Predicting Failure Modes of RC Beams Under Fire Using Machine Learning

The University of Sydney
School of Computer Science

Capstone Project – 2025

Supervisor: Dr. Faham Tahmasebinia

Name	SID	UniKey
Guanchen Pan	530301202	gpan0427
Zihan Liu	500176757	zliu7315
Pranav Anand	530606093	pana0377
Tran Duc Manh	510000079	dtra3691
Zhiyuan Na	530002077	zhna4143
Shengyuan Nie	530211565	snie9428

Abstract

FRP-strengthened RC beams are vulnerable under fire due to adhesive softening, interfacial debonding, and thermal gradients that couple serviceability and ultimate limit states. While prior experimental and numerical work has clarified mechanisms [1, 2, 3], and recent ML efforts predict fire-resistance time (FRT) [6, 5], there remains a gap in *mode-level* prediction with interpretable, safety-aware outputs for design decisions.

We develop an interpretable hybrid ML framework that jointly predicts **failure mode** (*Deflection Failure*, *Strength Failure*, *No Failure*) and **FRT** from raw physical features. The classifier blends **XGBoost + Targeted SMOTE** with an **LDAM–DRW XGBoost** leg trained with Borderline-SMOTE; probabilities are **isotonic-calibrated** and **per-class thresholded** to enforce a **safety-first** bias. On a held-out split from the Bhatt–Kodur–Naser dataset [6], we obtain **Accuracy** $\approx$ **0.906**, **Balanced Acc.** $\approx$ **0.824**, **Macro-F1** $\approx$ **0.818**. The FRT regressor achieves **MAE** $\approx$ **10.3 min**, $R^2 \approx 0.93$. SHAP analyses highlight insulation thickness/depth, concrete cover, FRP area, and load ratio as dominant drivers, consistent with fire mechanics and Eurocode EN 1992–1–2 framing.

Crucially, because *No Failure* is rare ($\sim$2–3%), interpretability provides *directional risk reduction* rather than a prescriptive “design-to-flip” guarantee. The tool supports engineering judgment; it does not replace it. Compared with prior approaches, our contribution is a **reproducible, calibrated, and interpretable** pipeline that integrates performance prediction with **design-oriented insight**.

1. Introduction

Fire-exposed RC beams strengthened with externally bonded FRP present coupled thermal–mechanical challenges: adhesive rheology shifts near T_g, bond degrades, reinforcement softens, and large deflections interact with strength loss. Eurocode EN 1992–1–2 provides system-level guidance but cannot fully capture project-specific geometries, anchorage schemes, and insulation details. Experimental programs [1, 2] show that failure pathway (deflection-governed vs strength-governed) depends sensitively on thermal protection and demand; high-fidelity numerical models [3, 4] offer insight but are resource-intensive and sensitive to uncertain inputs.

Recent ML studies predict FRT from curated datasets [5, 6], yet most *do not* classify failure *mode* nor provide calibrated probabilities and interpretable guidance that align with safety codes. We address this gap with a hybrid, safety-calibrated classifier plus a companion FRT regressor, both tied to SHAP-based explanations and an interactive dashboard.

2. Related Work and Comparative Analysis

2.1. Experimental and Numerical Baselines

Ahmed & Kodur [1] report that FRP debonding and steel softening dominate under elevated temperatures; the necessity of protecting anchorage and adhesive is underscored by Firmo et al. [2], who show that bond failure can be critical unless anchorage zones are insulated. Thermal–mechanical FE models (e.g., Zhang et al. [3]; Dong et al. [4]) reproduce temperature profiles and stiffness degradation but require extensive material characterisation and computational effort.

2.2. Data-Driven Approaches

Bhatt et al. [5] explored ML for FRT regression on a curated FRP-RC fire dataset; Bhatt [6] released a larger, documented dataset aggregating simulated and experimental cases. These works focus on predicting *duration*, not *mode*, and generally do not incorporate probability calibration, per-class safety thresholds, or SHAP-driven design guidance.

2.3. How Our Approach Differs

Our contribution relative to literature:

- **Target**: we classify *failure mode and* predict *FRT*, enabling both categorical risk assessment and time-based estimation (most prior ML focuses on FRT only).
- **Reliability**: we apply *per-class isotonic calibration* and *threshold tuning* that bias decisions toward safety, consistent with engineering duty-of-care (rare in prior ML work).
- **Interpretability**: we map SHAP attributions to *design levers* (e.g., insulation thickness/depth, cover), closing the loop between prediction and action.
- **Reproducibility**: config-driven pipeline; leak-safe validation; saved artefacts (figures/JSONs/models) and one-command runner.

