Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

Racing the Fire: Censored Survival Analysis for Wildfire Evacuation Forecasting

Prateeksha Hegde, Omkar Naik, Tridiv Guder, Soumya Ghateppagol
Department of Computer Science and Engineering (Artificial Intelligence)
KLE Technological University, Dr. M. S. Sheshgiri Campus, Belagavi, Karnataka, India

***Abstract*—Wildfires in wildland-urban interface regions have grown more frequent and destructive, placing increasing pressure on emergency responders to make rapid evacuation decisions. This paper presents our approach in the WiDS Datathon 2026, which posed wildfire evacuation-zone threat forecasting as a right-censored survival analysis problem using only the first five hours of perimeter observations for 316 real wildfire incidents (221 training, 95 test). The task required predicting, for four time horizons (12, 24, 48, and 72 hours), the probability that a fire would come within 5 km of an evacuation zone. We propose a hybrid pipeline combining an XGBoost Accelerated Failure Time (AFT) survival model with horizon-specific XGBoost regressors for evaluable horizons, with a fallback Laplace-smoothed base rate when a horizon contains only one class, trained under Inverse Probability of Censoring Weighting (IPCW), along with K-Means clustering, PCA-based dimensionality reduction, isotonic prob-ability calibration, and monotonicity-enforcing post-processing. The final submission blended horizon-specific regressor outputs (90%) with AFT-derived survival probabilities (10%), achieving a hybrid competition score of approximately 0.9598, ranking 192nd out of approximately 1,754 participating teams worldwide – placing the team in the top 12% globally. These results suggest that explicitly accounting for censoring can be beneficial for early-warning wildfire models trained on incomplete early-perimeter data.**

***Index Terms*—wildfire prediction, survival analysis, censored data, XGBoost, accelerated failure time, inverse probability of censoring weighting, isotonic regression, evacuation planning, ensemble learning**

I. INTRODUCTION

Wildfires in wildland-urban interface (WUI) regions have grown more frequent and destructive in recent years, placing increasing pressure on emergency responders to make rapid, well-informed evacuation decisions. A critical operational question during an active fire is not simply *whether* a community is at risk, but *when* that risk is likely to materialize. Emergency managers must sequence evacuation orders across multiple zones under severe time pressure and with limited information. In the earliest hours after ignition, only a handful of fire perimeter observations are typically available, and the eventual trajectory of the fire is highly uncertain.

The WiDS Datathon 2026, organized by Women in Big Data (WiBD) in collaboration with Watch Duty, framed this operational problem as a machine learning challenge: given only the first five hours of perimeter observations for a wildfire, predict the probability that the fire will come within 5 km of a given evacuation zone within 12, 24, 48, and 72 hours of ignition. Because a large fraction of fires in the dataset do not reach an evacuation zone within the 72-hour observation window, the problem is naturally formulated as a right-censored, time-to-event (survival analysis) task rather than a standard classification or regression problem.

Problem Formulation: Each wildfire is observed for only the first 5 hours post-ignition (the observation window). Using these early perimeter observations, the model must predict the probability $P(T \leq H)$ that the fire will come within 5 km of an evacuation zone at four future time horizons: $H \in \{12, 24, 48, 72\}$ hours from ignition. The 72-hour horizon represents the full observation limit; fires not reaching the zone by this time are considered censored.

This paper documents the approach developed for this challenge. Our pipeline combines: (i) dataset-provided wildfire behavior features augmented with unsupervised K-Means clustering and PCA; (ii) an XGBoost AFT survival model; (iii) rizon-specific XGBoost regressors with IPCW; (iv) isotonic regression for probability calibration; and (v) a fi-nal weighted ensemble with monotonicity enforcement. This approach achieved a hybrid evaluation score of approximately 0.9598 and a global rank of 192 out of approximately 1,754 teams (top 12%).

The remainder of this paper is organized as follows. Section II reviews related work. Section III describes the dataset and its censoring structure. Section IV presents the methodology. Section V reports experimental results. Section VI discusses the findings, and Section VII concludes with future directions.

II. RELATED WORK

Wildfire spread prediction has traditionally relied on physics-based simulators such as Rothermel's model and FARSITE [4], [9]. While accurate under well-characterized conditions, these simulators require detailed environmental inputs that are often unavailable in the first hours of an incident, and they are computationally expensive for real-time use. Moreover, they do not naturally handle uncertainty or integrate with observational data.

