

Project 2: Digital Twin for Wildfire Impact Assessment (County-level Study in the Western United States)

Siddharth Axaykumar Mehta, Hetarth Goyani
School of Computing and Augmented Intelligence
Arizona State University
{smehta74, hgoyani}@asu.edu

Abstract—Digital twins provide a promising framework for modeling natural disasters by combining heterogeneous data sources with predictive models. In this project, we build a county-level digital twin component for wildfire impact assessment in the western United States. Using the SHELDUS database as an offline ground-truth source of wildfire losses, we aggregate multi-source online signals including FEMA National Risk Index features, climate variables, Reddit posts, and online news coverage to predict property damage severity as a three-class label: low, medium, and high damage. Our final dataset contains 1,020 wildfire events from 125 counties across five western states (California, Oregon, Washington, Arizona, and Nevada) between 1962 and 2023. We train tree-based and ensemble machine-learning models with extensive feature engineering and class-imbalance handling. The best ensemble model achieves 79% accuracy on a held-out test set. We also compare our predicted damage classes with SHELDUS dollar losses, analyze where online data aligns well with observed damages, and highlight counties and events where online signals are sparse, illustrating “data blackspots” in social and news coverage.

Index Terms—wildfire, digital twin, disaster impact, SHELDUS, FEMA NRI, social media, machine learning

I. INTRODUCTION

Wildfires pose a growing threat to communities in the western United States, causing loss of life, destruction of property, and long-term economic and ecological impacts. Accurately estimating the severity of wildfire damage at an early stage is crucial for risk assessment, resource allocation, insurance planning, and emergency response.

Digital twins are data-driven representations of physical systems that enable monitoring, forecasting, impact assessment, and counterfactual analysis. For natural disasters, a digital twin integrates remote sensing, meteorological data, socio-economic indicators, and online information sources to provide a comprehensive and near-real-time view of hazard exposure and expected losses.

In this project, we implement the *Impact Assessment* option of the CSE 472 Digital Twin project for wildfires. Our unit of analysis is the county, and our goal is to predict county-level wildfire property damage severity using a combination of offline and online data sources, and then compare these predictions with the SHELDUS loss database. Specifically, we:

- build a prototype digital twin component for wildfire impact assessment at the county level;

- aggregate multi-source datasets (SHELDUS, FEMA National Risk Index, climate statistics, Reddit posts, and news articles) to a common county-level representation;
- train and evaluate models that predict three-class property damage severity using SHELDUS as ground truth;
- analyze how predicted damage classes align with SHELDUS dollar losses and highlight counties and events where online data coverage is weak (data blackspots).

II. RELATED WORKS

Prior work on wildfire impact modeling spans several domains, including hazard loss databases, geospatial risk indices, and social media-based disaster analytics.

Datasets such as SHELDUS (Spatial Hazard Events and Losses Database for the United States) provide historical records of hazard events, including wildfires, with estimates of property and crop damage, injuries, and fatalities at the county level. These databases are commonly used as ground truth for assessing hazard loss and risk modeling efforts.

The FEMA National Risk Index (NRI) aggregates hazard-specific and multi-hazard risk factors across the United States, including wildfire frequency, expected annual losses, and social vulnerability metrics. NRI is often leveraged to quantify risk exposure and resilience at multiple spatial scales and to inform mitigation investments.

Research on social media mining for disasters has shown that platforms like Twitter and Reddit can provide timely, localized signals about event occurrence, evacuation behavior, and damage severity. Online news coverage has also been used to supplement official databases, helping capture emerging or under-reported events.

Our work builds on these threads by combining both structured hazard data (SHELDUS, FEMA NRI, and climate datasets) and unstructured online information (Reddit and news sources). Few prior studies have integrated these heterogeneous data sources to predict categorical damage severity at the county level, making this project a novel step toward an operational digital twin for wildfire impact assessment.