Table 1: Positioning against representative literature (qualitative).

Study	**Approach**	**Target**	**Dataset Scope**	**Limitations vs Ours**
Ahmed & Kodur (2011) [1]	Experiments + analysis	Mechanisms/limits	Small sets	No rapid predictor; no ML interpretability/calibration
Firmo et al. (2012) [2]	Experiments (bond/anchorage)	Mechanisms	Focused	No predictive model; anchorage insights not generalised
Zhang et al. (2018) [3]	FE thermo-mechanical	Response histories	Case-based	High set-up cost; sensitive to material assumptions
Bhatt et al. (2021) [5]	ML regression	FRT	Moderate	No mode classification; limited interpretability
Bhatt (2023) [6]	Dataset + ML demos	FRT	Large curated	Primarily duration-focused; safety calibration not explicit
This work	**Hybrid ML + SHAP + calibration**	**Mode + FRT**	**∼21k+ curated**	**Adds safety thresholds, mode classification, design guidance**

3. Data and Pre-processing

3.1. Source Dataset

We use the open FRP-strengthened RC beam fire dataset compiled by Bhatt, Kodur, and Naser[1] [6]. It aggregates ∼21,384 numerically generated cases plus 49 experimental cases with geometry, materials, insulation, loading, and fire-response fields, validated by thermo–mechanical modelling in prior work [5].

3.2. Variables

Table 2 lists the raw physical variables ingested by our pipeline.

[1]Bhatt, P. (2023). *Fire resistance of FRP-strengthened beams* (Version 6) [Dataset]. Mendeley Data. doi:10.17632/3c2szhbdn5.6

Table 2: Variables used in modelling.

Symbol	Description
BN	Beam identifier
L	Span length (m)
Ac	Concrete area (mm^2)
Cc	Concrete cover (mm)
As	Steel area (mm^2)
Af	FRP area (mm^2)
t_{ins}	Insulation thickness at midspan (mm)
h_i	Insulation depth on sides (mm)
f_c	Concrete compressive strength (MPa)
f_y	Steel yield strength (MPa)
E_s	Steel modulus (MPa)
f_u	FRP tensile strength (MPa)
E_{frp}	FRP modulus (MPa)
T_g	Polymer glass transition temperature (°C)
k_{ins}	Insulation thermal conductivity (W/mK)
$\rho_{ins}c_{ins}$	Insulation volumetric heat capacity (J/°C m^3)
L_d	Applied load (kN)
LR	Load ratio (%)
F/EF	Failure or end of fire exposure (0/1)
FR	Fire resistance time (min)
δ_f	Midspan deflection at failure (mm)

3.3. Label formation

We standardise `LimitState` to: *Deflection Failure*, *Strength Failure*, *No Failure*. Unlabelled rows are dropped. A copy is stored as `target` for training scripts.

3.4. Cleaning, collinearity control, and features

We perform a single pass: strip headers; safe renames (`F/EF`→`F_to_EF`); coerce numerics; drop unlabeled. Feature set (classifier & regressor) uses **raw physical** inputs only:

$$\{L, Ac, Cc, As, Af, t_{ins}, h_i, f_c, f_y, E_s, f_u, E_{frp}, T_g, k_{ins}, \rho_{ins}c_{ins}, L_d, LR\}.$$

We prune collinearity by removing any column with $|r| > 0.97$ against a previously retained feature (upper-triangle scan). We **exclude engineered ratios** in production runs (improved minority F1 and reduced multicollinearity).

3.5. Imbalance and splits

The dataset is long-tailed: *Deflection Failure* ∼86.3%, *Strength Failure* ∼11.4%, *No Failure* ∼2.3% in the held-out split. We use an 80/20 stratified split (a grouped CV helper exists). Resampling is *train-only*.

3.6. Data integrity and leakage controls

All scalers/resamplers fit on train only; thresholds tuned on validation predictions; BN used for traceability only; FRT regressor trained on rows with valid FR labels; no target/FR leakage into classifier features.

4. Model Architecture and Methodology

4.1. Overview

We combine two complementary legs:

- **Leg A: XGBoost + Targeted SMOTE** — robust baseline on majority behaviour.
- **Leg B: LDAM–DRW XGBoost + Borderline-SMOTE** — emphasises minority modes near decision boundaries.

Probabilities are **soft-blended** and **isotonic-calibrated**; **per-class thresholds** are tuned to maximise macro-F1 while biasing away from unsafe false negatives (**safety-first**).

