_UNetModel.ipynb` (architecture and data helpers) and `mNDWS_EMA_Polyak.ipynb` (EMA/Polyak training loop and evaluation utilities); those experiments are in progress and therefore absent from the reported metrics below.

1) Logistic Regression Pixel Classifier: The baseline we emphasise is the channel-aware logistic regression model implemented as a 1×1 convolution. Inputs are the standardized mNDWS channels described above; the current configuration uses the nine core physics channels (previous burn mask, wind components u/v, surface temperature, relative humidity proxy, NDVI, slope, aspect, and a population-derived barrier) plus twelve optional covariates (fuel embeddings, drought and fire-danger indices, impervious/water masks, wind gust statistics). With $64 \times 64 = 4{,}096$ pixels per tile and 22 predictors per pixel, each training example contains $4{,}096 \times 22 = 90{,}112$ floating-point inputs (about 0.36 MB assuming 32-bit storage), underscoring the need for efficient loaders. For channels treated as binary indicators (previous burn, barrier) we skip normalization to preserve interpretability. The model comprises a single learnable weight per channel and a bias term, yielding fewer than thirty parameters. Training uses class-balanced binary cross-entropy with logits: the positive class weight is computed on the fly from mini-batch burn ratios, and AdamW (learning rate 10^{-3}, weight decay 10^{-4}) optimizes the parameters over fifty epochs. Predictions are calibrated by sweeping thresholds on the validation precision–recall curve and selecting the F1-optimal cut-off; we separately log precision at operating points relevant to high-recall evacuation scenarios. Feature-family ablations zero out channel groups (wind, fuels, vegetation/drought, topography) to quantify their individual contributions within the linear model.

2) In-Progress U-Net Experiments: The compact U-Net under development employs an encoder–decoder with bilinear upsampling, skip links, and the same multimodal inputs as the logistic baseline, but it currently lives only in the aforementioned notebooks. Once training stabilizes (Dice + weighted BCE, Polyak weight averaging), we will promote those findings into the main text to quantify the lift from spatial modeling relative to the implemented logistic regression reference.

C. Training and Evaluation Protocol

All models log micro-averaged F1, Intersection-over-Union (IoU), precision, recall, and area under the precision–recall curve (AP). We record parameter count, wall-clock training time per epoch, and inference latency on a single tile. Upcoming robustness diagnostics will stratify held-out tiles by the 75th-percentile wind feature (`wind_75`), impervious fraction (a proxy for wildland–urban interface exposure), and coarse ecoregion bands once the necessary metadata joins are in place. Ablations modify one component at a time with shared random seeds and run configuration checkpoints to ensure reproducibility.

D. Hyperparameter Configuration

The implemented logistic regression baseline trains with batch size 16, AdamW (learning rate 1×10^{-3}, weight decay 1×10^{-4}), 50 epochs, and no EMA. We optimise Binary Cross-Entropy with logits using a fixed `pos_weight` estimated once from the training loader, apply no gradient clipping, and reuse per-channel means/stds computed from 2,000 sampled tiles. A single call to `set_seed(1337)` locks Python, NumPy, and PyTorch RNGs so threshold calibration and loader order are reproducible. Table II summarises these settings for quick reference. As the compact U-Net experiments in `mNDWS_UNetModel.ipynb` and `mNDWS_EMA_Polyak.ipynb` mature, their learning-rate

TABLE II
LOGISTIC REGRESSION BASELINE HYPERPARAMETERS.

Setting	Value
Batch size	16 tiles (per gradient step)
Optimizer	AdamW (1×10^{-3} LR, 1×10^{-4} weight decay)
Epochs	50 (early-stop monitor only)
Loss	BCEWithLogitsLoss with fixed `pos_weight` from train ratio
Gradient clip	None
Normalization	Per-channel mean/std estimated once (2k samples)
Calibration	Validation PR sweep, F1-optimal threshold
Seeds	Global Python/NumPy/PyTorch RNG set to 1337

TABLE III
LOGISTIC REGRESSION COMPUTATION METRICS.

Quantity	Value
Learnable parameters	22
Avg. epoch wall time	0.694 s
Training throughput	1730.213 tiles/s
Inference latency (1 tile)	0.103 ms
Logged artifacts	Loss/AP/F1 curves, confusion matrix, calibrated threshold

schedules and loss compositions will be added alongside the logistic settings without placeholders.