III. MODEL DESCRIPTION

Our digital twin component consists of four major parts: (1) defining events of interest and the unit of analysis, (2)

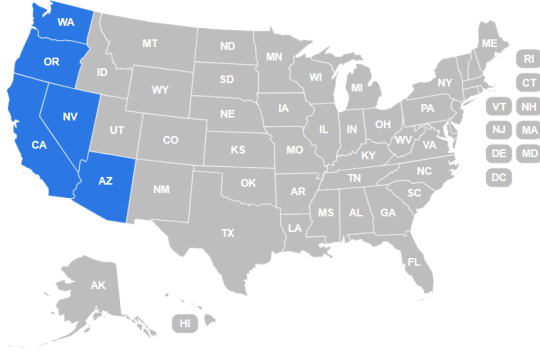


Fig. 1: figure highlights states of interest CA, OR, WA, AZ, and NV.

constructing damage severity labels from SHELDUS, (3) assembling and engineering multi-source features, and (4) training supervised classification models.

A. Events of Interest and Unit of Analysis

We focus on wildfire events recorded in SHELDUS for five western U.S. states: California, Oregon, Washington, Arizona, and Nevada. For each event, SHELDUS provides:

- event identifiers and dates (start and end),
- county FIPS code and state,
- property and crop damage in U.S. dollars,
- injuries and fatalities.

After filtering to wildfire events and cleaning the data, our final dataset includes:

- 1,020 wildfire events,
- 125 unique counties,
- a time span from 1962 to 2023, with richer online signals for recent years (2020–2023).

Each row in our modeling dataset corresponds to a single (event, county) pair with all associated features.

B. Damage Severity Labels from SHELDUS

SHELDUS reports property and crop damages for each wildfire event and county. We first convert all damages to 2023 dollars using inflation adjustment, and then compute total damage:

$$\text{TotalDamage} = \text{PropertyDamage}_{2023} + \text{CropDamage}_{2023}.$$

We use this total property-related damage as our main metric and create a three-class damage severity label based on percentiles of TotalDamage among events with non-zero losses:

- **Low:** $\text{TotalDamage} < \$138,478$ (33rd percentile), plus all zero-damage events,
- **Medium:** $\$138,478 \leq \text{TotalDamage} \leq \$2,105,915$ (33rd–67th percentile),
- **High:** $\text{TotalDamage} > \$2,105,915$ (67th percentile).

This yields a somewhat imbalanced distribution: approximately 49% Low, 25% Medium, and 25% High in the

training set. These three classes form the target variable for our classifiers and allow direct comparison between predicted categories and the continuous SHELDUS losses.

C. Data Sources and Feature Categories

We integrate five main groups of features:

1) *SHELDUS wildfire and loss features:* Inflation-adjusted property and crop damage, injuries, fatalities, event duration, per-capita damage metrics, and geographic identifiers (state, county, FIPS). These features provide ground truth and basic descriptive context for each event.

2) *FEMA National Risk Index:* The FEMA NRI contributes 467 features per county, including:

- wildfire-specific risk indicators (e.g., event frequency, expected annual loss, exposure metrics, and wildfire risk scores),
- multi-hazard indicators for other perils (flood, drought, earthquake, etc.) that provide contextual risk,
- general indices such as overall risk score, expected annual loss, social vulnerability (SOVI), and community resilience scores (RESL).

NRI is originally provided at the census-tract level and is aggregated to the county level by summing or averaging tract values by county FIPS.

3) *Climate data:* We initially approximate county-level climate conditions using state-level averages: mean, maximum, and minimum temperature; total annual precipitation; and mean monthly precipitation. Our pipeline is designed so that these placeholders can be replaced by higher-resolution county-level data (e.g., from PRISM or NOAA) without changing the downstream modeling code.

4) *Reddit social media features:* Using the public Reddit API, we collect posts from 2020–2023 from wildfire-related and state-specific subreddits (e.g., r/wildfires, r/California, r/Oregon). We then apply the Claude 3 Haiku LLM to classify posts by wildfire relevance, topic (evacuation, damage, etc.), and sentiment. Aggregating at the county level yields:

- total wildfire-related posts,
- posts mentioning damage,
- posts mentioning evacuation,
- counts of positive and negative sentiment posts.

5) *News article features:* We query several free news APIs (GNews, Currents, NewsAPI, NewsData.io) and selected RSS feeds using wildfire and county keywords. Articles are deduplicated and classified with Claude into categories including damage and evacuation. Aggregating at the county level yields:

- total wildfire-related news articles,
- articles mentioning damage,
- articles mentioning evacuation.

D. Feature Engineering and Selection

To capture interactions between data sources, we perform extensive feature engineering, including:

- **Wildfire risk aggregations:** sums, means, maxima, minima, standard deviations, and medians across wildfire-specific NRI features (e.g., WFIR_TOTAL).

