fire

MDPI

Article

Predicting Wildfire Damage Severity with Composite Indexing and Fire Weather Features: A Case Study in Gangwon Province, South Korea

Jaeun Choi [1], Wonseok Yang [2], Seokju Kim [2], Ahyeon Jeong [2], Jiwoo Baek [2], Nanggyun Ko [2], Chumni Jeon [3] and Eun Sang Jung [4,*]

[1] Major in Big Data Management, Division of Business Administration, Kwangwoon University, Seoul 01897, Republic of Korea; juchoi@kw.ac.kr
[2] Division of Business Administration, Kwangwoon University, Seoul 01897, Republic of Korea
[3] Corporate Relations Office, Korea Telecom, Seoul 03155, Republic of Korea; cm.jeon@kt.com
[4] Department of Bioenvironmental Energy, Pusan National University, Miryang 50463, Republic of Korea
* Correspondence: esjung@pusan.ac.kr

Abstract

Accurate wildfire prediction increasingly determines whether emergency resources arrive before a disaster becomes uncontrollable, yet the dominant paradigm reduces the problem to binary occurrence, offering no estimate of the severity that drives suppression planning. This study develops a machine-learning framework for four-class wildfire severity prediction, conditional on ignition, from weather-station observations and calendar terms alone. We construct a composite severity index (CSI) by applying principal component analysis to five damage dimensions (burned area, suppression equipment, personnel, duration, and property loss) recorded for 868 wildfires in Gangwon Province, South Korea (2011–2022) and pair standard observations with effective humidity and six indices of the Canadian Forest Fire Weather Index (FWI) System. Under a leakage-safe protocol, the strongest tree ensembles reach a macro F1 of 0.46 to 0.50 (recommended configuration: 0.41 ± 0.03 across 20 repeated splits) against a four-class chance level of 0.25, and the recommended Random Forest attains an extreme-class recall of 0.474; the CSI target outperforms burned area by 5.5 macro-F1 points under identical inputs. A weather-only screen separates extreme from non-extreme events with an ROC AUC of 0.758, capturing 47% of extreme events at a 20% alert budget. We also quantify how oversampling misplaced before the train-test split inflates the macro F1 to 0.65–0.83, a cause for caution for the severity-prediction literature.

Keywords: wildfire damage severity; composite severity index; principal component analysis; random forests; effective humidity; Canadian forest fire weather index system; Gangwon province; operational forecasting

Academic Editors: Washington Franca-Rocha, António Vieira and Marcos Francos

Received: 29 May 2026
Revised: 14 July 2026
Accepted: 17 July 2026
Published: 20 July 2026

1. Introduction

Wildfires are among the most destructive natural disasters of the 21st century. Each year, global fire activity drives deforestation, carbon emissions, and losses of human and animal life [1], with cascading effects on air quality, biodiversity, and the carbon cycle. Reanalysis records show that fire weather seasons lengthened across roughly a quarter of the Earth's vegetated surface between 1979 and 2013 [2]. Climate projections indicate that fire-favorable conditions, such as prolonged drought, elevated temperatures, and high wind variability, will intensify across temperate and boreal regions in the coming

decades [3]. In South Korea, this risk is concentrated in Gangwon Province. Situated along the eastern slope of the Taebaek Mountain range, Gangwon is subject to strong Föhn-type winds and dry continental air that drive fast-moving spring fires through the dense coniferous forests [4]. The 2019 Goseong–Sokcho fire illustrated the catastrophic potential: it forced the overnight evacuation of more than 4000 residents, and the east-coast fire complex it belonged to left 1524 residents displaced and a confirmed 129.1 billion KRW in property damage, prompting a nationwide mobilization of fire-service resources [5].

Substantial progress has been achieved in the prediction of binary fire occurrence. Machine learning models such as Random Forest, gradient boosting, and deep learning routinely exceed 90% accuracy in fire-versus-no-fire tasks when provided with comprehensive meteorological, topographic, and vegetation inputs [6,7]. These models serve fire risk mapping well but share a fundamental limitation: once ignition is assumed, they provide no estimate of the severity of the resulting event. For emergency managers, the critical question is not whether a fire will start but how large, how damaging, and how resource-intensive it will be. The distinction between a quickly contained brush fire and a multiday conflagration requiring inter-regional mobilization is the information needed to allocate suppression assets, issue evacuation orders, and activate mutual aid agreements [3].

A few studies that moved beyond binary prediction have typically targeted burned areas as a continuous outcome [8]. The burned area is widely recorded and easily measured; however, it is an incomplete proxy for the overall impact. Two fires of equal spatial extent can differ drastically in their suppression burden, property damage, and threat to life, depending on the terrain, proximity to settlements, and response effectiveness [9,10]. The multidimensional aspects of fire impact (personnel, equipment, duration, and economic damage alongside the physical extent) are routinely documented in post-event reports but have seldom been integrated into a unified prediction target.

This study addresses this research gap in two ways. The first is the construction of a composite severity index (CSI). Principal component analysis (PCA) is applied to five post-event damage dimensions (burned area, suppression equipment deployed, personnel mobilized, suppression duration, and property damage), and the resulting score is discretized into four ordinal classes: low, medium, high, and extreme. This target captures the overall operational and socioeconomic burden of a wildfire rather than any single aspect. Ablation supports its use: under identical inputs, the composite target is more predictable than burned area alone (macro F1 0.465 versus 0.410), while measuring what suppression planning actually needs. These four classes mirror the tiers of Korea's graded mobilization protocol, and Section 5.2 maps each predicted class to concrete emergency actions.

The second contribution is the systematic assembly of a meteorological feature set that augments raw weather observations with derived fire-weather indicators. Station-level measurements are supplemented by effective humidity, a cumulative drying metric used operationally by the Korea Meteorological Administration, and by six indices of the Canadian Forest Fire Weather Index (FWI) System, which encode fuel moisture dynamics from hourly to seasonal timescales [11]. Ablation attributes the largest consistent gain to the FWI-system indices: removing them lowers macro F1 by 5.2 percentage points, and a model restricted to raw meteorological and temporal inputs loses 5.7 points.

Seven candidate models and an ordinal variant were evaluated on 868 wildfire records from Gangwon Province (2011–2022). Under a leakage-safe protocol, Extra Trees achieved the best overall scores (accuracy 0.506, macro F1 0.497), and the tuned Random Forest was selected for operational use on the strength of its balanced class profile and extreme-class recall of 0.474, identifying roughly half of the catastrophic events at the time of ignition, before the eventual scale of the event materialized. A complementary weather-only screen separates extreme from non-extreme events with an ROC AUC of 0.758. Every input

feature can be computed in real time from standard weather-station data and the incident timestamp, requiring no satellite imagery or terrain models, so the framework can support operational severity assessment conditional on ignition.

The remainder of this paper is organized as follows: Section 2 reviews related work on fire-weather indices, machine learning for wildfire prediction and severity assessment, and the Korean context. Section 3 describes the data, target construction, feature engineering, and the experimental design. Section 4 presents the results in six parts: model comparison, ablation, feature importance, per-class performance, robustness and sensitivity analyses, and extreme-event screening. Section 5 discusses the implications and limitations of this study. Section 6 concludes.

2. Related Work

2.1. Fire Weather Indices and Meteorological Predictors

The relationship between meteorological conditions and wildfire behavior has been studied for over a century. Among fire danger rating systems, the Canadian Forest Fire Weather Index (FWI) System developed by Van Wagner [11] remains the most widely adopted. It decomposes fire weather into sub-indices that track fine fuel moisture (FFMC), intermediate duff moisture (DMC), deep soil drought (DC), fire spread potential (ISI), buildup of available fuel (BUI), and an overall intensity rating (FWI). This system has been operationally validated across Canada, the European Union, and numerous other jurisdictions [12]. Operational systems now generate FWI-based danger ratings at continental to global scale from numerical weather prediction forcings, with demonstrated skill in flagging large-fire conditions [13], and reanalysis-based FWI records document a global lengthening of fire weather seasons [2].