4.2. Leg A: XGBoost (Baseline Learning)

Gradient-boosted trees capture nonlinear interactions among t_{ins}, h_i, Cc, LR, A_f, etc. **Targeted SMOTE** up-samples *Strength Failure* to ∼100% of majority and *No Failure* to ∼70% (train-only). Core hyperparameters: `n_estimators=650`, `learning_rate=0.06`, `max_depth=6`, `subsample=0.9`, `colsample_bytree=0.9`, `min_child_weight=10`, `tree_method='hist'`, `random_state=42`.

4.3. Leg B: LDAM–DRW XGBoost (Long-tailed Learning)

We emulate LDAM–DRW behaviour via a margin-aware, class-reweighted objective. Inputs are z-scored; **Borderline-SMOTE** targets ambiguous regions. Hyperparameters:

`eta=0.05`, `max_depth=5`, `subsample=0.9`, `colsample_bytree=0.9`, `lambda=1.0`, trained in two phases (200+300 rounds).

4.4. Fusion, calibration, and thresholds

We blend distributions with weight $\alpha \in [0.50, 0.90]$ (best $\sim$0.7). We apply **per-class isotonic calibration** on validation scores, then grid-search **class thresholds** (0.25–0.60) to maximise macro-F1 subject to *safety bias* (tolerate more false positives in at-risk classes to reduce false negatives).

4.5. Deployment

We serialise the calibrated ensemble and regressor (joblib), plus SHAP metadata, into a model pack for a Streamlit dashboard. The UI supports dataset mode (ground-truth comparison) and manual mode (design sandbox), and provides Eurocode-style exposure+margin framing.

5. Results and Evaluation

5.1. Failure-Mode Classification

On the held-out validation split:

$$\text{Accuracy} \approx 0.906, \qquad \text{Balanced Acc.} \approx 0.824, \qquad \text{Macro-F1} \approx 0.818.$$

Table 3: Per-class metrics on the validation fold.

Class	**Precision**	**Recall**	**F1-score**
Deflection Failure	0.947	0.945	0.946
No Failure	0.869	0.887	0.878
Strength Failure	0.613	0.619	0.616
Overall accuracy		0.907	
Macro average F1		0.813	
Balanced accuracy		0.819	

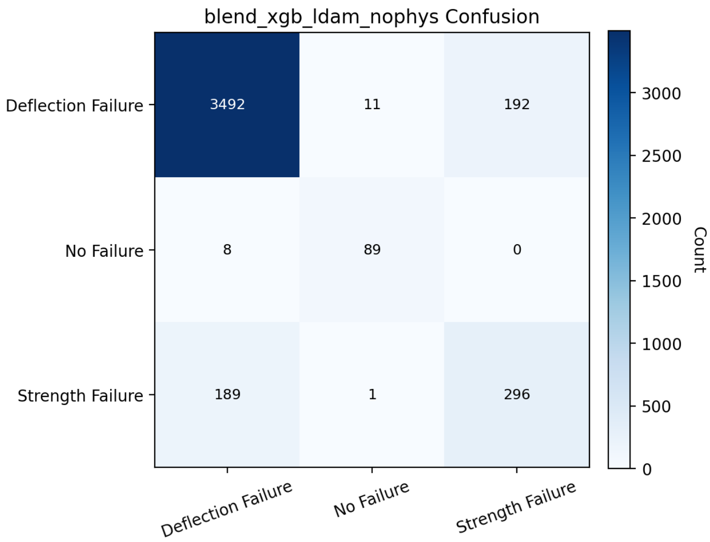

Figure 1: Confusion matrix for the blended, calibrated classifier (held-out validation). Class order follows the model pack.