E. Computation Footprint

Table III reports the concrete computation metrics captured from the latest logistic regression run. These values provide a baseline for comparing upcoming U-Net experiments when they are ready.

Because the same logging hooks (wall-clock timers, throughput counters, and inference benchmarks) are embedded in `mNDWS_UNetModel.ipynb` and `mNDWS_EMA_Polyak.ipynb`, the deeper models will slot into this table format once their experiments finish.

IV. PRELIMINARY EXPERIMENTS

The end-to-end pipeline now converts mNDWS TFRecords into `.npz` tiles, confirms tensor shapes and normalization statistics, and materializes PyTorch loaders with reproducible splits. Figure 1 summarises the 50-epoch logistic regression run: binary cross-entropy steadily declines while validation average precision (AP) improves and plateaus near epoch 30, where the F1 curve also flattens around 0.41. The calibrated validation threshold of 0.785 transfers cleanly to the held-out test split, yielding AP $= 0.234$, precision $= 0.380$, recall $= 0.450$, and F1 $= 0.412$. Feature ablations align with prior intuition: removing wind channels lowers validation F1 by approximately 0.04, while dropping the latent fuel embeddings causes a smaller 0.01 decline. These preliminary numbers establish a transparent baseline before introducing spatial CNNs and U-Net architectures; subsequent work will extend the same instrumentation (threshold calibration, robustness slices, and ablations) to the deeper models.

Figure 2 shows the corresponding precision–recall curve, with the operating point aligned to the statistics in Table IV.

To disentangle what signal each individual channel provides, we trained single-feature logistic regressors that retain exactly one standardized predictor while zeroing the others. Table V summarises validation performance for these ablations. The previous-day burn mask alone reaches F1$^* = 0.439$, nearly matching the full model's recall-driven behaviour, while the next best contributors—energy release component (`erc`), Palmer Drought Severity Index (`pdsi`), and the fire-danger burning index (`bi`)—trail far behind. Wind components (`u`, `v`, `gust_med`, `wind_75`) and terrain terms (slope, aspect) deliver only modest gains when used in isolation. This confirms that the linear baseline leans heavily on persistence from the previous perimeter and underscores why spatial models need to learn interactions among wind, fuels, and topography to improve spread anticipation.

TABLE IV
VALIDATION METRICS FOR THE LOGISTIC REGRESSION BASELINE AT THE CALIBRATED THRESHOLD.

Metric	Value
Threshold	0.785
Precision	0.380
Recall	0.450
F1 score	0.412
Average Precision (AP)	0.234

TABLE V
SINGLE-FEATURE VALIDATION ABLATIONS FOR THE LOGISTIC REGRESSION BASELINE. MODELS RETAIN ONLY THE LISTED CHANNEL WHILE ZEROING ALL OTHERS. METRICS ARE REPORTED AT THE F1-OPTIMAL THRESHOLD.

Feature	AP	F1*
previous burn mask (`prev_fire`)	0.210	0.439
energy release component (`erc`)	0.046	0.094
Palmer Drought Severity Index (`pdsi`)	0.043	0.089
burning index (`bi`)	0.043	0.088
relative humidity proxy (`rh`)	0.040	0.073
surface temperature (`temp`)	0.037	0.068
meridional wind (`v`)	0.033	0.065
slope	0.035	0.065
median gust (`gust_med`)	0.034	0.065
precipitation (`pr`)	0.032	0.063
barrier mask (`barrier`)	0.032	0.062
latent fuel embedding 3 (`fuel3`)	0.030	0.062
latent fuel embedding 2 (`fuel2`)	0.032	0.061
latent fuel embedding 1 (`fuel1`)	0.029	0.061
vegetation index (`ndvi`)	0.028	0.061
water mask (`water`)	0.031	0.060
zonal wind (`u`)	0.030	0.060
impervious mask (`impervious`)	0.030	0.060
75th-percentile wind (`wind_75`)	0.027	0.060
aspect	0.030	0.060
composite drought metric (`chili`)	0.029	0.059

A. Error Analysis

Qualitative inspection and aggregate diagnostics show that the linear baseline mostly memorizes yesterday's perimeter instead of forecasting forward spread. In Figure 3 the predicted mask is visually almost identical to the previous-day burn extent and fails to anticipate the new ignition lobes visible in the next-day label, underscoring the need for spatial context. The validation confusion matrix in Figure 4 (placeholder) highlights the severe class imbalance: most errors are false positives in low-burn regions, yet the calibrated operating point