In South Korea, the Korea Meteorological Administration adopted effective humidity as a complementary operational indicator [14]. This metric applies an exponentially weighted moving average to past humidity values, thereby capturing cumulative fuel drying rather than instantaneous atmospheric moisture. Korean fire studies consistently rank effective humidity as the strongest meteorological predictor of fire occurrence on the Peninsula [14]. Beyond these indices, broader studies have established temperature, humidity, wind speed, and precipitation as the primary meteorological drivers of ignition and spread [4,15,16]. Shmuel and Heifetz [1] showed that machine-learning models substantially outperform regression baselines for global wildfire occurrence and burned-area prediction, whereas Kim et al. [17] showed that ensemble-merged fire weather indices significantly improved the wildfire forecasting accuracy in Gangwon Province.

2.2. Machine Learning for Wildfire Prediction

Machine learning has rapidly displaced earlier statistical approaches to wildfire forecasting. Binary occurrence prediction has received the most attention: Prapas et al. [6] applied deep learning for Mediterranean fire-danger forecasting, Sayad et al. [18] combined satellite imagery with machine learning for occurrence modeling, Xu et al. [19] developed an ensemble-based fire detection system, and gradient boosting variants (XGBoost, LightGBM, CatBoost) consistently outperform standalone trees and logistic regression in fire risk assessment [7,20,21].

Despite this progress, an overwhelming majority of ML-based wildfire studies remain confined to binary tasks. The few studies that have moved beyond this have focused on continuous burned-area regression [8] or post-fire severity mapping from satellite imagery [22,23]. Multiclass severity classification based on meteorological conditions observable at ignition, the approach taken here, remains largely unexplored.

Among the model families, tree-based ensembles dominate tabular wildfire datasets. Studies reviewed by Jain et al. [3] confirmed that Random Forest and gradient boosting models frequently achieve the highest prediction accuracy among classical ML methods for fire occurrence and risk mapping. The strength of these methods lies in their ability to handle mixed-type inputs, capture nonlinear interactions, and perform well on datasets of modest size [20,24], which are characteristics that align closely with regional wildfire data.

2.3. Wildfire Severity Assessment and Target Variable Construction

The definition of wildfire severity has generated substantial discussion. The most common approach equates it with the burned area [1,8]; however, multiple authors have argued that this is incomplete. San-Miguel-Ayanz et al. [9] emphasized that socioeconomic consequences, such as property damage, infrastructure loss, and suppression costs, can vary enormously among fires of similar size. Tedim et al. [10] advocated multidimensional severity frameworks that integrated physical, ecological, and socioeconomic indicators.

In post-fire assessment, remote sensing indices such as dNBR are standard for landscape-scale burn mapping [22]; however, these require post-fire satellite passes and are unavailable for pre-event prediction. Keeley's distinction among fire intensity, fire severity, and burn severity [25] underlines that no single physical metric captures impact, and recent mapping studies that grade severity from Sentinel-2 imagery with random forests [26] remain post-event products by construction. Conceptually, the need for composite severity grades parallels other hazard domains (the Saffir–Simpson scale for hurricanes and the Modified Mercalli scale for earthquakes), in which a single rating integrates multiple impact dimensions. Beyond named scales, composite indicators that weight multiple dimensions, frequently through PCA, are standard instruments in hazard and vulnerability research [27,28]; the CSI follows this tradition with weights estimated from the damage data themselves.

Related work has modeled the drivers of fire outcomes rather than composite severity itself. Sevinc et al. [29] modeled possible fire causes using Bayesian networks. Costafreda-Aumedes et al. [30] reviewed human-caused fire occurrence models and highlighted the importance of integrating human and physical resource variables into predictive frameworks. However, to the best of our knowledge, no prior study has combined PCA-based index construction with multiclass classification and systematic ablation to validate the predictive advantage of a composite target in a wildfire context.

2.4. Wildfire Research in South Korea

Korean wildfire research reflects the peninsula's distinctive fire regime: a pronounced spring fire season (March–May), driven by dry continental air and strong Föhn-type winds descending the Taebaek range [31]. Lim and Chae [31] applied the FWI System to fire danger assessments across South Korea and found that FWI values exhibited strong seasonal variations with significant increases in spring. Kim et al. [17] extended wildfire forecasting in the Gangwon Province through ensemble-based index merging, whereas Jeon et al. [32] explored deep-learning architectures that incorporate human factors for occurrence prediction in the same region. Studies on Korean FWI adaptation [31] examined Canadian sub-indices under local fuel types and climates.

Despite this body of work, Korean wildfire research has not systematically addressed multiclass severity predictions. Existing studies have predicted binary occurrences or estimated single-target burned area. The absence of a composite severity framework that accounts for suppression burden, resource deployment, and property damage represents a gap that this study aims to address.

3. Materials and Methods

This section describes the methodological pipeline of the proposed framework. Each design decision, from target variable construction to model selection, is presented alongside the reasoning that motivated it so that the rationale is transparent and the pipeline is reproducible in other regional contexts.

3.1. Study Area and Data Sources

Gangwon Province occupies the northeastern corner of South Korea and is the country's most wildfire-prone region (Figure 1). Its vulnerability is not merely a matter of fire frequency. Gangwon fires are disproportionately severe because the province's topographic and climatic conditions amplify fire behavior once ignition occurs. The Taebaek Mountain range channels strong Föhn-type winds down its eastern slope and rapidly desiccates surface fuel over extensive coniferous forests. This combination of steep terrain, dense fuel loading, and sudden wind events means that fires in Gangwon escalate faster and demand larger suppression responses than fires with comparable ignition conditions on the Korean Peninsula [4,31]. Therefore, selecting Gangwon as a case study maximized the dynamic range of severity outcomes available for classification, from minor brush fires extinguished within hours to multiday conflagrations requiring inter-provincial mobilization.

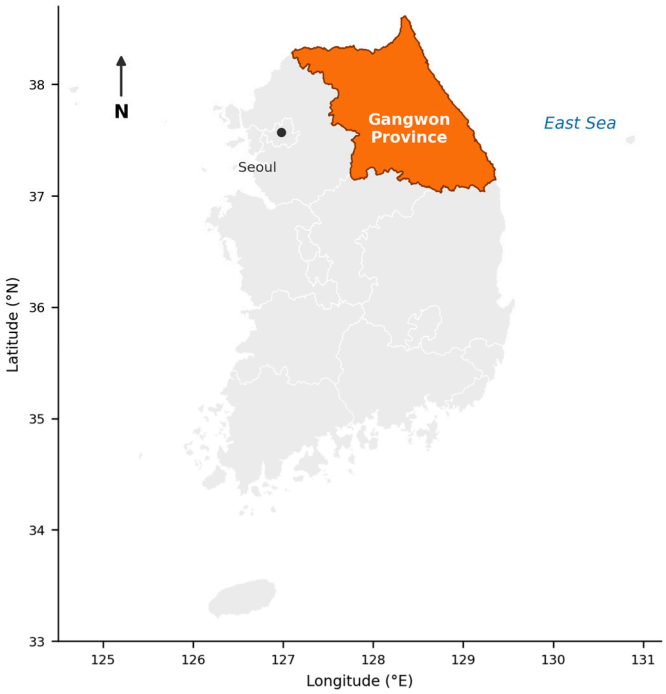

Figure 1. Location of the study area, Gangwon Province (orange), in the northeastern Republic of Korea.

This study draws on 868 wildfire records spanning 12 consecutive fire seasons (2011–2022). This window captured substantial year-to-year climatic variability, including both anomalously dry and relatively wet spring seasons, while remaining within a period of consistent data collection standards. Wildfire damage records were obtained from the Fire Safety Big Data Platform operated by the National Fire Agency of Korea (dataset: Wildfire Status Data of Gangwon Province, 2011–2022). Each record documents not only burned area but also four additional operational attributes collected during and after suppression: the number of equipment units deployed, total personnel mobilized, suppression duration, and estimated property damage (in nominal KRW, as recorded). The co-availability of these five dimensions in a single, consistently coded dataset makes composite severity indexing

feasible. Most international fire databases record only an area, or an area plus a single cost estimate, which constrains analysis to a less expressive target.

Meteorological observations were retrieved from the Korea Meteorological Administration's Automated Surface Observation System (ASOS) for the station nearest to the ignition point of each fire at the recorded time of occurrence. The raw variables included temperature, ground surface temperature, wind speed, dew point temperature, vapor pressure, sea-level pressure, local atmospheric pressure, relative humidity, daily precipitation, snow depth, horizontal visibility, and categorical indicators of precipitation and snowfall in the preceding hour. Two families of derived indicators were computed from these inputs, as described in Section 3.3.

