CHAPTER 3

RESEARCH DESIGN

3.1 Overview of the Research Process

This chapter introduces the overall research process adopted in this study. Since the objective is to construct a wildfire risk prediction model with practical value, the research design was developed based on both the existing literature and some of the identified limitations in previous work. In general, the research process includes five main steps: data collection, data preprocessing, feature selection, model training, and model evaluation.

First, a series of publicly available datasets related to environmental conditions and human activities were collected. These datasets came from different sources, so data preprocessing was necessary to address missing values and to align the time and spatial attributes. Once the data was ready, several supervised learning models were trained using selected variables. These models included both relatively simple and more advanced algorithms. Their performance was compared using commonly used evaluation indicators such as accuracy and precision. Based on the results, one or two models with better performance were selected for further analysis. Throughout the process, efforts were made to balance model complexity with interpretability and efficiency, in the hope that the final model can be applied in practical wildfire monitoring and early warning scenarios.

3.2 Data Collection and Preprocessing

This study uses the "California Wildfire Damage (2014–Feb 2025)" dataset from Kaggle as the main source of data. Covering over ten years of wildfire activity in California, this dataset includes essential information such as the name and location of each fire, the date it occurred, the area burned, the number of structures affected, and other relevant attributes. These records form the basis for understanding how different environmental and human-related factors might be linked to wildfire events.

Compared to synthetic or simulated datasets, this type of real-world record has stronger practical relevance. What makes it particularly valuable is that it doesn't just include natural factors, but also human-related ones—such as the suspected cause of the fire, whether it was arson, lightning, or accidental ignition—giving us a broader view of how wildfires behave in real settings.

Before modeling, the dataset went through several cleaning and transformation steps. Incomplete records were first dealt with—some rows had missing fields like containment dates or the number of structures damaged. Depending on the importance of the field and the proportion of missing data, we either filled in values based on similar samples or dropped the row altogether. Categorical variables such as fire cause were converted into numerical values, so that models could recognize them during training. We also extracted time-related features—like month or season—from date fields, to reflect the seasonal patterns that often exist in wildfire occurrences. As for spatial features, we kept the original latitude and longitude. While the dataset doesn't include information like elevation or urban proximity, such factors could be considered later by combining with GIS sources, if needed. For the target variable, we set a binary label: "1" means a wildfire occurred, and "0" means it didn't. Since the dataset mainly includes positive samples (i.e., actual fire events), we generated

synthetic negative samples by randomly picking date-location combinations where no fire was recorded. This helped create a balanced dataset for model training.

Altogether, after filtering, cleaning, and constructing key features, the dataset was ready for modeling. The next section will explain how specific features were selected and transformed for use in machine learning algorithms.

3.3 Feature Selection and Engineering

After the data was cleaned and organized, the next step was to decide which variables might actually be useful for predicting wildfire risk. Not every field in a dataset offers helpful information—some may provide strong signals, while others could introduce noise or redundancy. So this stage involved selecting a group of features that were both meaningful in practice and analytically reasonable.

Some features were fairly straightforward to include. For instance, “fire size” reflects the scale of an event and can also signal the underlying severity of fire conditions. The “cause” of the fire was also retained, since it helps distinguish between natural and human-related fire events. From the time-related data, we extracted information such as the month of discovery and whether it occurred during a fire-prone season. Past studies often highlight seasonal patterns in wildfire outbreaks, so these time-based features helped the model capture such tendencies more effectively.

To prepare categorical variables like fire cause for modeling, we used one-hot encoding, which allows the model to recognize each category as a separate input without assuming any specific order. In terms of spatial features, latitude and longitude were included directly, providing basic geographic context for each record.

Although additional spatial features—like vegetation cover or proximity to roads—could be helpful, they were not available in this version of the dataset and may be added later through GIS integration.

We also created a few new features by transforming existing ones. For example, we added a binary variable to flag whether the fire was caused by human activity. We also calculated the logarithmic value of fire size to reduce the impact of outliers, especially extremely large fires that might skew the model's learning. These steps helped improve the informativeness and balance of the dataset.

Taken together, this phase aimed to build a feature set that reflects domain understanding while also supporting the model's ability to learn meaningful patterns. In the next section, we will describe how the modeling process was carried out using these selected variables.

3.4 Model Selection and Training Strategy

After preparing the dataset and constructing the relevant features, the next important step is choosing the right model and deciding how to train it. Since the task of wildfire prediction involves classifying whether or not a fire might occur under certain conditions, this study treats it as a binary classification problem.

At the beginning, several classic machine learning models were considered, including logistic regression, decision trees, and support vector machines. These models are relatively easy to implement and interpret, which makes them a good starting point. However, after some early-stage testing, we found that more advanced models, especially ensemble methods like Random Forest and XGBoost, performed

better with the data, especially when dealing with more complex feature combinations.