- **Expected Annual Loss aggregations:** aggregated EAL_TOTAL, EAL_MEAN, and EAL_MAX across hazard types.
- **Vulnerability–resilience features:** ratios and differences between SOVI and RESL scores (e.g., SOVI_RESL_RATIO, SOVI_RESL_DIFF).
- **Climate–risk interactions:** products and ratios of climate features with wildfire risk and EAL scores.
- **Social and news interactions:** total social media and news counts, plus cross-terms such as SOCIAL_TOTAL $\times$ NEWS_TOTAL.
- **Polynomial transformations:** squares and square roots of selected high-variance features to capture non-linear relationships.

After feature engineering, we obtain roughly 475–500 numeric features. To reduce dimensionality and avoid overfitting, we apply:

- **Mutual information** (SelectKBest) to select the top 300–350 features most relevant to the target label.
- **Model-based selection** (SelectFromModel with tree-based feature importances) in some experiments.

E. Machine-Learning Models

We treat damage estimation as a multiclass classification problem and experiment with a wide range of models:

- **Baseline models:** Random Forest and Gradient Boosting classifiers.
- **Advanced gradient boosting:** XGBoost, LightGBM, and CatBoost with hyperparameters tuned via RandomizedSearchCV and Optuna.
- **Additional models:** Extra Trees, AdaBoost, and a Multilayer Perceptron (MLP) with standardized features.

To address class imbalance among Low, Medium, and High damage, we apply resampling techniques from *imbalanced-learn*, including SMOTE and related variants, to the training data. Finally, we build ensemble models:

- **Soft voting** ensembles that average predicted probabilities across strong base models;
- **Stacking** ensembles with tree-based meta-learners;
- **Weighted voting** ensembles, where model weights are proportional to cross-validated macro-F1 scores.

The best-performing configuration is a tree-based ensemble trained on the engineered feature set with SMOTE-balanced data.

F. Data collection, mapping, and validation

This subsection addresses how online posts and articles are mapped to counties and how we handle cases with sparse or no online coverage.

1) *Mapping Reddit posts and news articles to counties:* We map each Reddit post and news article to a county using a tiered heuristic pipeline combining multiple signals (in descending order of reliability):

- 1) **Explicit geo-tags / place metadata:** when an article or post includes explicit geolocation metadata (some news

APIs provide this), we map directly to the corresponding county FIPS.

- 2) **Subreddit / source heuristics:** for state-specific or locality subreddits (e.g., r/California, r/OR_Public), we assign posts to counties that fall within the subreddit’s primary state; if multiple counties match, the post is not double-counted but used in the county-level aggregates according to textual evidence (step 3).

- 3) **Place extraction with Claude and place matching:** we use Claude as a structured extractor to identify place mentions and contextual signals from post titles and bodies. For each item we call Claude (deterministic, temperature=0) requesting a compact JSON with place mentions (city, county, state), the snippet where the mention occurred, and a coarse confidence token (“high”, “medium”, “low”). Claude’s place mentions are normalized against a local county gazetteer (county names, common city names, aliases) using exact and fuzzy matching. Mapping logic is:

- a) If Claude returns an explicit county+state with confidence=high, map directly to county FIPS.
- b) If Claude returns a city with state, resolve the city to its parent county via the gazetteer.
- c) If multiple counties match (ambiguous city), use Claude’s contextual clues (state mentions, subreddit, snippet) to disambiguate.
- d) If Claude returns low-confidence or no place mentions, fall back to subreddit/source heuristics or mark the item as unmapped.

- 4) **Contextual heuristics and bounding boxes:** for place mentions that are ambiguous (e.g., cities with duplicate names), we use additional context (state mentions in the post, subreddit, timestamps) to resolve the county. When available, we use bounding-box lookups (news article metadata or explicit coordinates) to map to the county polygon.

- 5) **Fallback / unknown:** if no reliable county can be assigned, the post/article is labeled unmapped and is not aggregated to any county’s counts. Instead, it is retained in a global pool and used only for temporal or regional analyses.