3.2. Composite Severity Index

The choice of target variable is arguably the most consequential design decision in any prediction study because it determines what the model learns to distinguish. In wildfire research, the default target is the burned area, a natural choice given its measurability, but one that collapses a multidimensional phenomenon into a single spatial metric. Two fires of identical hectarages can impose radically different burdens on emergency services depending on their proximity to settlements, terrain accessibility, and fire intensity. A 50-hectare fire in a remote alpine grassland may require a handful of personnel for a few hours; a 50-hectare fire on the urban-wildland interface of Gangwon's east coast can mobilize thousands of responders, destroy residential property, and persist for days. Predicting "50 hectares" for both events provides identical output where the operational reality is entirely different.

To address this, we constructed a CSI that integrates all five recorded damage variables (burned area, equipment, personnel, duration, and property damage) into a single score using PCA [33]. PCA is chosen over alternative aggregation strategies for a specific reason: it lets the data determine the relative weighting of each dimension rather than imposing expert judgment. An equal-weighting scheme treats a marginal increase in equipment deployment as equivalent to the same proportional increase in burned area, ignoring the fact that these variables have different variance structures and relationships with the overall severity. Although defensible in principle, expert weighting introduces subjective choices that are difficult to reproduce across regions and datasets. PCA avoids both problems by extracting the orthogonal axes of the maximum variance from the standardized damage matrix.

Each variable is first standardized via Z-score normalization:

$$z_i = (x_i - \mu_i)/\sigma_i \quad (1)$$

Because wildfire damage variables are heavily right-skewed, each variable was transformed with log(1 + x) before standardization. The PCA was then applied to the standardized matrix. PC1 explains 60.0% of total variance (eigenvalue 3.00) and loads almost evenly on burned area, personnel, equipment, and property loss, capturing the overall scale of mobilization. PC2 explains a further 16.2% (eigenvalue 0.81) and is dominated by suppression duration, separating events that burn long relative to their mobilization scale. Only PC1 satisfies the Kaiser criterion (eigenvalue > 1.0); we retain two components because they jointly account for 76.2% of the variance and because the second axis carries a distinct operational meaning. The composite index was computed as the variance-weighted sum of the retained components:

$$\mathrm{CSI} = \Sigma_j \, (w_j \cdot \mathrm{PC}_j) \quad (2)$$

where w_j is the proportion of variance explained by the jth component. Variance weighting ensures that the dominant mobilization axis drives the index, whereas the prolonged-suppression axis provides discriminative refinement.

The continuous CSI was then discretized into four ordinal classes. The choice of four rather than three or five reflects an explicit mapping of the operationally distinct response tiers. Emergency management in South Korea operates on a graded mobilization protocol: small incidents handled by local stations (low), multistation responses within a district (medium), provincial-level coordination involving specialized assets (high), and national-level mobilization with inter-provincial mutual aid (extreme) [34]. Aligning severity classes with these operational tiers ensures that the model's output corresponds to actionable resource decisions rather than arbitrary statistical bins.

Class boundaries were set via a percentile-based scheme informed by the heavy right skew of wildfire damage distributions; the top 10% of CSI scores defined the extreme class, and the remaining 90% were divided equally into low, medium, and high (each approximately 30%). The resulting cut points are the 30th, 60th, and 90th percentiles of the CSI distribution, which in this sample fall at -0.589, -0.009, and $+1.103$, yielding class counts of 261, 260, 260, and 87 across the 868 records. Because the scheme is percentile-based, transfer to another region uses that region's own percentiles rather than these absolute values. The 10% threshold for extreme isolates the catastrophic tail, events whose suppression burden is qualitatively different from the rest of the distribution, while preserving enough samples (87) for the model to learn from. The equal subdivision of the lower 90% avoids introducing an artificial class imbalance where none exists in the underlying severity continuum.

Throughout this paper, damage severity denotes the phenomenon itself, the multidimensional burden a fire imposes, whereas the CSI denotes our operational measure of it. A contrast from the dataset illustrates how the five indicators jointly determine the class. A 1.0 ha fire in April 2014 was handled by 112 personnel with 23 equipment units and left no recorded property loss; its CSI of -0.79 places it in the low class. A 0.5 ha fire in April 2020, half the area, mobilized 313 personnel and 51 equipment units over a 5.6 h suppression with recorded property loss, and its CSI of 1.95 places it in the extreme class. The smaller fire is the more severe event on every dimension except area, which is precisely the information a single burned-area target discards.

3.3. Feature Engineering

The feature set is designed around a central insight that differentiates severity prediction from occurrence prediction: while ignition depends primarily on instantaneous atmospheric conditions (a single hot, dry, windy hour can start a fire), severity is shaped by the cumulative state of the fuel complex leading up to the event. A fire that ignites after weeks of progressive drying behaves fundamentally differently from a fire that starts after a brief dry spell, even if the weather at the moment of ignition is identical. This distinction motivated the inclusion of temporal moisture-memory features alongside standard meteorological observations. Twenty-one predictor features were assembled from four categories: raw meteorological observations (including precipitation, snow, and visibility terms), derived humidity, six FWI-system indices, and temporal terms. Table 1 summarizes the complete set.

Raw meteorological variables. Twelve observations were obtained directly from KMA ASOS: temperature, ground surface temperature, wind speed, dew point temperature, vapor pressure, sea-level pressure, local atmospheric pressure, daily precipitation, snow depth, horizontal visibility, and categorical indicators of precipitation and snowfall in the preceding hour. Together, these capture the atmospheric state governing fuel dryness and

convective fire behavior, along with recent wetting events that suppress spread. The set includes two pressure variables (sea-level and local), which serve as proxies for synoptic-scale weather patterns: high-pressure ridges associated with Föhn events in the Taebaek range are among the most consistent precursors to severe fire weather in Gangwon [4,17].

Table 1. Summary of predictor features.

No.	Feature	Category	Description
1	Temperature	Raw Met.	Air temperature (°C)
2	Ground Surface Temp.	Raw Met.	Ground surface temperature (°C)
3	Wind Speed	Raw Met.	Average wind speed (m/s)
4	Dew Point Temperature	Raw Met.	Dew point temperature (°C)
5	Vapor Pressure	Raw Met.	Atmospheric vapor pressure (hPa)
6	Sea-Level Pressure	Raw Met.	Sea-level atmospheric pressure (hPa)
7	Local Atm. Pressure	Raw Met.	Local atmospheric pressure (hPa)
8	Precipitation	Raw Met.	Daily precipitation (mm)
9	Snow Depth	Raw Met.	Snow depth on the ground (cm)
10	Visibility	Raw Met.	Horizontal visibility (10 m units)
11	Precip. Category (t − 1 h)	Raw Met.	Precipitation category one hour earlier
12	Snow Category (t − 1 h)	Raw Met.	Snowfall category one hour earlier
13	Effective Humidity	Derived	Weighted moving average of relative humidity (%)
14	FFMC	FWI Sub-index	Fine Fuel Moisture Code
15	DMC	FWI Sub-index	Duff Moisture Code
16	DC	FWI Sub-index	Drought Code
17	ISI	FWI Sub-index	Initial Spread Index
18	BUI	FWI Sub-index	Buildup Index
19	FWI	FWI Sub-index	Fire Weather Index (composite rating)
20	Month of Occurrence	Temporal	Month when the wildfire occurred (1–12)
21	Hour of Occurrence	Temporal	Hour of day when the wildfire occurred (0–23)

Effective humidity. The instantaneous relative humidity fluctuates on an hourly time scale and poorly represents the cumulative moisture state of the surface fuels. The Korea Meteorological Administration's effective humidity addresses this limitation by applying an exponentially decaying weighted average over the preceding days as follows:

$$\mathrm{EH} = (H_0 + 0.7H_1 + 0.49H_2 + 0.343H_3 + \cdots)/(1 + 0.7 + 0.49 + 0.343 + \cdots) \quad (3)$$

where H_0 is the current day's minimum relative humidity and H_t is the value t days prior. The decay factor of 0.7 halves the weight roughly every two days, retains about one-third at three days, and leaves negligible weight beyond a week, creating a multiday moisture memory tuned to the drying time scale of fine surface fuels in Korean forests. Values below 50% are empirically associated with elevated fire danger in Korean operational practice [14]. The inclusion of this feature rests on the hypothesis that severity depends not only on how dry the air is at the moment of ignition but on how long and how deeply the fuel bed has desiccated beforehand; under the corrected protocol its marginal contribution is modest and within split-to-split noise (Section 4.2), and we retain it for its operational familiarity and its distinct temporal scale.