In this study, we finally selected Random Forest as the main model. The reason is that Random Forest tends to handle nonlinear relationships well, is less likely to overfit compared to single decision trees, and offers some level of interpretability. Each tree in the forest learns from a random subset of the data and features, which helps reduce bias and variance. Moreover, it handles missing data and variable importance rankings quite naturally.

As for the training strategy, we divided the dataset into training and testing sets in an 80:20 ratio. The training set is used to build the model, while the testing set evaluates its generalization ability. To further improve the model's robustness, we also adopted 5-fold cross-validation during training. This means the training data is split into five parts, and the model is trained five times, each time using four parts for training and one for validation. The average performance across the five runs is then used as the evaluation result.

To avoid overfitting, we also fine-tuned some key hyperparameters of the Random Forest, including the number of trees, the maximum depth of each tree, and the minimum number of samples required to split a node. These were selected based on grid search and cross-validation results.

3.5 Modeling Methods for Wildfire Prediction

In this section, the focus shifts from preparing the dataset to building the predictive models themselves. Based on the insights gained from exploratory analysis

and data preprocessing, we now turn to selecting suitable machine learning algorithms and designing a training strategy that aligns with the goals of this study.

3.5.1 Model Selection Rationale

Given that the task is to predict wildfire occurrence (binary classification: fire or no fire), this study adopts a supervised learning approach. Based on existing literature and the structure of the Kaggle dataset—which includes both categorical and continuous features like temperature, precipitation, vegetation index, and human activity data—three main algorithms are selected: Random Forest, Logistic Regression, and XGBoost. Each of these models has its own strengths: Random Forest handles non-linear interactions well and is robust to overfitting, Logistic Regression provides interpretability and serves as a baseline, while XGBoost often performs well with structured tabular data and can handle missing values effectively.

3.5.2 Model Training Plan & Parameters

All models will be trained on the processed dataset using a unified pipeline. Basic hyperparameters will be set as follows: for Random Forest, 100 trees will be used; for Logistic Regression, L2 regularization will be applied; for XGBoost, a learning rate of 0.1 and a maximum tree depth of 6 will be used initially. These settings are commonly used in similar wildfire studies and serve as a reasonable starting point. Further tuning will be considered if performance needs to be improved.

3.5.3 Data Splitting Strategy

The dataset will be divided into a training set and a test set using a 70:30 ratio. Since the data spans over a decade (2014–Feb 2025), we will ensure temporal

consistency by assigning earlier years to the training set and more recent data to the test set. This setup better mimics a real-world scenario where past data is used to predict future risks.

3.5.4 Evaluation Metrics & Validation Design

Model performance will be assessed using standard classification metrics, including accuracy, precision, recall, F1-score, and AUC-ROC. These indicators help measure the balance between correctly identifying fires and minimizing false alarms. Since wildfire datasets often show imbalance (i.e., fewer positive fire samples), F1-score and AUC-ROC are especially important. To enhance robustness, 5-fold cross-validation will be applied during training.

These evaluation strategies are not only aimed at measuring how well the models perform on known data but also at ensuring that they generalize to unseen situations—a key consideration when dealing with unpredictable natural events like wildfires. With the model design and training plan in place, the next step is to summarize the entire modeling workflow and reflect on how each component fits together in the broader research framework.

3.6 Summary of Workflow

This chapter outlined the overall design of the wildfire risk prediction study, from understanding the problem to developing machine learning models. The research began by clarifying the problem statement and research objectives, which provided a foundation for the entire modeling process. Based on this, relevant data were collected from an open-source wildfire damage dataset, including geographic, environmental,

and fire occurrence records. A brief dataset overview and metadata description helped to identify useful features and potential modeling challenges.

Through exploratory data analysis, we examined the basic distribution of key variables, visualized patterns, and checked for correlations that might indicate underlying relationships between factors. The following data preprocessing stage focused on cleaning, integrating, and transforming the raw dataset into a structured format that could be effectively used by machine learning algorithms. Attention was given to handling missing values, standardizing features, and selecting those most relevant to wildfire risk.

In the final stage of this chapter, we proposed a model development plan. Different machine learning methods were considered, and the rationale behind model selection was explained. The training strategy, data splitting approach, and evaluation metrics were discussed in order to ensure that the model could provide both accurate predictions and generalizability.

Altogether, the workflow outlined in this chapter serves as a roadmap for the empirical implementation described in Chapter 4. It connects problem understanding, data preparation, and algorithm design into a logical sequence and provides a clear reference for future replication or improvement.