5.2. Precision–Recall and ROC

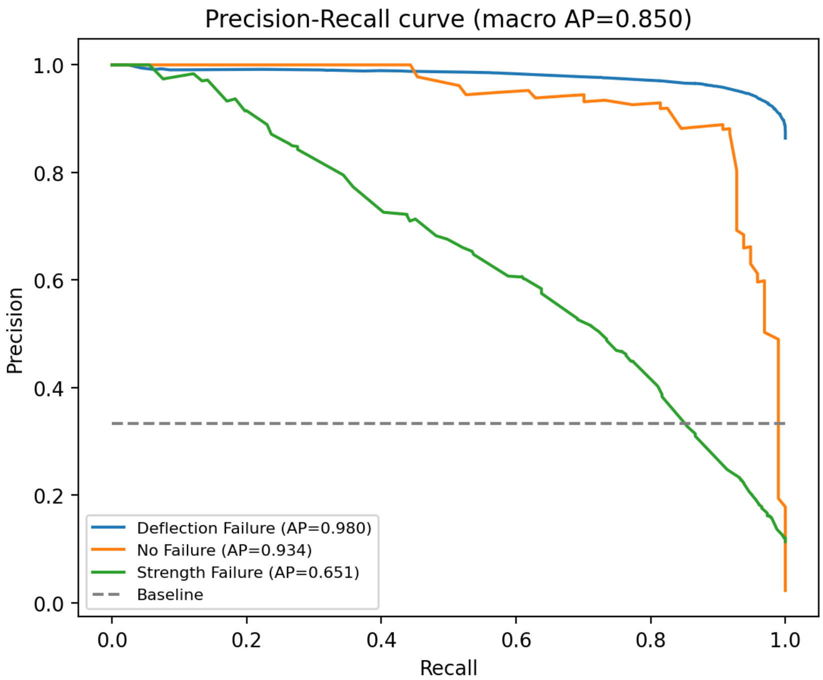

Figure 2: Precision–Recall curves. Macro average precision (AP) ≈ 0.850. Strong separability in *Deflection Failure* and *No Failure*; *Strength Failure* remains partially entangled, consistent with known coupling between strength degradation and large-deflection behaviour [2, 3].

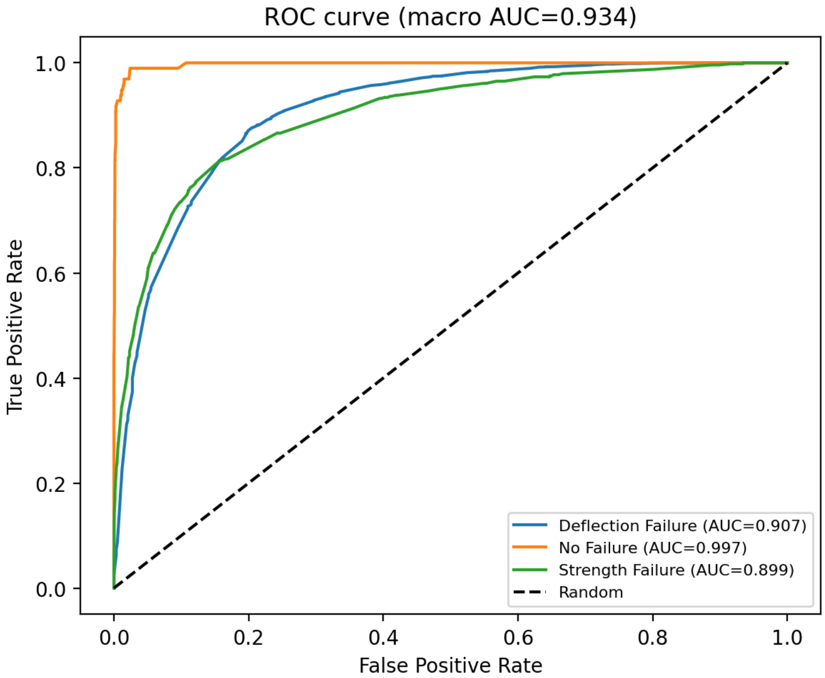

Figure 3: ROC curves. Macro AUC ≈ 0.934. Calibrated thresholds enable a conservative operating point to reduce unsafe false negatives.

5.3. FRT Regression

Table 4: FRT regressor performance.

Dataset Split	**MAE (min)**	**RMSE (min)**	R^2
Training	5.99	9.86	0.982
Validation	10.27	19.41	0.931