FWI-system indices. Six indices of the Canadian Forest Fire Weather Index System [11] were computed from daily meteorological data. Each encodes the moisture state or expected behavior of the fuel bed at a distinct depth and temporal scale: FFMC tracks fine surface litter with a response time of hours; DMC reflects intermediate duff moisture over approximately 12 days; DC captures deep soil drought on a seasonal scale (about 52 days); ISI integrates FFMC with wind speed to estimate the initial spread rate; BUI combines DMC and DC into the total fuel available for combustion; and FWI integrates ISI and BUI

into an overall intensity rating. These indices have been validated globally and calibrated for Korean conditions [14,31]. Their value for severity prediction lies in the longer-memory codes (DC, DMC, and BUI); when deep organic layers are thoroughly dry, fires burn hotter, penetrate deeper into the ground, and resist suppression more stubbornly, all factors that escalate severity beyond what surface conditions alone would predict.

Temporal features. The month of occurrence (1 to 12) captures the seasonal pattern of fire activity: in Gangwon Province, both frequency and severity peak during March to May, when low humidity, strong winds, and dry fuel accumulated over winter converge. The hour of occurrence (0 to 23) adds the diurnal cycle of temperature, humidity, and human activity. Including these terms lets the model use seasonal and diurnal priors even when the meteorological variables for a given day do not clearly signal elevated risk.

3.4. Candidate Models

Seven classifiers spanning three algorithmic families were evaluated, each with its hyperparameters tuned under the protocol of Section 3.5: the gradient boosting machine (GBM) [20], LightGBM [21], XGBoost [24], Random Forest [35], Extra Trees, a support vector machine (SVM) with a radial basis function kernel, and a feedforward artificial neural network (ANN). The primary formulation is nominal: tree ensembles with nominal targets are the established default for tabular severity data, impose no assumptions about the spacing of adjacent severity thresholds, and allow direct control of per-class errors, which the asymmetric-cost criterion of Section 3.5 requires. A Frank and Hall ordinal decomposition of the best tree ensemble was added as an eighth, ordinal reference model to address the ordered nature of the severity classes.

The emphasis on tree-based ensembles is deliberate. Prior work has consistently demonstrated their superiority on structured, tabular data of moderate size [20,24], and the characteristics of our dataset (21 mixed-type features, 868 records, and no spatial or sequential structure) fall squarely within this regime. Tree ensembles natively handle heterogeneous feature types, require no feature scaling (a practical advantage when combining raw meteorological values with index-derived scores at different scales), and are robust against irrelevant predictors. SVM and ANN are not included as competitive contenders, but as reference points to quantify the ensemble advantage on this specific task, a comparison frequently requested but rarely reported in regional wildfire studies.

3.5. Experimental Protocol

The 868 records were partitioned into 80% training (694 samples) and 20% testing (174 samples) via a nonstratified random split with a fixed seed (random_state = 190) for reproducibility. The decision to use nonstratified rather than stratified splitting is intentional and warrants further explanation. Stratified splitting preserves the exact class distribution in both partitions, which is the standard practice for a balanced evaluation. However, in an operational deployment scenario, the distribution of incoming wildfire events in any given season does not mirror the training distribution; some years have more extreme events, whereas others have fewer. Nonstratified splitting introduces this distributional variability into the test set, providing a more realistic estimate of how the model would perform under genuine deployment conditions.

The prediction is issued at the time of ignition detection, using observations from the nearest station at that moment; it estimates the eventual severity class of a fire that has already started, not the probability of ignition. The lead time is therefore the interval between detection and the realization of the final damage, within which mobilization decisions are made: the median recorded suppression duration rises from 0.3 h for the low class to 6.8 h for the extreme class, and multi-day events extend well beyond it.

Class imbalance was addressed through two mechanisms: class-weighting options included in each model's hyperparameter search where the learner supports them (the selected Random Forest uses balanced-subsample weighting), and SMOTE [36] applied at a uniform 2× factor, which augments every class, including the minority, while preserving class proportions. Oversampling is confined to the training side of every evaluation: it is applied inside each cross-validation fold during model selection and to the training partition alone for the final fit, so no synthetic sample ever derives from a test observation. Section 4.5 reports a sensitivity analysis over oversampling factors from none to 5×, together with a quantification of the inflation that arises when this placement rule is violated. The 2× setting was retained because performance is insensitive to the factor under the correct placement.

Model performance was assessed on the held-out test set using overall accuracy and macro-averaged precision, recall, and F1-score. Macro averaging (computing each metric independently per class and taking the unweighted mean) was chosen over micro or weighted averaging because it treats all four severity levels as equally important, preventing majority-class dominance from masking poor minority-class performance. Per-class recall receives particular attention for the extreme class, reflecting the asymmetric cost structure of wildfire management; failing to anticipate a catastrophic event carries far greater consequences than over-preparing for one that proves moderate. Hyperparameters for every model were tuned by randomized search (25 candidate configurations per model) with five-fold cross-validation on the training set, using macro F1 as the selection criterion; the test set played no role in any selection decision. To characterize split-to-split variability, the selected configuration was refit on 20 additional random 80/20 splits (Section 4.5). Finally, a binary extreme-versus-rest screen was trained on the same features with LightGBM and balanced class weights, tuned by average precision under the same protocol, to support the alert-budget analysis of Section 4.6.

All experiments were implemented in Python (version 3.9), using scikit-learn (version 1.6.1) [37] for GBM, Random Forest, and Extra Trees, the XGBoost (version 2.1.4) [24] and LightGBM (version 4.6.0) [21] libraries, and imbalanced-learn (version 0.12.4) for SMOTE [36]. Missing feature values were imputed with training-set medians inside each pipeline. No feature scaling was applied to tree-based models, consistent with their scale-invariant splitting mechanism; for the SVM and ANN, standard scaling was fit on the training partition and applied to the test partition.

4. Results

This section reports the results in six parts: overall model comparison, ablation study, feature importance analysis, per-class performance, robustness and sensitivity analyses, and screening for extreme-class events, followed by a summary of key findings. All results come from the corrected, leakage-safe protocol described in Section 3.5.

4.1. Overall Model Comparison

Table 2 summarizes test-set performance for the seven candidate models and a Frank and Hall ordinal variant of the recommended model, all trained and selected under the corrected protocol.

Extra Trees attains the best overall scores (accuracy 0.506, macro F1 0.497), followed by Random Forest and LightGBM at an accuracy of 0.466. On macro F1, Random Forest reaches 0.465 with the highest extreme-class recall (0.474), and its per-class F1 spans a narrow band from 0.460 to 0.475, the most balanced profile in the table. Because failing to anticipate an extreme event is the costliest error in this application (Section 3.5), we

recommend the tuned Random Forest as the operational model; the ablation, importance, and per-class analyses that follow use this model.

Table 2. Performance comparison of candidate models. Bold values indicate the best score per column.

Model	Accuracy	Macro F1	CV Acc.	Extreme Recall	Remark
GBM	0.397	0.397	0.434	0.368	
Random Forest	0.466	0.465	0.431	**0.474**	Best extreme recall; recommended
Extra Trees	**0.506**	**0.497**	0.435	0.421	Best accuracy and macro F1
XGBoost	0.443	0.409	0.438	0.211	
LightGBM	0.466	0.440	**0.445**	0.263	
SVM	0.437	0.384	0.422	0.105	
ANN	0.437	0.413	0.431	0.263	
Ordinal (Frank–Hall RF)	0.431	0.413	—	**0.474**	Ordinal reference

The margin between ensemble and non-ensemble methods is narrower than binary-occurrence studies typically report: SVM and ANN trail the best ensemble by 0.08 to 0.11 macro F1 (0.384 and 0.413, respectively). Four-class severity discrimination from meteorological inputs alone leaves limited headroom for any model family, which compresses the differences between families.