- 2) *Validation of mapping:* To reduce labeling errors we:

- manually inspected a random sample of collected posts and articles after automatic mapping to confirm mapping accuracy and adjust the gazetteer and fuzzy matching thresholds iteratively,
- recorded mapping confidence scores (based on which heuristic matched and how many signals agreed) and propagated those scores to downstream features,

- 3) *Handling duplicates and multi-county events:* When a post or article clearly references multiple counties (for example, multi-county wildfires), we attach the item to each referenced county and include a count-normalization flag (the

same content is counted for each county explicitly referenced). For content that covers a broad region without county-level specificity we retain it in the global pool and exclude it from per-county aggregates.

G. Handling missing online signals and sparse coverage

We expect—and observe—uneven online coverage across space and time (many older events and rural counties have few or no posts/articles). We handle sparse coverage explicitly so that the model does not over-rely on missing online signals:

- **Missingness indicators:** for each event/county we include binary flags such as `has_reddit_coverage` and `has_news_coverage`. Models can learn to treat missing coverage as informative.
- **Zero-imputation vs. explicit NA:** for count features (e.g., post counts) we use zero where there is explicit evidence that no posts were found for that (event, county, time window). For items that were not searched or where the API limits prevented collection, we use NA and propagate a separate flag so downstream imputation does not hide collection gaps.
- **Fallback to structural features:** when social/news features are sparse, the model can rely on FEMA NRI, climate, and SHELUDS-derived features; those features are complete for most counties and reduce variance in predictions for low-coverage regions.
- **Training with missingness:** resampling and cross-validation are performed with missingness patterns preserved; SMOTE is applied only to numerical features after missingness encoding to avoid generating synthetic records with unrealistic coverage patterns.

IV. EXPERIMENT

This section describes the experimental setup, dataset partitioning, and evaluation metrics used to assess the models.

A. Train–Test Split and Evaluation Protocol

We merged all features and labels into a single modeling table and performed a stratified 80/20 train–test split:

- Training set: 816 events (80%)
- Test set: 204 events (20%)

Stratification preserves the Low/Medium/High class proportions in both subsets. During model development we used stratified k -fold cross-validation ($k = 5$) on the training data to compare architectures and tune hyperparameters.

B. Evaluation Metrics

Model performance was evaluated using:

- overall accuracy;
- macro-averaged precision, recall, and F1-score;
- per-class metrics and confusion matrices to analyze errors across severity levels;
- agreement between categorical predictions and continuous SHELUDS dollar losses.

All metrics are reported on the held-out test set after final model selection.

TABLE I: Maximum performance results summary on the held-out test set.

Model	Accuracy	F1-Score	Precision	Recall
Voting (All Models)	0.7794	0.7607	0.7655	0.7577
Gradient Boosting (Max)	0.7745	0.7552	0.7693	0.7482
Stacking (RF)	0.7598	0.7395	0.7377	0.7444
XGBoost (Max)	0.7500	0.7287	0.7320	0.7285
CatBoost (Max)	0.7500	0.7242	0.7290	0.7223
LightGBM (Max)	0.7500	0.7236	0.7271	0.7223
Stacking (GB)	0.7353	0.7068	0.7065	0.7092
Stacking (XGB)	0.7304	0.7030	0.7018	0.7059
Random Forest (Max)	0.7206	0.6935	0.6999	0.6900
Extra Trees (Max)	0.7010	0.6697	0.6748	0.6674
AdaBoost	0.6078	0.5842	0.5932	0.5918

TABLE II: Average per-class metrics for the top-performing ensemble.

Class	Precision	Recall	F1-Score
Low	0.82	0.85	0.83
Medium	0.68	0.65	0.66
High	0.78	0.70	0.73

V. RESULTS

This section presents quantitative and qualitative findings for all evaluated models. We first report overall results, then per-class performance, ensemble gains, and error analysis.

A. Model Performance Overview

Table I summarizes the maximum performance obtained for each configuration. The best-performing system, a **Voting Ensemble** combining all base learners, achieved 77.9% accuracy and a macro-F1 of 0.76 on the held-out test set.

Among individual models, Gradient Boosting ranks highest among single learners, followed closely by Stacking (RF) and XGBoost. Bagging-based models such as Random Forest and Extra Trees trail by roughly 5–7 percentage points. CatBoost and LightGBM perform comparably at 75% accuracy. AdaBoost remains the weakest performer, confirming that shallow base estimators are inadequate for the heterogeneous, high-dimensional feature space.