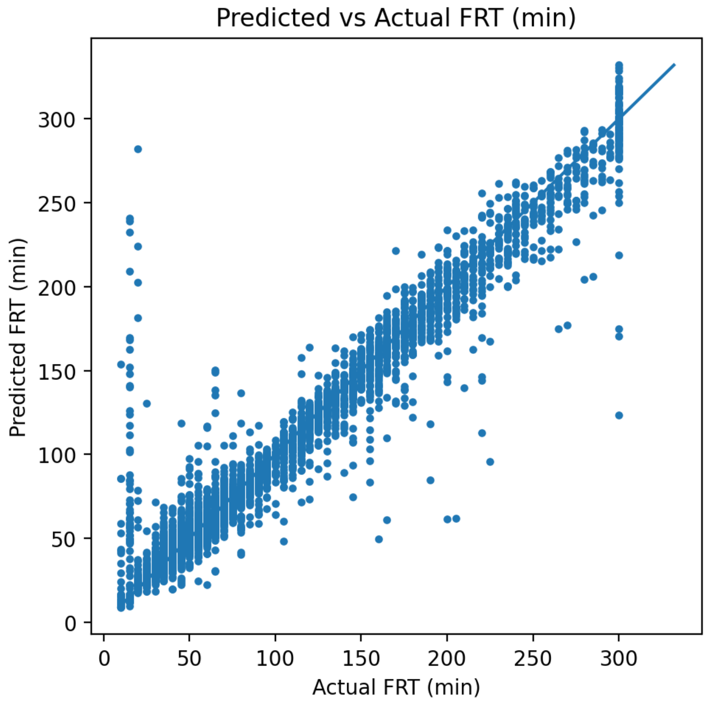

Figure 4: Predicted vs actual FRT (validation). Near 1:1 with widening variance at long durations (120 min), reflecting fewer high-endurance samples and greater thermal-path variability.

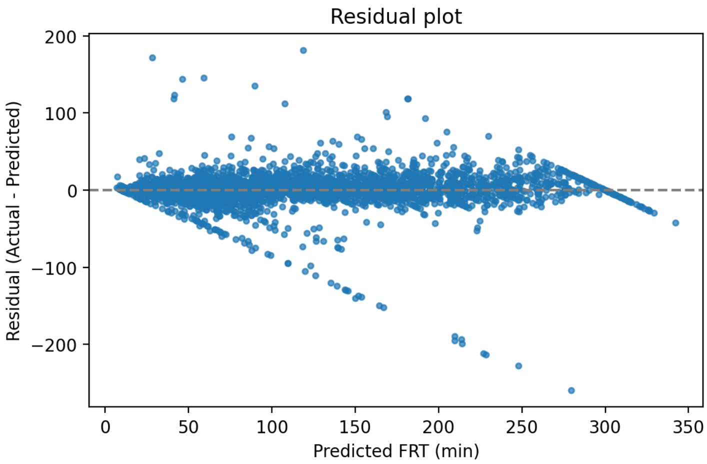

Figure 5: FRT residuals. Heteroscedasticity at high FRT suggests scope for stratified modelling or physics-informed constraints.

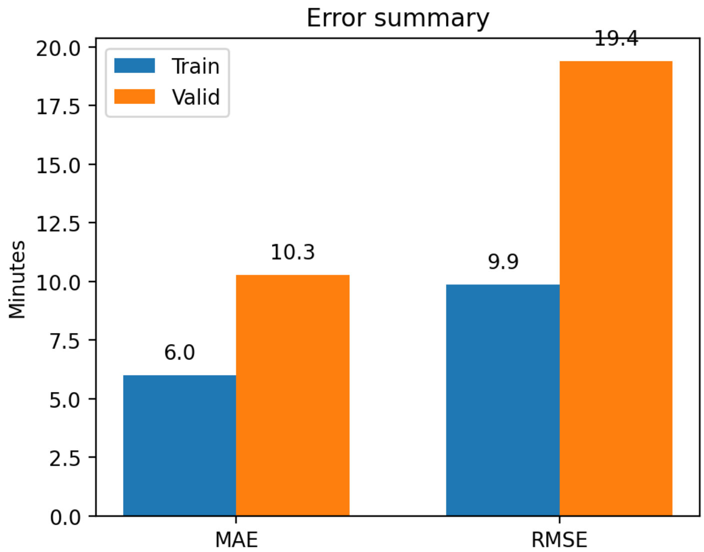

Figure 6: Error summary (MAE/RMSE). Validation error remains within a practical envelope for screening-level decisions.