Cross-validation accuracies on the training set fall between 0.422 and 0.445 for all models, while the strongest test scores sit somewhat above their CV counterparts; Section 4.5 therefore reports 20 repeated random splits alongside the single-split values, so that neither is read in isolation. In short, Extra Trees is the strongest overall performer, and Random Forest offers the most balanced class profile together with the best extreme-class recall.

4.2. Ablation Study

Table 3 reports the effect of removing feature groups and of replacing the target variable, using the tuned Random Forest under the corrected protocol. ΔF1 is measured against the full 21-feature configuration (macro F1 0.465).

Table 3. Ablation study results. ΔF1 is the change in macro F1 relative to the full configuration.

Configuration	No. Feat.	Accuracy	Macro F1	ΔF1	Remark
Full (proposed, 21 features)	21	0.466	0.465	—	
w/o Effective Humidity	20	0.483	0.480	+0.015	Within split-to-split noise
w/o FWI sub-indices	15	0.431	0.413	−0.052	Largest group effect
Raw meteorological + temporal only	14	0.437	0.408	−0.057	
Original 13-feature set (as submitted)	13	0.466	0.456	−0.009	Submitted feature set
Single target: burned area	21	0.408	0.410	−0.055	Target replaced (see text)

Engineered features. The six FWI-system indices form the largest consistent contribution: removing them lowers macro F1 by 5.2 points, and removing all derived indices (raw meteorology plus temporal terms only) costs 5.7 points. The effective-humidity result requires a candid note. Under the corrected protocol its removal raises macro F1 by 1.5 points, a change within the split-to-split noise band of about ±3 points (Section 4.5), so the large single-feature contribution reported in the original submission does not survive the correction. The expanded 21-feature set improves on the original 13-feature set by a modest 0.9 points.

Composite target. With identical features and model, predicting burned-area classes instead of the CSI lowers macro F1 from 0.465 to 0.410. The composite target is therefore not

only more operationally meaningful but also more predictable from meteorological inputs, although the advantage (5.5 points) is far smaller than the originally reported collapse, which Section 4.5 identifies as an artifact of the flawed oversampling placement. Extreme-tail detection is broadly comparable across the two targets (recall 0.474 versus 0.526 on their respective class definitions), so the case for the CSI rests on overall discrimination and on what the target measures, not on a dramatic performance gap.

4.3. Feature Importance Analysis

Figure 2 ranks the 21 features by permutation importance for the tuned Random Forest, measured as the decrease in test-set macro F1 when a feature is shuffled (mean and standard deviation over 20 permutations).

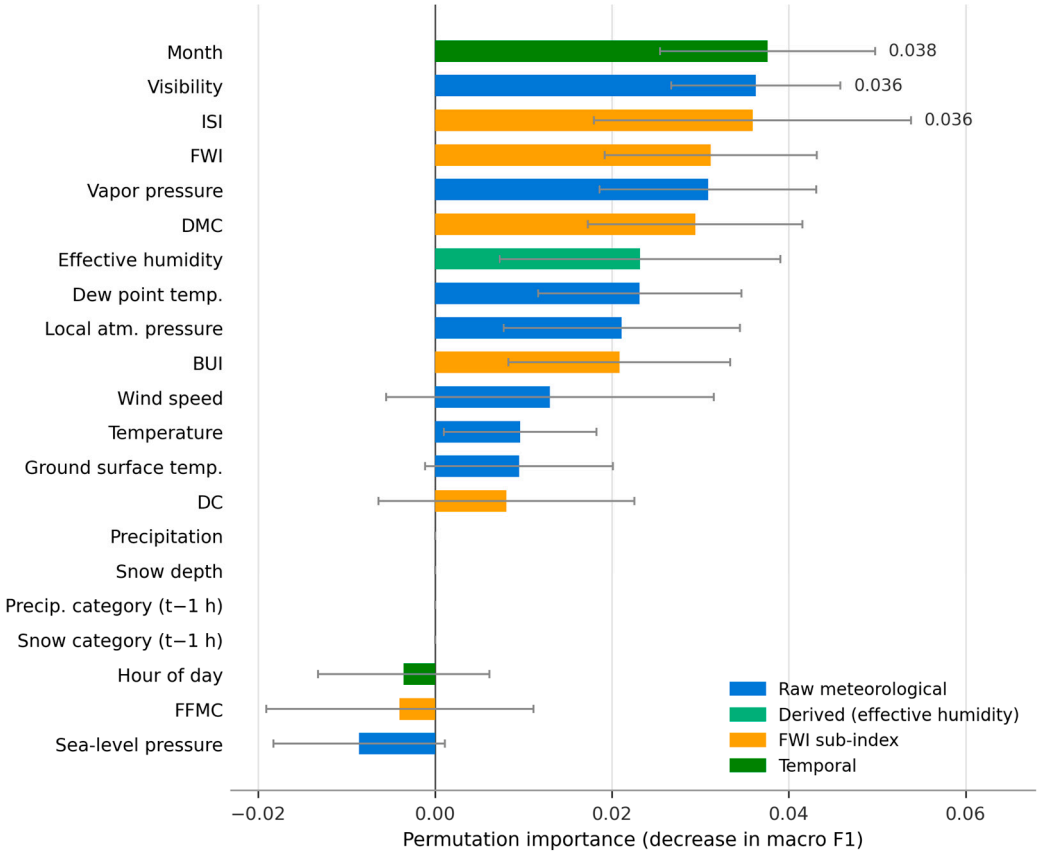

Figure 2. Permutation importance of the 21 features for the tuned Random Forest, measured as the mean decrease in test-set macro F1 over 20 permutations; whiskers show one standard deviation, and slightly negative values indicate features whose shuffling did not degrade performance. Colors denote feature categories.

Month of occurrence leads (0.038), followed by visibility (0.036), ISI (0.036), the composite FWI rating (0.031), and vapor pressure (0.031). The prominence of ISI, FWI, DMC, and BUI is consistent with the ablation result that the FWI family carries the largest group contribution: these indices summarize wind-driven spread potential and the moisture state of progressively deeper fuel layers. Visibility, new in the expanded feature set, plausibly reflects the dry, dust-laden air masses that accompany severe fire weather in the region; we treat this interpretation as tentative.

Effective humidity ranks seventh (0.023 ± 0.016), and several features contribute little or slightly negatively after permutation, which is expected when correlated predictors share information. Together with the ablation table, the importance profile supports one

consistent conclusion: no single feature dominates the corrected model, and the predictive signal is spread across seasonal timing, spread-potential indices, and atmospheric moisture.

4.4. Per-Class Performance

Table 4 lists per-class precision, recall, and F1 for all seven models and the ordinal variant, and Table 5 shows the confusion matrix of the recommended Random Forest.

Table 4. Per-class precision, recall, and F1-score for the seven candidate models and the ordinal variant on the held-out test set (n = 174).

Model	Class	Precision	Recall	F1-Score	Remark
GBM	Low	0.460	0.434	0.447	
	Medium	0.386	0.423	0.404	
	High	0.333	0.340	0.337	
	Extreme	0.438	0.368	0.400	
Random Forest	Low	0.489	0.434	0.460	
	Medium	0.490	0.462	0.475	
	High	0.431	0.500	0.463	
	Extreme	0.450	0.474	0.462	Best Extreme F1
Extra Trees	Low	0.561	0.434	0.489	
	Medium	0.518	0.558	0.537	Best Medium F1
	High	0.459	0.560	0.505	Best High F1
	Extreme	0.500	0.421	0.457	
XGBoost	Low	0.471	0.453	0.462	
	Medium	0.464	0.500	0.481	
	High	0.426	0.460	0.442	
	Extreme	0.308	0.211	0.250	
LightGBM	Low	0.490	0.453	0.471	
	Medium	0.455	0.481	0.467	
	High	0.466	0.540	0.500	
	Extreme	0.417	0.263	0.323	
SVM	Low	0.447	0.396	0.420	
	Medium	0.473	0.500	0.486	
	High	0.403	0.540	0.462	
	Extreme	0.400	0.105	0.167	
ANN	Low	0.579	0.415	0.484	
	Medium	0.417	0.481	0.446	
	High	0.400	0.480	0.436	
	Extreme	0.312	0.263	0.286	
Ordinal (Frank–Hall RF)	Low	0.450	0.679	0.541	Best Low F1
	Medium	0.484	0.288	0.361	
	High	0.395	0.300	0.341	
	Extreme	0.360	0.474	0.409	

Table 5. Confusion matrix of the recommended model (tuned Random Forest with SMOTE 2×) on the held-out test set (n = 174). Rows are true classes and columns are predicted classes; diagonal entries are in bold.