In recent years, data-driven approaches using satellite- and sensor-derived perimeter observations have gained traction as a complement to physics-based simulation. For instance, Coen et al. [3] coupled weather and fire models in the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

WRF-Fire system, while Papadopoulos et al. [1] provided a comprehensive review of machine learning methods for wildfire spread prediction, highlighting the growing role of ML in operational forecasting. These methods often treat the problem as regression or classification at fixed time steps, ignoring the censored nature of the data.

Survival analysis methods have been widely used in medical and reliability domains to model time-to-event outcomes with censored observations. The Cox proportional hazards model [8] and the Kaplan-Meier estimator [7] are foundational. More recently, tree-based survival models, particularly those using gradient boosting, have become popular. Chen and Guestrin [2] introduced XGBoost. XGBoost supports accelerated failure time (AFT) objectives for censored survival modelling, allowing nonlinear relationships between tabular covariates and the predicted survival-time distribution through gradient-boosted trees.

Inverse Probability of Censoring Weighting (IPCW) is a technique from causal inference [6] that corrects bias due to informative censoring. It has been adapted for converting censored survival problems into a sequence of debiased binary classification problems at fixed time horizons. IPCW assigns higher weights to observations that are less likely to be censored, thereby recovering the true distribution of event times.

Isotonic regression [5] is a standard nonparametric calibration technique for correcting miscalibrated probability outputs from tree-based models. It produces a monotonic mapping from predicted scores to probabilities, which often improves Brier scores and reliability diagrams.

Survival-analysis and censoring-aware techniques have also seen growing use in adjacent environmental and disaster-forecasting settings, where outcomes are frequently right-censored by the observation window [6], [7]. Our approach builds on this line of work, adapting censored survival modelling to the wildfire evacuation-threat setting and combining a global AFT model with discrete-horizon IPCW-corrected classification models – a hybrid strategy that, to our knowledge, has not been applied to short-horizon wildfire evacuation forecasting.

III. DATASET DESCRIPTION

The dataset, provided through the WiDS Datathon 2026 Kaggle competition [10], contains 316 real wildfire incidents (221 training, 95 test), each described using only the first five hours of early perimeter observations after ignition (t_0 to t_0 + 5h). For each incident, 33 numerical features span six categories: temporal coverage, growth, centroid kinematics, distance to evacuation zones, directionality, and temporal metadata. Table I summarises the feature categories.

The prediction target consists of two fields: `time_to_hit_hours` (time from t_0 + 5h until the fire perimeter comes within 5 km of an evacuation zone, bounded at 72 hours) and `event` (binary censoring indicator). Table II reports training set statistics. Two properties make this dataset challenging: (1) small sample size (221 training incidents) relative to dimensionality (33 features), and (2) majority (68.8%) censoring – naive models ignoring this structure will be systematically biased. Additionally, the features are derived from a limited time window, so the signal-to-noise ratio is low.

TABLE I: Feature Categories in the WiDS 2026 Wildfire Dataset

Category	Example Features
Temporal coverage	`num_perimeters_0_5h`, `dt_first_last_0_5h`
Growth	`area_first_ha`, `area_growth_rate_ha_per_h`
Centroid kinematics	`centroid_speed_m_per_h`, `spread_bearing_sin/cos`
Distance	`dist_min_ci_0_5h`, `closing_speed_m_per_h`
Directionality	`alignment_cos`, `alignment_abs`
Temporal metadata	`event_start_hour`, `event_start_month`

TABLE II: Training Set Descriptive Statistics

Statistic	Value
Training incidents	221
Test incidents	95
Events observed (≤72h)	69 (31.2%)
Censored observations	152 (68.8%)
Mean time-to-hit (hours)	37.6 (SD = 25.9)
Raw features	33

IV. PROPOSED METHODOLOGY

The complete pipeline proceeds from raw perimeter data to the final calibrated, monotonic submission. Figure 1 provides an overview of the system architecture, and Algorithm 1 summarises the step-by-step procedure.

A. Preprocessing and Dimensionality Reduction

Missing values were imputed using per-column medians (scikit-learn `SimpleImputer`). The imputer, K-Means clustering, and PCA projection were each fit once on the full training set and then applied identically to the test set, so no information from the held-out test partition was used at any stage. Two derived representations were appended: K-Means cluster assignment ($k = 6$) and 15-component PCA projection. The choice of $k = 6$ and 15 components was based on preliminary experiments and the need to balance information retention with overfitting. The augmented feature matrix X_{aug} concatenates imputed raw features, cluster id, and 15 PCA components. We note that because these preprocessing steps were fit on the full training set rather than refit within each individual cross-validation fold, a small amount of fold-to-fold information sharing is possible; this does not affect the test set but means out-of-fold performance estimates (Section V) should be interpreted as a mild upper bound rather than a fully independent estimate. Only the isotonic calibrators (Section IV-C) were fit strictly on out-of-fold predictions.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