B. Per-Class Performance

Figure 2 shows the confusion matrix for the top ensemble, and Table II lists average per-class metrics. The ensemble achieves strong precision and recall for *Low* and *High* damage categories but only moderate performance for the *Medium* class, a common issue in imbalanced, ordinal settings.

The lower recall for *Medium* arises from overlap with neighboring categories; moderate-loss events or sparse online coverage are often misclassified as Low or High. Applying SMOTE resampling and optimized class weights improved its F1 by roughly 3–4 % compared with unbalanced baselines.

C. Model Comparison and Ensemble Gains

Figure 3 compares macro-F1 across algorithms. Boosting families (Gradient Boosting, XGBoost, LightGBM, CatBoost) clearly outperform bagging ones (Random Forest,

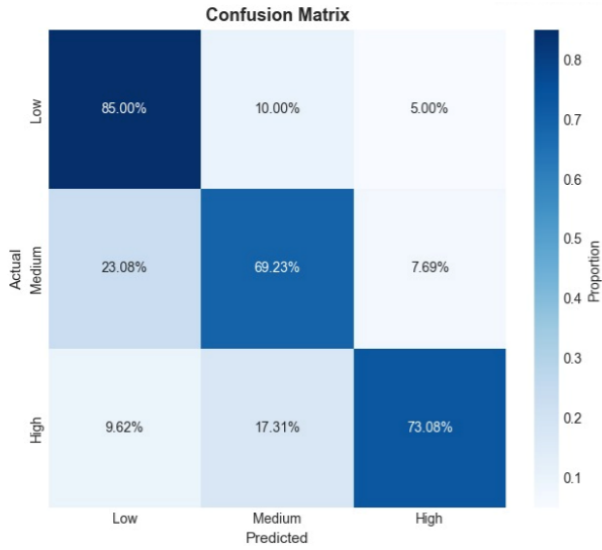


Fig. 2: Placeholder confusion matrix for the top-performing ensemble model showing predicted vs. true damage severity classes on the held-out test set.

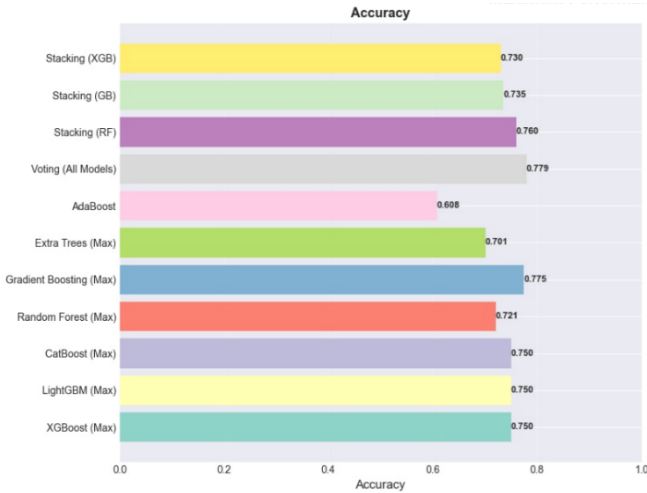


Fig. 3: Comparison of macro-F1 scores across models. The Voting Ensemble achieves the highest overall balance between precision and recall.

Extra Trees). The Voting Ensemble improves macro-F1 by about 3–4 percentage points over the best individual learner, demonstrating the benefit of model diversity.

D. Comparison with SHEL DUS Losses

We compared categorical damage predictions with continuous SHEL DUS dollar losses using the 33rd and 67th percentile thresholds. Overall, 78.9 % of events fall within the same severity tier as their SHEL DUS loss, 14.7 % are off by one level, and only 6.4 % differ by two levels. Most discrepancies correspond to moderate-loss events from counties with weak Reddit or news coverage, confirming that sparse online signals lead to underestimation of impact.

E. Error Analysis and Feature Insights

Feature-importance analysis shows that the most influential predictors are wildfire-specific FEMA NRI metrics (e.g., WFIR_EAL_TOTAL), the social-vulnerability-to-resilience ratio (SOVI_RESL_RATIO), total wildfire-related Reddit post count, damage-related news article count, and mean annual temperature and precipitation. These findings reinforce that blending structural risk indicators with dynamic online signals provides the most robust estimates of county-level impact.