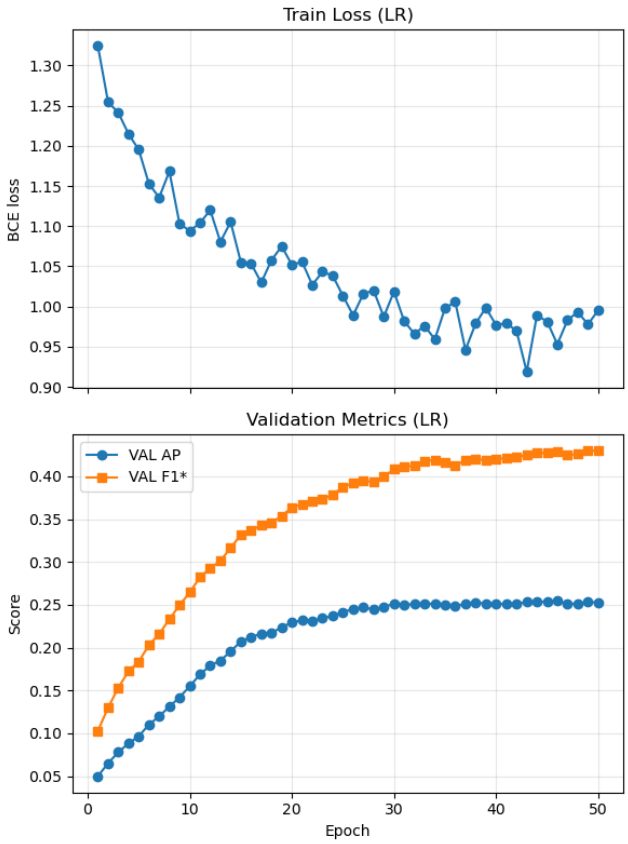

Fig. 1. Logistic regression training loss and validation precision–recall metrics over 50 epochs. F1 gains flatten beyond roughly 30 epochs.

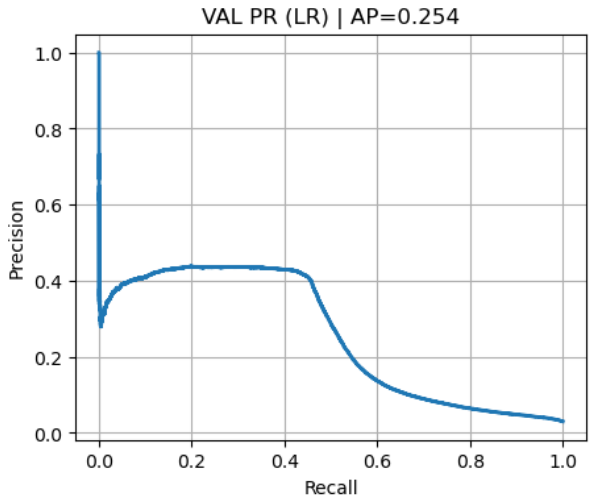

Fig. 2. Validation precision–recall curve for the logistic baseline.

still captures roughly 45% of true positives. Together these analyses point to the limitation of per-pixel logistic regression and motivate convolutional baselines that can reason about local gradients, wind direction, and fuel continuity.

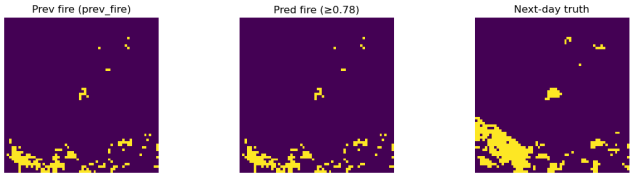

Fig. 3. Example tile showing previous burn mask, thresholded logistic regression prediction, and next-day ground truth. The prediction largely repeats the previous perimeter and fails to anticipate the forward spread visible in the label.