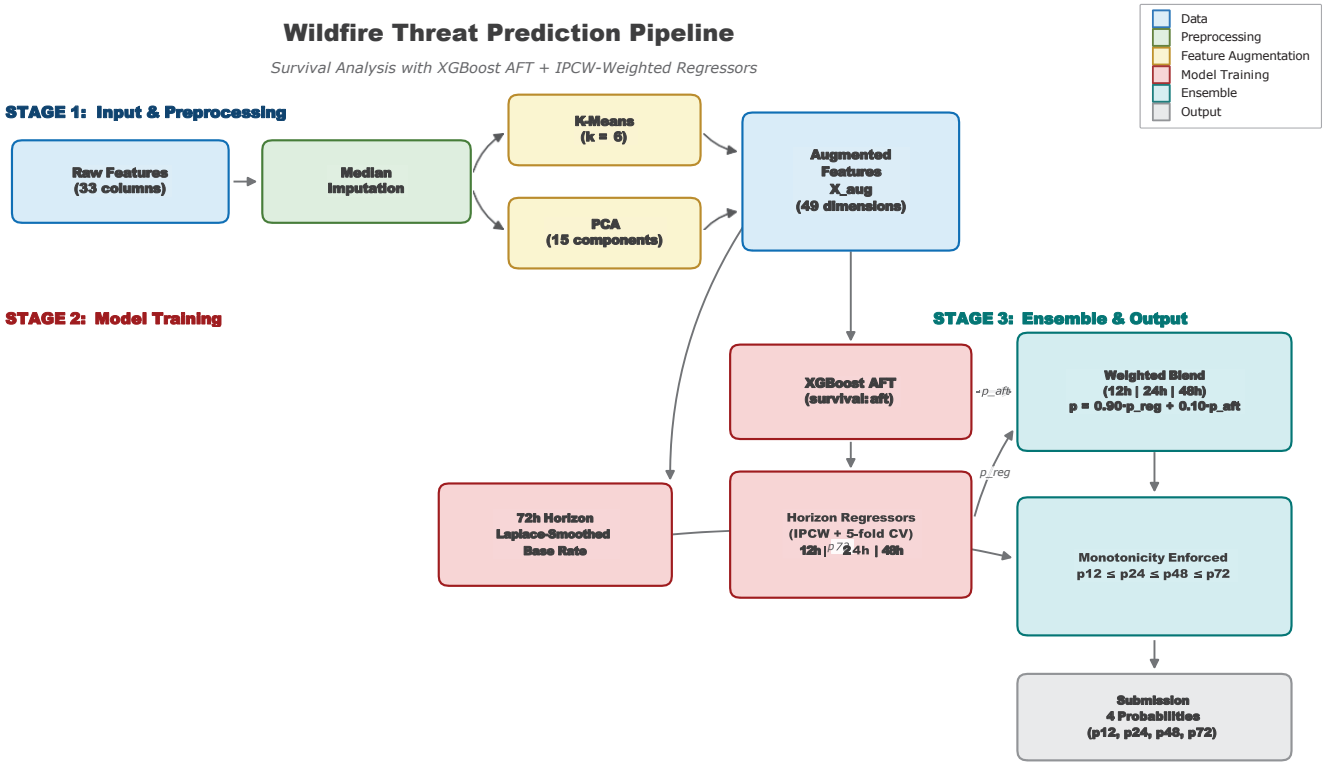

Fig. 1: Complete pipeline architecture. Raw features are imputed, augmented with K-Means (k=6) and PCA (15 components), then processed through (1) an AFT survival model and (2) horizon-specific regressors with IPCW and isotonic calibration. Results for the 12h, 24h, and 48h horizons are blended (90% regressor + 10% AFT), while the 72h horizon uses a Laplace-smoothed base rate; all outputs are subsequently post-processed to enforce monotonicity.

B. Accelerated Failure Time (AFT) Survival Model

An XGBoost model was trained on the censored time-to-event labels using the `survival:aft` objective. For each training fold, the lower bound of the survival interval was set to the observed `time_to_hit_hours`, and the upper bound
was set to infinity for censored observations and to the observed time otherwise. Hyperparameters: 400 boosting rounds, learning rate 0.04, max depth 5, subsample/colsample 0.8/0.7, and 5-fold stratified cross-validation (stratified on whether `time_to_hit_hours` $\leq$ 48h). Predicted AFT times $\hat{y}$ were converted to horizon-specific probabilities using a log-normal survival link:

$$P(T \leq H) = \Phi(\log(H) - \log(\hat{y})) \quad (1)$$

where Φ is the standard normal cumulative distribution function. This link assumes that the log of survival times follows a normal distribution, which is a common and flexible choice for time-to-event data.