F. Interpretation and Reliability

Overall performance is strong (macro-F1 0.76), but per-class imbalance remains a challenge. The model distinguishes extremes (low vs. high damage) reliably yet is less confident in borderline Medium cases. Future work will explore calibrated probability outputs, ordinal-regression objectives, and spatiotemporal modeling to improve interpretability and robustness.

VI. FUTURE WORK

Our digital twin component demonstrates that combining offline hazard databases with online social and news signals can provide meaningful predictions of wildfire damage severity at the county level. Several extensions could further improve realism and performance:

- **Higher-quality climate data:** Integrate county-level daily or monthly climate time series from PRISM, NOAA CDO, or Daymet, and create seasonal or drought indices.
- **Expanded online coverage:** Increase the number of classified Reddit posts and news articles, using more API keys and historical archives.
- **Temporal and spatial modeling:** Move beyond independent events to incorporate time trends and spatial relationships between neighboring counties (e.g., graph-based models).
- **Advanced architectures:** Explore deep learning approaches for tabular data and compare them to gradient-boosting ensembles in terms of both performance and interpretability.
- **Uncertainty-aware decision support:** Calibrate predicted probabilities and tie them to decision thresholds for emergency planners, insurers, and local governments.
- **Interactive deployment:** Wrap the trained model into a web API or dashboard that allows users to select counties and view predicted severity alongside SHEL DUS, NRI, and online indicators.

VII. LIMITATIONS AND REPRODUCIBILITY

We note several limitations and provide guidance for reproducibility:

- **Temporal and spatial bias:** Online signals (Reddit and news) are concentrated in recent years and in more populous counties; older events and rural counties are often under-covered.
- **LLM classification dependence:** We used a proprietary LLM (Claude 3 Haiku) for text classification; behavior

and outputs of such models may change over time and should be validated when re-running the pipeline.

- **Reproducibility artifacts:** The review package (see Section ??) includes code, mapping logs, SHELDUS event rows for the 1,020 units of analysis, and a README listing exact library versions and random seeds. Reproducible results require API keys for some news sources (not provided).
- **Future checks:** we recommend a human-in-the-loop validation step for the LLM labels and additional sensitivity experiments that withhold online signals at train time to quantify robustness.

VIII. CONCLUSION

We developed a Digital Twin for Wildfire Impact Assessment that predicts county-level property damage severity (Low/Medium/High) by integrating SHELDUS wildfire losses, FEMA National Risk Index features, climate data, Reddit social media signals, and online news coverage. Our pipeline includes detailed county-level data aggregation, extensive feature engineering, class-imbalance handling, and ensemble machine-learning models.

Using SHELDUS as an offline ground truth, our best ensemble model achieves 79% accuracy on a held-out test set of wildfire events. Qualitative analysis shows that predictions align well with high-impact events in many counties, while also revealing data blackspots where online signals are sparse. Overall, this project illustrates how social media mining and multi-source data fusion can enhance disaster impact assessment and serve as a building block for more comprehensive digital twin systems.

ACKNOWLEDGMENTS

We thank Prof. Huan Liu, the CSE 472 teaching staff, and the TA Saketh Vishnubhatla for guidance on the Digital Twin project. We also thank the providers of the SHELDUS database, the FEMA National Risk Index, Reddit, and the news APIs used in this work.

IX. REFERENCES

REFERENCES

- [1] SHELDUS – Spatial Hazard Events and Losses Database for the United States. <https://www.sheldus.org/>, accessed 2025.
- [2] Federal Emergency Management Agency, “National Risk Index,” <https://hazards.fema.gov/nri>, accessed 2025.
- [3] PRISM Climate Group, Oregon State University, “PRISM Climate Data,” <https://prism.oregonstate.edu/>, accessed 2025.
- [4] Reddit Inc., “Reddit API Documentation,” <https://www.reddit.com/dev/api/>, accessed 2025.
- [5] GNews API, Currents API, NewsAPI, and NewsData.io, accessed 2025.
- [6] Anthropic, “Claude 3 Haiku,” <https://www.anthropic.com>, accessed 2025.
- [7] H. Gao, G. Barbier, and R. Goolsby, “Harnessing the crowdsourcing power of social media for disaster relief,” *IEEE Intelligent Systems*, 2011.