True\Predicted	Low	Medium	High	Extreme
Low	**23**	14	14	2
Medium	13	**24**	13	2
High	9	9	**25**	7
Extreme	2	2	6	**9**

The Random Forest reaches an extreme-class recall of 0.474 (9 of 19 test events) at a precision of 0.450. The confusion matrix adds an operationally relevant pattern: 15 of the 19 extreme events (79%) are predicted High or Extreme, and only 2 are predicted Low. Even when the exact class is missed, severe events are rarely mistaken for minor ones, so a High prediction still activates the provincial readiness tier from which escalation is fast.

The ordinal Frank and Hall variant matches the extreme recall (0.474) but loses overall F1 (0.413 versus 0.465), so ordinal structure alone does not add skill on this dataset.

Averaged over models, the Extreme class is the most difficult (mean F1 0.335), which reflects its 87-sample size rather than a qualitative difference in signal; the three lower classes average between 0.449 and 0.471. The recommended Random Forest keeps all four class-wise F1 scores within 0.460 to 0.475, so its extreme-class performance does not come at the expense of the majority classes.

4.5. Robustness and Sensitivity Analyses

Table 6 addresses two questions about the oversampling step: whether performance depends on the oversampling factor, and what happens when oversampling is misplaced before the train-test split. With the correct placement, macro F1 varies only between 0.459 and 0.482 across factors from none to 5×, so SMOTE is not a material driver of the corrected results. With the flawed placement, the identical pipeline returns 0.647 at factor 2× and 0.830 at factor 5×. This range brackets the 0.696 reported in the original submission and identifies it as a leakage artifact: synthetic copies of test-set neighborhoods enter training, and the model is partly evaluated on information it has already seen. We keep the two flawed-placement rows in the table as a caution, because oversampling before splitting remains a common implementation error in the applied severity-prediction literature.

Table 6. Sensitivity of test performance to the oversampling factor and to the placement of oversampling relative to the train–test split. The flawed placement (oversampling before splitting) is reported only to quantify the leakage artifact that affected the originally submitted results; every other result in this paper uses the correct placement.

Placement	Factor	Accuracy	Macro F1	Extreme Recall
Correct (after split)	none	0.466	0.459	0.421
Correct (after split)	1.5×	0.477	0.481	0.526
Correct (after split)	2×	0.466	0.465	0.474
Correct (after split)	3×	0.483	0.482	0.474
Correct (after split)	5×	0.483	0.479	0.474
Flawed (before split)	2×	0.664	0.647	0.571
Flawed (before split)	5×	0.832	0.830	0.802

Two further checks probe the stability of the reported results. First, refitting the recommended configuration on 20 random 80/20 splits yields macro F1 0.411 $\pm$ 0.032, accuracy 0.429 $\pm$ 0.028, and extreme recall 0.350 $\pm$ 0.112; the single-split values in Table 2 sit in the upper tail of these bands and should be read together with, not in place of, the repeated-split means. Second, re-estimating the PCA underlying the CSI on the two halves of the record (2011 to 2016, n = 410; 2017 to 2022, n = 458) leaves the dominant loading structure essentially unchanged (Tucker congruence 0.997 for PC1 and 0.850 for PC2, with PC1 explained variance of 50.7% versus 64.1%), indicating that the mobilization-scale axis is a stable property of the damage-generating process rather than an artifact of one period.

4.6. Screening for Extreme-Class Events

The four-class formulation treats every misclassification boundary alike, but the decision that matters most in practice is binary: does an incident warrant highest-tier readiness or not. We therefore trained a separate extreme-versus-rest screen on the same weather-only features (LightGBM with balanced class weights, tuned by average precision under the protocol of Section 3.5). On the held-out test set the screen attains an ROC AUC of 0.758 and an average precision of 0.366 against a base rate of 0.109, a 3.4-fold enrichment. Table 7 translates this into alert budgets: flagging the top 20% of incidents captures 9 of the

19 extreme events (recall 0.474). With only 19 positive events in the test set these estimates carry wide uncertainty, but they indicate that these routinely available inputs contain a usable pre-positioning signal for the highest response tier.

Table 7. Screening performance for extreme-class events (extreme vs. rest) with weather-only features on the held-out test set (ROC AUC = 0.758; base rate 10.9%). Each alert budget is the share of incidents flagged for highest-tier readiness.

Alert Budget	Alerts Issued	Extreme Events Captured	Recall
Top 10%	17	6 of 19	0.316
Top 15%	26	8 of 19	0.421
Top 20%	34	9 of 19	0.474

4.7. Summary of Key Findings

Four principal findings emerge under the corrected protocol. First, tree-based ensembles remain the strongest family, with the best models reaching macro F1 of 0.46 to 0.50 against a four-class chance level of 0.25; Extra Trees scores highest overall, and the tuned Random Forest is recommended for operations on the strength of its balanced class profile and extreme recall of 0.474. Second, the composite severity index is a more predictable target than burned area under identical inputs (macro F1 0.465 versus 0.410). Third, the FWI sub-indices provide the largest consistent feature-group contribution (5.2 points), whereas the effective-humidity contribution reported in the original submission is within split-to-split noise. Fourth, a weather-only screen separates extreme from non-extreme events with an ROC AUC of 0.758, capturing 47% of extreme events at a 20% alert budget.

5. Discussion

5.1. Interpretation of Results

These results should be interpreted in light of the inherent difficulty of the task. Unlike binary occurrence prediction, where accuracies above 0.90 are common [6], four-class severity classification demands discrimination among ordinal damage levels using only conditions observable at ignition. Wildfire behavior is stochastic; identical weather can produce different outcomes depending on ignition timing, terrain interaction, and suppression response [15,16]. In this context, a macro F1 of 0.46 to 0.50 against a four-class chance level of 0.25 constitutes a real but modest predictive signal, and the operationally meaningful part of it is concentrated where it matters: extreme-class recall reaches 0.474, and 79% of extreme events are placed at High or above (Section 4.4). Our corrected figures are lower than several results published for adjacent tasks; Section 4.5 illustrates one mechanism by which inflated scores can arise, which argues for caution when comparing across studies.

The comparison with the non-ensemble baselines is more nuanced than in binary-occurrence studies. The strongest ensembles hold a clear margin over the SVM, while the ANN overlaps with the weaker boosting configurations on macro F1, which mirrors the limited headroom of the task noted in Section 4.1. With 694 training samples and 21 mixed-type features, tree-based ensembles remain the sensible default for data of this scale, and the bagging variants (Extra Trees, Random Forest) finish slightly ahead of the boosting variants on this test set.

5.2. Value of the Composite Severity Index

The value of the composite index appears in two forms. Methodologically, with identical features and model, the CSI target is more predictable than burned-area classes (macro F1 0.465 versus 0.410). Burned area depends heavily on terrain, vegetation connectivity,

and suppression effectiveness [4,15], none of which are visible to a weather-only model; by folding suppression resources, duration, and economic damage into the target, the composite index aligns the outcome more closely with the meteorological drivers available as inputs. The advantage is real but moderate, and we no longer claim the dramatic gap reported in the original submission, which Section 4.5 attributes to a preprocessing artifact.

Practically, a composite severity measure better serves emergency management because resource pre-positioning requires estimates of the total suppression burden (personnel, equipment, and time), not only expected burned area [3]. The four classes map onto Korea's graded mobilization protocol as follows. A low prediction corresponds to single-station response under local command. A medium prediction corresponds to a multi-station response within a district, with neighboring stations placed on standby. A high prediction corresponds to provincial-level coordination, deployment of specialized ground crews and helicopters, and precautionary review of evacuation needs. An extreme prediction corresponds to national-level mobilization, inter-provincial mutual aid, and large-scale evacuation orders. The screening analysis of Section 4.6 supports the top of this ladder directly: at an alert budget of 20% of incidents, a weather-only screen captures 47% of extreme events, usable lead information for pre-positioning scarce national assets. This aligns with recent calls for holistic damage metrics that capture socioeconomic impacts alongside physical extent [9,10].