C. Horizon-Specific Regressors with IPCW

For each horizon $H \in \{12, 24, 48, 72\}$, a binary threat label was constructed:

$$y_H = \begin{cases} 1 & \text{if event} = 1 \text{ and } t \leq H, \\ 0 & \text{if } t > H, \\ \text{undefined} & \text{otherwise (censored before } H). \end{cases} \quad (2)$$

Observations with undefined labels were excluded from that horizon's training subset, because their true status at time H is unknown. To correct for the informative censoring that remains after exclusion, each retained observation was reweighted using Inverse Probability of Censoring Weighting (IPCW). Weights were derived from a Kaplan-Meier estimate $\hat{G}(t)$ of the censoring distribution, computed once per horizon from the full training set:

$$w_i = \begin{cases} 1/\hat{G}(t_i) & \text{if event at } t_i \leq H, \\ 1/\hat{G}(H) & \text{if censored and } t_i > H, \end{cases} \quad (3)$$

The retained observations were weighted using IPCW, and the nonzero weights were normalized to have unit mean within the horizon-specific training subset.

An XGBoost regressor with `reg:squarederror` objec-tive was then trained per horizon under 5-fold stratified cross-validation, using 400–600 estimators (more for the 48h hori-zon), learning rate 0.04, max depth 5, and the IPCW sample

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

weights. Out-of-fold predictions were used to fit an isotonic regression calibrator per horizon, which was then applied to the corresponding test-set predictions. This calibration step is crucial because the raw regressor outputs are not necessarily well-calibrated probabilities.

At the 72-hour horizon, every retained training observation carries label $y_{72} = 1$ (Section V), so y_H has fewer than two classes and no regressor, isotonic calibrator, or AFT blend (Eq. 4) is fit for this horizon. Instead, a Laplace-smoothed base rate $\hat{p} = (\sum y_H + 1)/(n + 2)$ is assigned identically to every test incident at 72h, consistent with Algorithm 1. This constant value is only subsequently adjusted, if at all, by the monotonicity enforcement step (Eq. 5) against the 48h prediction. Consequently, the 72h probability carries no incident-specific discriminative signal from the regressor or AFT model, unlike the other three horizons.

D. Ensembling and Monotonicity Enforcement

The calibrated regressor output p_H^{reg} was combined with the AFT-derived survival probability p_H^{aft} (Eq. 1) in a fixed weighted blend, applied at the 12h, 24h, and 48h horizons (see Section IV-C for the 72h exception):

$$p_H = 0.90\, p_H^{\text{reg}} + 0.10\, p_H^{\text{aft}}. \tag{4}$$

The 90:10 ratio was set manually rather than selected through a systematic weight search: we reasoned that the horizon-specific regressor, being trained directly on binary labels for that horizon, should dominate over the AFT model's indirect, survival-time-derived probability, and fixed the weighting accordingly. No cross-validated sweep over alternative blend ratios was performed, so the sensitivity of results to this choice is untested; we identify a learned or tuned blend weight as future work (Section VII).

Because the four horizons are nested in time, the probability outputs are constrained to be non-decreasing:

$$p_H \leftarrow \max(p_H, p_{H-1}), \qquad H \in \{24, 48, 72\}. \tag{5}$$

This enforces the logical condition that if a fire has reached the zone by 12h, it has certainly reached by 24h, etc. Without this post-processing, independent predictions might violate monotonicity due to noise. Final probabilities were clipped to $[0, 1]$ before being written to the submission file.

E. Hyperparameter Tuning

All hyperparameters were fixed manually based on preliminary, non-nested experimentation on the training set, rather than through an inner cross-validation loop for automated model selection. The number of boosting rounds (400 for AFT, 400-600 for regressors), learning rate (0.04), tree depth (5), and subsample ratios were chosen to balance performance and overfitting. The choice of 5-fold CV was a practical compromise given the small dataset; more folds would reduce training size further, while fewer folds would increase variance. Because hyperparameter selection was not embedded in a nested (nested + inner-loop) validation procedure, the out-of-fold metrics reported in Section V should be read as evaluation of a fixed, manually-tuned configuration rather than as an unbiased estimate of a fully automated model-selection pipeline.