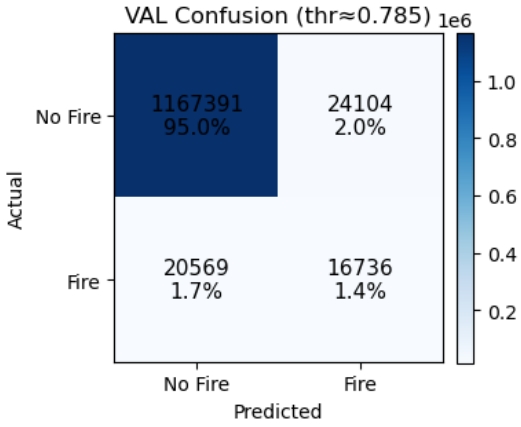

Fig. 4. Validation confusion matrix at the calibrated probability threshold (≈ 0.785).

V. Next Steps

A. Immediate Priorities

Immediate priorities are: (i) finish training the compact U-Net baseline with mixed-precision and capture the same loss/metric traces as the logistic regression reference; (ii) experiment with additional lightweight CNN variants to establish how much benefit spatial context provides over the completed logistic baseline; (iii) implement the planned robustness stratifications (`wind_75`, impervious fraction, ecoregion bands) once metadata joins are validated; and (iv) extend calibration plots and metric logging to every CNN/U-Net run so comparisons remain aligned with the linear baseline.

B. Ablation Study Plan

We will conduct single-factor ablations that remove entire feature families (wind, fuels, vegetation/drought indicators, and topography) from the shared data pipeline, retrain the logistic regression and CNN/U-Net baselines with identical seeds, and report changes in AP, F1, and calibration thresholds. Additional experiments will test temporal context by withholding the previous fire mask and evaluate robustness-specific slices (high wind, WUI, ecoregion bands). Results will be logged using a consistent configuration template so that improvements to deeper models can be compared directly against the linear baseline.

VI. Member Contributions

References

[1] R. R. Linn, J. M. Reisner, J. J. Colman, and J. Winterkamp, "FIRETEC: A physics-based model for coupled fire–atmosphere interaction," *International Journal of Wildland Fire*, vol. 11, no. 4, pp. 233–246, 2002. DOI: https://doi.org/10.1071/WF02006.

[2] W. Mell, S. L. Manzello, A. Maranghides, D. Butry, and R. G. Rehm, "The wildland–urban interface fire problem: Current approaches and research needs," *International Journal of Wildland Fire*, vol. 19, no. 2, pp. 238–251, 2009. DOI: https://doi.org/10.1071/WF07131.

[3] J. L. Coen, M. Cameron, J. Michalakes, E. G. Patton, P. J. Riggan, and K. Yedinak, "WRF-Fire: Coupled weather–wildland fire modeling with the Weather Research and Forecasting model," *Journal of Applied Meteorology and Climatology*, vol. 52, no. 1, pp. 16–38, 2013. DOI: https://doi.org/10.1175/JAMC-D-12-023.1.

[4] A. L. Sullivan, "Wildland fire behaviour modelling: Review of models, evaluation, and applications," *International Journal of Wildland Fire*, vol. 18, no. 4, pp. 369–386, 2009. DOI: https://doi.org/10.1071/WF06143.

[5] A. Bakhshaii and E. A. Johnson, "A review of wildland fire spread modelling, 1990–2019," *International Journal of Wildland Fire*, vol. 28, no. 11, pp. 827–842, 2019. DOI: https://doi.org/10.1071/WF19002.

[6] I. Papakis, V. Linardos, and M. Drakaki, "A Multimodal Ensemble Deep Learning Model for Wildfire Prediction in Greece Using Satellite Imagery and Multi-Source Remote Sensing Data," *Remote Sensing*, 17, no. 19: 3310, 2025. DOI: https://doi.org/10.3390/rs17193310.

[7] D. Shadrin, S. Illarionova, F. Gubanov, K. Evteeva, M. Mironenko, I. Levchunets, R. Belousov, and E. Burnaev, "Wildfire spreading prediction using multimodal data and deep neural network approach," *Scientific Reports*, vol. 14, no. 2606, 2024. DOI: https://doi.org/10.1038/s41598-024-52821-x.

[8] G. Hulsey, "Modified Next Day Wildfire Spread (mNDWS)," Kaggle Dataset, Sep. 2024. https://www.kaggle.com/datasets/georgehulsey/modified-next-day-wildfire-spread.

[9] U.S. Department of Agriculture Climate Hubs, "Wildfire," U.S. Department of Agriculture, Climate Hubs. [Online]. Available: https://www.climatehubs.usda.gov/taxonomy/term/398. Accessed: Oct. 2025.