5.3. Role of Engineered Features

Under the corrected protocol, the FWI-system indices carry the largest consistent contribution among engineered features: removing the six indices costs 5.2 macro-F1 points, and permutation importance places ISI, the composite FWI rating, DMC, and BUI in the upper half of all features. This gradient favoring fuel-moisture and spread-potential codes makes physical sense; severe events in Gangwon develop over days to weeks of drying, so the state of intermediate and deep fuel layers is more informative for eventual severity than any instantaneous reading [4,11]. The effective-humidity result is more sobering: its removal changes performance by less than split-to-split noise (Section 4.2). This does not contradict the operational usefulness of effective humidity for occurrence forecasting reported in Korean studies [14,38]; it indicates that, for severity conditional on ignition and in the presence of six FWI indices spanning similar timescales, its marginal information is largely redundant.

A practical property of the feature set is that every variable, raw and derived, can be computed in real time from standard weather-station data and the incident timestamp. This distinguishes the framework from burn-severity approaches that rely on post-fire remote sensing [22,23] and positions it as a tool usable at the moment of ignition detection, before the eventual scale of the event has materialized.

5.4. Comparison with Related Work

Positioning these results within the literature requires attention to task specifications. Binary occurrence studies routinely report accuracies above 0.90 [6], but binary detection is fundamentally easier than multiclass severity discrimination. Among multiclass studies, Jain et al. [3] observed that Random Forest models dominate tabular wildfire prediction tasks, although most reviewed studies address binary occurrence; a stacked-ensemble study of satellite-derived severity mapping [23] likewise highlights the difficulty of ordinal classification. Two lessons from our correction bear on such comparisons. First, weather-only severity prediction has a genuine ceiling, and results far above it deserve scrutiny of the evaluation protocol. Second, oversampling placement alone can move macro F1 by up

to 0.35 (Section 4.5), which exceeds most differences reported between competing models in this literature.

In South Korean wildfire research, most studies have addressed binary occurrence [16,31] or single-target area regression [8]. To the best of our knowledge, this is the first study to propose a PCA-based composite severity index for Korean wildfire data and to validate it through multiclass classification with systematic ablation. The framework's transferable design (standard meteorological inputs, PCA-based target, model-agnostic protocol) makes it applicable to other fire-prone regions with basic weather monitoring infrastructure.

5.5. Limitations and Future Work

This study has several limitations. The dataset covers a single province over 12 fire seasons (868 records), and generalizability to other regions and climates is untested; transfer requires re-estimating both the CSI (on local damage records, using local percentiles) and the classifier. The feature set excludes topographic variables (slope, aspect, elevation), vegetation data (fuel type, canopy density), and human factors (road proximity, land use), which may be needed to sharpen intermediate-class discrimination where meteorological signals are least distinct.

Second, the target is built from administrative records. All five damage fields come from a single national system with a consistent schema across 2011 to 2022, but property damage is recorded in nominal KRW, and gradual changes in reporting practice over twelve years cannot be excluded. The CSI's rank-based construction (log transform followed by percentile boundaries) limits sensitivity to monotone drift in recorded magnitudes; it does not remove such bias, and severity labels inherit whatever error the underlying reports contain. The index was also constructed from post-event records, so the prediction task estimates what the recorded severity will be given conditions at ignition, which is operationally useful for pre-positioning but distinct from real-time severity tracking during an active fire.

Third, the feature-severity relationship may not be stationary under a changing climate. The sub-period analysis of Section 4.5 shows that the structure of the index itself is stable across halves of the record, but relationships between weather and severity can drift as the fire regime intensifies [4], and a 12-season window limits the trends that can be detected. Operational deployment should therefore retrain on a rolling window and monitor calibration. Fourth, misclassification costs are asymmetric: a missed extreme event is far costlier than an over-prepared response. We address this by prioritizing extreme recall and by the alert-budget framing of Section 4.6, but a full cost-sensitive treatment, including ordinal loss functions that penalize distant errors more heavily, remains future work.

Finally, the framework treats each wildfire as independent, ignoring temporal autocorrelation (persistent dry spells) and spatial clustering (adjacent areas sharing weather); sequence-aware architectures may capture these dependencies given larger datasets. Class boundaries (30/30/30/10 percentiles) were chosen to mirror response tiers, and alternative schemes would change the difficulty of the task; the boundary values reported in Section 3.2 make this choice fully reproducible.

6. Conclusions

This study presents a framework for predicting the wildfire damage severity in Gangwon Province, South Korea, using only meteorological observations and calendar terms. Two methodological contributions were introduced: a PCA-based composite severity index integrating five damage dimensions into four ordinal classes, and an engineered feature set

combining raw weather observations with effective humidity and six Canadian Forest Fire Weather Index (FWI) System indices.

An evaluation of 868 wildfire records (2011–2022) under a leakage-safe protocol showed that the strongest tree ensembles reach a macro F1 of 0.46 to 0.50 (recommended configuration: 0.41 ± 0.03 across 20 repeated splits) against a four-class chance level of 0.25. The recommended Random Forest attains an extreme-class recall of 0.474, and a weather-only screen separates extreme from non-extreme events with an ROC AUC of 0.758. The ablation study supported both contributions in moderated form: the composite index is more predictable than a burned-area target under identical inputs (macro F1 0.465 versus 0.410), and the FWI-system indices supply the largest consistent feature-group gain (5.2 percentage points). We also quantify how oversampling misplaced before the train-test split inflates the macro F1 to between 0.65 and 0.83 on this dataset, a caution we document for the severity-prediction literature.

From a practical standpoint, every input feature can be computed in real time from standard weather stations and the incident timestamp, enabling severity assessments, conditional on ignition, to enter routine fire-weather briefings alongside ignition-risk products. The composite target provides emergency managers with a more operationally relevant signal than burned-area prediction alone, thereby supporting the pre-positioning of suppression resources before escalation.

The framework is region agnostic by design. PCA-based target construction requires only standard post-event damage records, and FWI sub-indices have already been computed operationally in dozens of countries. Applying the pipeline to other fire-prone regions requires substituting only local data, with no structural changes to the methodology. Future work should extend the dataset geographically, incorporate topographic and vegetation features, explore temporal models that capture the evolving nature of fire-season conditions, and couple the severity model with ignition-probability models to yield unconditional risk estimates.

Author Contributions: Conceptualization, J.C.; methodology, J.C.; software, W.Y. and S.K.; validation, W.Y. and S.K.; formal analysis, W.Y. and S.K.; investigation, W.Y., S.K. and C.J.; resources, E.S.J.; data curation, W.Y. and S.K.; writing—original draft preparation, A.J. and J.B.; writing—review and editing, J.C., N.K. and C.J.; visualization, W.Y. and S.K.; supervision, J.C.; project administration, J.C.; funding acquisition, E.S.J. All authors have read and agreed to the published version of the manuscript.

Funding: The present Research has been conducted by the Research Grant of Kwangwoon University in 2023.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The wildfire damage records analyzed in this study were obtained from the Fire Safety Big Data Platform operated by the National Fire Agency of Korea (dataset: Wildfire Status Data of Gangwon Province, 2011–2022), and the meteorological observations from the Korea Meteorological Administration (KMA) Automated Surface Observation System (ASOS). The corrected analysis code and complete experimental outputs are available from the corresponding author upon reasonable request.