Algorithm 1 Time-to-Threat Prediction Procedure

Training features X_{tr} with censored labels $(t, e)_{tr}$, test features X_{te} Monotonic threat probabilities $\{p_{12}, p_{24}, p_{48}, p_{72}\}$

Step 1: Impute missing values using per-column medians fitted on X_{tr}; apply to X_{te}.

Step 2: Fit K-Means ($k = 6$) and PCA (15 components) on X_{tr}; append cluster id and PCA components to form X_{aug} (and $X_{\text{te,aug}}$).

Step 3: Train the XGBoost AFT model (`survival:aft`) under 5-fold CV on X_{aug} with censored labels $(t, e)_{tr}$; convert predicted survival times to p^{aft} via Eq. (1) for each horizon H.

Step 4: for each horizon $H \in \{12, 24, 48, 72\}$ **do**

Step 4a: Construct binary label y_H via Eq. (2); drop rows with undefined labels.

Step 4b: if y_H has fewer than 2 classes **then**

Step 4b(i): Assign the Laplace-smoothed base rate $\hat{p} = (\sum y_H + 1)/(n + 2)$ to every test row for this H; skip Steps 4c–4f (no regressor, calibration, or AFT blend).

else

Step 4c: Compute IPCW weights w via Eq. (3); normalize nonzero weights to unit mean.

Step 4d: Train an XGBoost regressor (`reg:squarederror`) with weights w under 5-fold CV; collect out-of-fold predictions and test predictions.

Step 4e: Fit an isotonic regression calibrator on the out-of-fold predictions; apply it to the test predictions to obtain p_H^{reg}.

Step 4f: Blend $p_H \leftarrow 0.90\, p_H^{\text{reg}} + 0.10\, p_H^{\text{aft}}$ (Eq. (4)).

end if

end for

Step 5: Enforce monotonicity: $p_H \leftarrow \max(p_H, p_{H-1})$ for $H \in \{24, 48, 72\}$ (Eq. (5)).

Step 6: Clip all p_H to $[0, 1]$ and write the final submission file.

V. EXPERIMENTS AND RESULTS

All experiments were run on a single workstation with an Intel Core Ultra 9 185H processor (5.1 GHz max turbo) and 32 GB RAM, with an NVIDIA GeForce RTX 4060 (8 GB VRAM) for GPU-accelerated stages. The pipeline was implemented in Python using XGBoost, scikit-learn, SciPy, and lifelines, with a fixed random seed (42) for reproducibility. The final submission was evaluated by the competition organisers using a hybrid metric combining probabilistic ac-curacy (Brier-score-style scoring across the four horizons) and rank-based concordance. Our pipeline achieved a hybrid score of approximately **0.9598**, corresponding to a global rank of

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

TABLE III: Final Leaderboard Result

Metric	Value
Hybrid competition score	≈ 0.9598
Global rank	192
Total participating teams	≈ 1, 754
Percentile	Top 12%

TABLE IV: Out-of-Fold Model Comparison (Mean Over 12h/24h/48h)

Model	Acc.	F1	AUC	Brier
AFT-only	0.944	0.904	0.989	0.041
XGB-raw	0.939	0.895	0.977	0.048
XGB-calibrated	0.950	0.903	0.983	0.038
Ensemble (proposed)	**0.950**	**0.908**	**0.991**	**0.036**

192 out of approximately 1,754 participating teams – placing the team in the **top 12%** worldwide (Table III).

To evaluate the contribution of each pipeline component, we compared four model variants under the same 5-fold stratified cross-validation protocol with out-of-fold evaluation: AFT-only (Eq. 1), XGB-raw (uncalibrated horizon regressor), XGB-calibrated (with isotonic calibration), and Ensemble (the full blended pipeline). Table IV reports out-of-fold Accuracy, weighted F1, ROC-AUC, and Brier score for the 12h, 24h, and 48h horizons. The 72h horizon is omitted because every valid training incident has label $y_{72} = 1$ (no negatives), making discriminative metrics undefined – this is a structural property of the dataset, and as described in Section IV-C, the 72h submission is generated via a constant base rate rather than a trained model.

Figure 2 shows (a) the Accuracy and F1 bar charts, and (b) the ROC curves for the Ensemble model across 12h, 24h, and 48h. The Ensemble achieves strong discrimination across all evaluable horizons, with AUC increasing monotonically from 0.983 at 12h to 0.993 at 24h and 0.996 at 48h – consistent with more elapsed perimeter history providing a stronger fire-growth signal at later horizons. The improvement over XGB-raw confirms that calibration and blending are beneficial.

Figure 3 shows the confusion matrices. The 12-hour horizon, with the fewest positive examples (49 of 215, 22.8%) and the least perimeter history, has the highest miss rate (10 false negatives), while 24h and 48h show very few errors (1 and 0 false negatives, respectively). This trend is intuitive: with more elapsed time, the fire growth signal becomes stronger, making classification easier – consistent with the monotonically increasing AUC noted above.

A. Ablation Study

To quantify the contribution of K-Means and PCA, we conducted an ablation under the same 5-fold stratified cross-validation protocol using the XGB-calibrated regressor with four feature configurations: raw only, + K-Means, + PCA, and both (proposed). Table V reports the results. PCA provides the primary benefit, improving mean AUC from 0.974 to 0.983 and reducing Brier from 0.040 to 0.036. K-Means alone offers minimal improvement (AUC 0.975, Brier 0.042). The full combination achieves the best Brier score (0.034) and ties for best AUC (0.983). This indicates that PCA is effective at compressing correlated features, while K-Means adds only a weak regularizing effect.

TABLE V: Ablation Study (Mean Over 12h/24h/48h)

Feature configuration	Acc.	F1	AUC	Brier
Raw only	0.949	**0.909**	0.974	0.040
+ K-Means cluster id	0.946	0.902	0.975	0.042
+ PCA(15) only	**0.950**	0.906	**0.983**	0.036
+ K-Means + PCA (proposed)	**0.950**	0.903	**0.983**	**0.034**

VI. DISCUSSION

The experimental results demonstrate that the proposed ensemble consistently outperforms individual components. Several design choices contributed to this success:

1) **Explicit modelling of censoring** avoided systematic bias for the 68.8% of training incidents whose true time-to-threat was unobserved. By treating censoring as informative and using IPCW, we recovered a more accurate training signal.
2) **IPCW reweighting** allowed each horizon-specific classifier to be trained on a properly debiased subset. This proved empirically more stable than training a single shared model across all four horizons, because each horizon has a different censoring pattern.
3) **Isotonic calibration** reduced the Brier score at every evaluable horizon relative to the raw regressor (Table IV), confirming it as an effective safeguard against over-confident predictions given the small training set (221 incidents, only 69 observed events).
4) **Monotonicity enforcement** acted as an implicit information-sharing mechanism between horizons, improving the 72-hour predictions in particular, where the number of confidently labelled positive examples was smallest. This post-processing step is simple yet powerful.

We acknowledge several limitations. First, the reported hyperparameters were fixed manually rather than selected through a nested cross-validation procedure with an inner tuning loop; the out-of-fold metrics in Section V therefore evaluate this fixed configuration rather than provide an unbiased estimate of a fully automated pipeline, and some risk of informal overfitting to the dataset through manual iteration cannot be ruled out. Second, the imputer, K-Means, PCA, and IPCW weighting were each fit once on the full training partition rather than refit inside every individual cross-validation fold; while this introduces no test-set leakage, it means the out-of-fold performance estimates in Table IV and Table V may be mildly optimistic due to fold-to-fold information sharing in these shared preprocessing steps. Third, the 72-hour horizon has no observed negative examples among valid training