Acknowledgments: During the preparation of this manuscript, the authors used Claude (Anthropic) for the purposes of language editing and manuscript formatting. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: Author Chumni Jeon is an employee of Korea Telecom. Korea Telecom had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results. The remaining authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Shmuel, A.; Heifetz, E. Global wildfire susceptibility mapping based on machine learning models. *Forests* **2022**, *13*, 1050. [CrossRef]
2. Jolly, W.M.; Cochrane, M.A.; Freeborn, P.H.; Holden, Z.A.; Brown, T.J.; Williamson, G.J.; Bowman, D.M.J.S. Climate-induced variations in global wildfire danger from 1979 to 2013. *Nat. Commun.* **2015**, *6*, 7537. [CrossRef] [PubMed]
3. Jain, P.; Coogan, S.C.P.; Subramanian, S.G.; Crowley, M.; Taylor, S.; Flannigan, M.D. A review of machine learning applications in wildfire science and management. *Environ. Rev.* **2020**, *28*, 478–505. [CrossRef]
4. Chang, D.Y.; Jeong, S.; Park, C.-E.; Park, H.; Shin, J.; Bae, Y.; Park, H.; Park, C.R. Unprecedented wildfires in Korea: Historical evidence of increasing wildfire activity due to climate change. *Agric. For. Meteorol.* **2024**, *348*, 109920. [CrossRef]
5. Ministry of the Interior and Safety. *2019 Gangwon East Coast Wildfire White Paper*; Ministry of the Interior and Safety: Sejong, Republic of Korea, 2019. (In Korean)
6. Prapas, I.; Kondylatos, S.; Papoutsis, I.; Camps-Valls, G.; Ronco, M.; Fernández-Torres, M.-A.; Piles, M.; Carvalhais, N. Deep learning methods for daily wildfire danger forecasting. *arXiv* **2021**, arXiv:2111.02736.
7. Choi, S.; Son, M.; Kim, C.; Kim, B. A forest fire prediction model based on meteorological factors and the multi-model ensemble method. *Forests* **2024**, *15*, 1981. [CrossRef]
8. Cortez, P.; Morais, A. A data mining approach to predict forest fires using meteorological data. In *New Trends in Artificial Intelligence, Proceedings of the 13th Portuguese Conference on Artificial Intelligence (EPIA 2007), Guimarães, Portugal, 3–7 December 2007*; Neves, J., Santos, M.F., Machado, J., Eds.; APPIA: Guimarães, Portugal, 2007; pp. 512–523.
9. San-Miguel-Ayanz, J.; Durrant, T.; Boca, R.; Maianti, P.; Libertà, G.; Artés-Vivancos, T.; Oom, D.; Branco, A.; De Rigo, D.; Ferrari, D.; et al. *Forest Fires in Europe, Middle East and North Africa 2020*; EUR 30862 EN; Publications Office of the European Union: Luxembourg, 2021.
10. Tedim, F.; Leone, V.; Amraoui, M.; Bouillon, C.; Coughlan, M.R.; Delogu, G.M.; Fernandes, P.M.; Ferreira, C.; McCaffrey, S.; McGee, T.K.; et al. Defining extreme wildfire events: Difficulties, challenges, and impacts. *Fire* **2018**, *1*, 9. [CrossRef]
11. Van Wagner, C.E. *Development and Structure of the Canadian Forest Fire Weather Index System*; Forestry Technical Report 35; Canadian Forestry Service: Ottawa, ON, Canada, 1987.
12. San-Miguel-Ayanz, J.; Schulte, E.; Schmuck, G.; Camia, A.; Strobl, P.; Liberta, G.; Giovando, C.; Boca, R.; Sedano, F.; Kempeneers, P.; et al. Comprehensive monitoring of wildfires in Europe: The European Forest Fire Information System (EFFIS). In *Approaches to Managing Disaster—Assessing Hazards, Emergencies and Disaster Impacts*; Tiefenbacher, J., Ed.; InTech: Rijeka, Croatia, 2012. [CrossRef] [PubMed]
13. Di Giuseppe, F.; Pappenberger, F.; Wetterhall, F.; Krzeminski, B.; Camia, A.; Libertà, G.; San Miguel, J. The potential predictability of fire danger provided by numerical weather prediction. *J. Appl. Meteorol. Climatol.* **2016**, *55*, 2469–2491. [CrossRef]
14. Won, M.S.; Lee, M.B.; Lee, W.K.; Yoon, S.H. Prediction of forest fire danger rating over the Korean Peninsula with the digital forecast data and Daily Weather Index. *Korean J. Agric. For. Meteorol.* **2012**, *14*, 1–10. [CrossRef]
15. Kim, J.; Kim, T.; Lee, Y.E.; Im, S. Spatial and temporal variability of forest fires in the Republic of Korea over 1991–2020. *Nat. Hazards* **2025**, *121*, 9801–9821. [CrossRef]
16. Lee, C.; Choi, E.H.; Han, Y.; Lee, Y. Year-round daily wildfire prediction and key factor analysis using machine learning: A case study of Gangwon State, South Korea. *Sci. Rep.* **2025**, *15*, 29910. [CrossRef] [PubMed]
17. Kim, S.Y.; Jun, C.; Na, W. Ensemble-based forecasting of wildfire potentials using relative index in Gangwon Province, South Korea. *Ecol. Inform.* **2025**, *86*, 103021. [CrossRef]
18. Sayad, Y.O.; Mousannif, H.; Al Moatassime, H. Predictive modeling of wildfires: A new dataset and machine learning approach. *Fire Saf. J.* **2019**, *104*, 130–146. [CrossRef]
19. Xu, R.; Lin, H.; Lu, K.; Cao, L.; Liu, Y. A forest fire detection system based on ensemble learning. *Forests* **2021**, *12*, 217. [CrossRef]
20. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
21. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*; Curran Associates, Inc.: Long Beach, CA, USA, 4–9 December 2017; pp. 3149–3157.
22. Collins, L.; Griffioen, P.; Newell, G.; Mellor, A. The utility of Random Forests for wildfire severity mapping. *Remote Sens. Environ.* **2018**, *216*, 374–384. [CrossRef]

23. Nguyen Van, L.; Lee, G. Optimizing stacked ensemble machine learning models for accurate wildfire severity mapping. *Remote Sens.* **2025**, *17*, 854. [CrossRef]
24. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016*; ACM: New York, NY, USA, 2016; pp. 785–794. [CrossRef]
25. Keeley, J.E. Fire intensity, fire severity and burn severity: A brief review and suggested usage. *Int. J. Wildland Fire* **2009**, *18*, 116–126. [CrossRef]
26. Gibson, R.; Danaher, T.; Hehir, W.; Collins, L. A remote sensing approach to mapping fire severity in south-eastern Australia using sentinel 2 and random forest. *Remote Sens. Environ.* **2020**, *240*, 111702. [CrossRef]
27. Cutter, S.L.; Boruff, B.J.; Shirley, W.L. Social vulnerability to environmental hazards. *Soc. Sci. Q.* **2003**, *84*, 242–261. [CrossRef]
28. Greco, S.; Ishizaka, A.; Tasiou, M.; Torrisi, G. On the methodological framework of composite indices: A review of the issues of weighting, aggregation, and robustness. *Soc. Indic. Res.* **2019**, *141*, 61–94. [CrossRef]
29. Sevinc, V.; Kucuk, O.; Goltas, M. A Bayesian network model for prediction and analysis of possible forest fire causes. *For. Ecol. Manag.* **2020**, *457*, 117723. [CrossRef]
30. Costafreda-Aumedes, S.; Comas, C.; Vega-Garcia, C. Human-caused fire occurrence modelling in perspective: A review. *Int. J. Wildland Fire* **2017**, *26*, 983–998. [CrossRef]
31. Lim, C.J.; Chae, H.M. Application of the Canadian Fire Weather Index for forest fire danger assessment in South Korea. *Forests* **2025**, *16*, 1058. [CrossRef]
32. Jeon, G.; Cho, C.; Kim, K.; Choi, W. Deep-learning-based wildfire occurrence prediction with human factors in Gangwon Province, South Korea. *Ecol. Inform.* **2025**, *92*, 103519. [CrossRef]
33. Jolliffe, I.T. *Principal Component Analysis*, 2nd ed.; Springer Series in Statistics; Springer: New York, NY, USA, 2002. [CrossRef]
34. Korea Forest Service. *Practical Manual for Forest Fire Disaster Crisis Response*; Korea Forest Service: Daejeon, Republic of Korea, 2024. (In Korean)
35. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
36. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
37. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
38. Kang, Y.; Jang, E.; Im, J.; Kwon, C.; Kim, S. Developing a new hourly forest fire risk index based on CatBoost in South Korea. *Appl. Sci.* **2020**, *10*, 8213. [CrossRef]