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

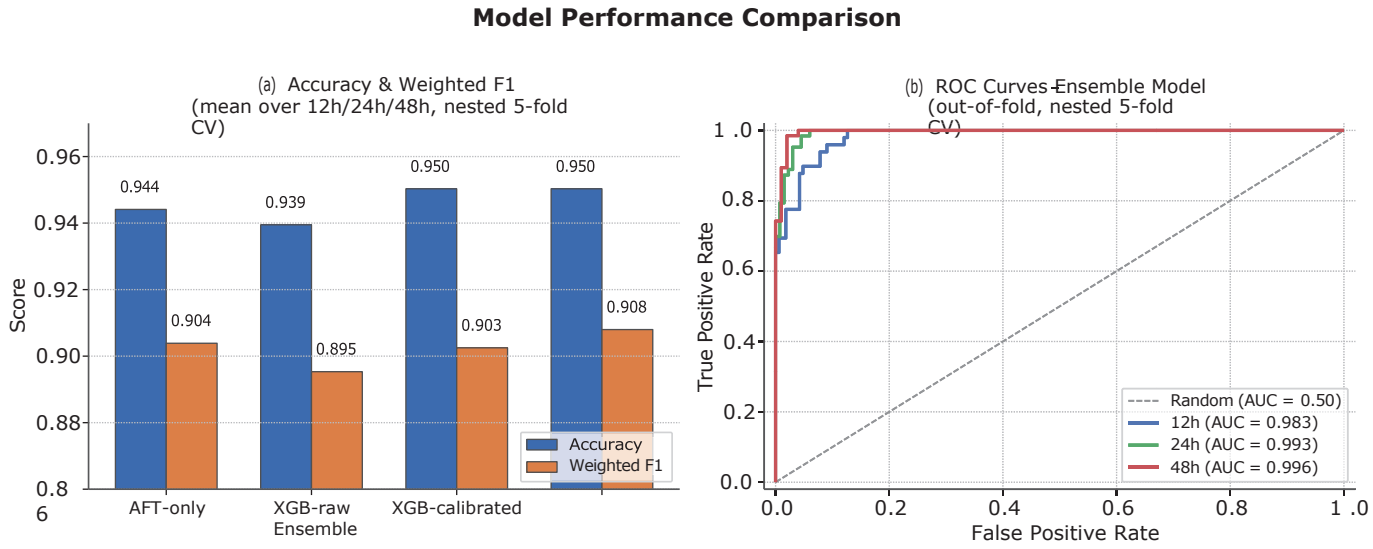

Fig. 2: (a) Out-of-fold Accuracy and weighted F1 for all model variants; (b) ROC curves for the proposed Ensemble model at each evaluable horizon. The 72h horizon is omitted because discriminative metrics (AUC, F1) are undefined due to the absence of negative examples – a structural property of the dataset.

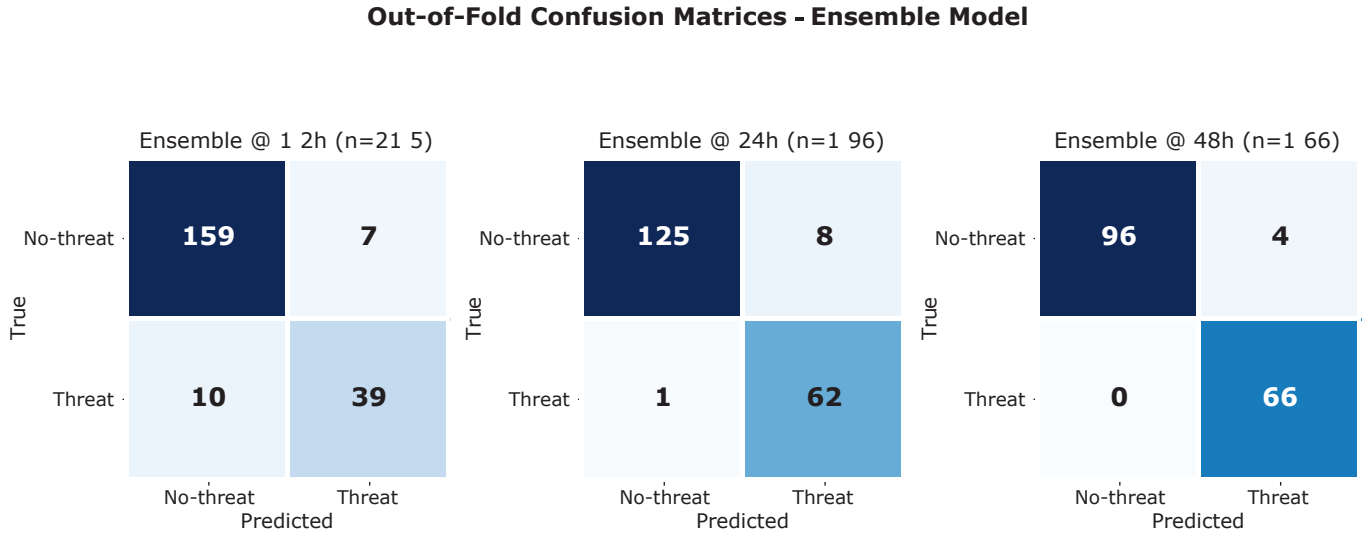

Fig. 3: Out-of-fold confusion matrices for the proposed Ensemble model at the 12h, 24h, and 48h horizons. The 72h horizon is omitted because discriminative evaluation is undefined (no negative examples). The 12-hour horizon shows the highest miss rate (10 false negatives), while 24h and 48h show very few errors.

incidents, so its prediction reduces to a constant Laplace-smoothed base rate with no learned regressor, isotonic calibration, or AFT contribution; its accuracy depends entirely on the monotonicity constraint against the 48h prediction, making it substantially weaker and less informative than the other three horizons. Fourth, the 90:10 regressor/AFT ensemble weight (Eq. 4) was set manually rather than tuned via a systematic search, so its optimality and sensitivity remain unverified. Fifth, PCA and K-Means showed limited standalone signal; their main contribution appeared to be modest regularization rather than a strong predictive signal. Sixth, the AFT model's log-normal survival link (Eq. 1) is a simplifying parametric assumption; a more flexible distribution (e.g., log-logistic) or a fully non-parametric survival forest may better capture the true event-time distribution, especially given the heavily right-skewed censoring pattern.

A further limitation is that the dataset is geographically limited to 316 incidents, all from the same region. The model's generalizability to other wildfire regimes remains untested. Additionally, the features are derived solely from perimeter data; incorporating weather, fuel, and terrain variables could further improve performance.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

VII. CONCLUSION AND FUTURE WORK

This paper presented a hybrid survival-analysis pipeline for predicting the probability of wildfire threat to evacuation zones across multiple time horizons, developed for the WiDS Datathon 2026 challenge. By combining an XGBoost AFT survival model, horizon-specific IPCW-weighted regressors, isotonic calibration, and a monotonicity-enforcing ensemble, the approach achieved a hybrid evaluation score of approximately 0.9598 and a global rank of 192 out of roughly 1,754 teams, placing the team in the top 12% worldwide. The results suggest that explicitly modelling censoring – rather than discarding or naively labelling censored observations – is an important design consideration for early-warning wildfire models trained on incomplete early-perimeter data.

Future work could extend this approach in several directions:

- **Nested hyperparameter tuning:** Embedding hyperparameter selection within an inner cross-validation loop, and refitting preprocessing steps (imputation, PCA, K-Means, IPCW weighting) strictly within each outer fold, would provide a fully unbiased estimate of pipeline performance.
- **Incorporating exogenous covariates:** Weather (wind speed, humidity, temperature) and fuel-moisture data could complement the perimeter-derived kinematic features, potentially improving long-horizon predictions.
- **Learned ensemble weights:** The fixed 90:10 blend could be replaced with a meta-model, or a cross-validated search over blend weights, that learns the optimal com-bination per incident or per horizon.
- **Non-parametric survival models:** Random survival forests or DeepSurv-style neural architectures may better handle the non-linearities and censoring patterns without strong parametric assumptions.
- **Larger-scale validation:** Testing the pipeline on a larger, more geographically diverse set of wildfire incidents would assess generalizability and robustness.

ACKNOWLEDGMENT

The authors thank Women in Big Data (WiBD) and Watch Duty for organising the WiDS Datathon 2026 and providing the wildfire perimeter dataset, and KLE Technological Univer-sity for institutional support.

REFERENCES

[1] A. Papadopoulos, G. Petropoulos, and D. Kalivas, “Machine learning approaches for wildfire spread prediction: A comprehensive review,” *Remote Sens.*, vol. 16, no. 3, p. 482, 2024, doi: 10.3390/rs16030482.

[2] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[3] J. L. Coen, M. Cameron, J. Michalakes, E. G. Patton, P. J. Riggan, and K. M. Yedinak, “WRF-Fire: Coupled weather-wildland fire modeling with the Weather Research and Forecasting model,” *J. Appl. Meteorol. Climatol.*, vol. 52, no. 1, pp. 16–38, 2013, doi: 10.1175/JAMC-D-12-023.1.

[4] M. A. Finney, “FARSITE: Fire Area Simulator – model development and evaluation,” USDA Forest Service, Res. Paper RMRS-RP-4, 1998, doi: 10.2737/RMRS-RP-4.

[5] T. Robertson, F. T. Wright, and R. L. Dykstra, *Order Restricted Statistical Inference*. New York: Wiley, 1988, doi: 10.1002/9780470316702.

[6] M. J. van der Laan and J. M. Robins, *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer, 2003, doi: 10.1007/978-0-387-21700-0.

[7] E. L. Kaplan and P. Meier, “Nonparametric estimation from incomplete observations,” *J. Am. Stat. Assoc.*, vol. 53, no. 282, pp. 457–481, 1958, doi: 10.1080/01621459.1958.10501452.

[8] D. R. Cox, “Regression models and life-tables,” *J. R. Stat. Soc. Ser. B*, vol. 34, no. 2, pp. 187–220, 1972, doi: 10.1111/j.2517-6161.1972.tb00899.x.

[9] R. C. Rothermel, “A mathematical model for predicting fire spread in wildland fuels,” USDA Forest Service, Res. Paper INT-115, 1972, doi: 10.2737/INT-RP-115.

[10] Women in Big Data and Watch Duty, “WiDS Global Datathon 2026: Predicting Time-to-Threat for Evacuation Zones Using Survival Analysis,” Kaggle, 2026. [Online]. Available: https://www.kaggle.com/competitions/WiDSWorldWide_GlobalDatathon26