5.4. Data Geometry (Correlation Context)

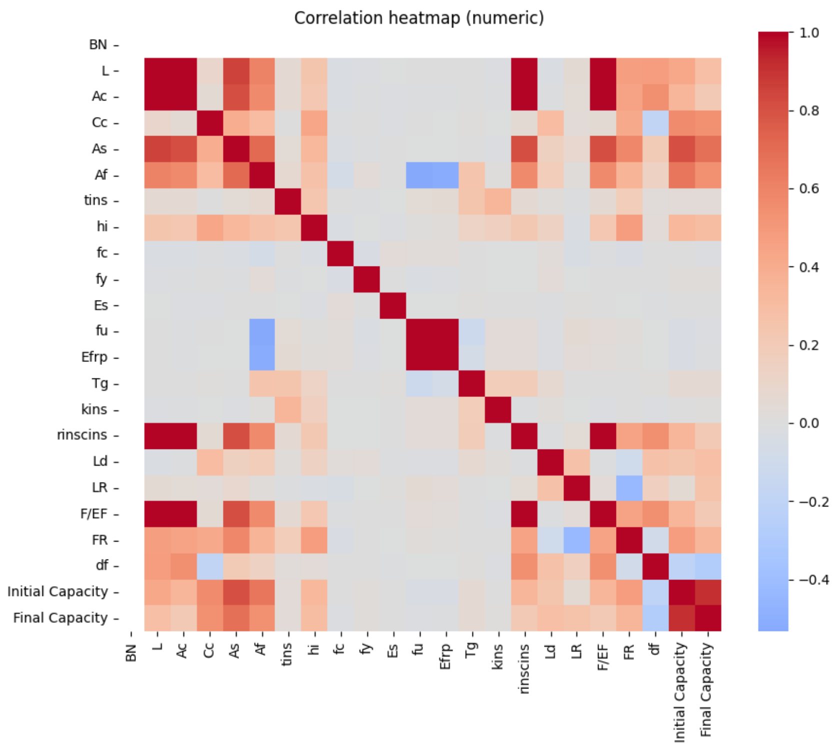

Figure 7: Feature correlation heatmap (train split). Clusters around geometric and material groups motivate our collinearity pruning ($|r| > 0.97$).

6. Interpretability and Design Guidance

6.1. Global Trends (SHAP)

Global SHAP analyses (not shown due to space) reveal:

- **Insulation thickness/depth** (t_{ins}, h_i) $\uparrow \Rightarrow$ shifts toward *No Failure* / away from *Strength Failure*.
- **Concrete cover** (C_c) $\uparrow \Rightarrow$ reduced strength-governed risk under heating.
- **Load ratio** (LR) $\uparrow$ and **applied load** (L_d) $\uparrow \Rightarrow$ higher failure likelihood unless insulation is increased.
- **FRP area** (A_f) interacts with thermal protection: large A_f without adequate protection increases risk of FRP/bond-related strength loss [1, 2].

6.2. From Explanation to Action (Safety-Aware)

We embed Eurocode-like exposure+margin logic in the dashboard and calibrate probabilities to **favour safety** (reduce false negatives in at-risk classes). Recommendations are phrased as *risk-reducing adjustments* (e.g., increase h_i to 80–100 mm; increase C_c to 35–40 mm; reduce LR where feasible) rather than deterministic guarantees.

6.3. Critical Limitation: Rare *No Failure* Class

Because *No Failure* represents $\sim$2–3% of data, SHAP explanations in that region have wider uncertainty. As such, **do not** interpret SHAP as a prescriptive "flip to safe" oracle; treat it as indicating *directional trends*. This aligns with the conservative posture in structural fire engineering: use the tool to *support*—not replace—engineering judgment [3, 2].

7. Discussion and Future Work

Our pipeline advances prior ML work by (i) adding **mode classification**, (ii) **calibrated** probabilities with **safety thresholds**, and (iii) **interpretability-to-design** mapping. Compared with experimental/numerical baselines [1, 2, 3, 4], it offers rapid, screening-level assessments consistent with Eurocode framing, suitable for exploring what-if designs.

Limitations. (1) Heavy reliance on simulated cases in the dataset may bias boundaries to its modelling assumptions; (2) default split is stratified but not grouped by experimental series; (3) ECE/Brier calibration diagnostics are not yet reported; (4) interpretability near *No Failure* is constrained by rarity.

Future Work. (i) External blind validation from independent tests; (ii) grouped splits by series; (iii) physics-informed constraints and rule extraction; (iv) calibration diagnostics (ECE/Brier) and bootstrap CIs; (v) extend taxonomy beyond 3 classes (debonding vs rupture vs shear).

Conclusion

We present a reproducible, interpretable, and **safety-calibrated** ML framework that predicts failure mode and FRT for FRP-strengthened RC beams under fire. Results are competitive (Accuracy ∼0.91, Macro-F1 ∼0.82; FRT MAE ∼10 min) and explanations align with known physics (insulation, cover, demand). The system is a *decision-support* tool: it communicates risk and design direction while respecting the conservative posture of structural fire engineering.

References

[1] Ahmed, A. & Kodur, V. (2011). *Behavior of FRP-strengthened reinforced concrete members under fire exposure* (representative experimental/analytical study). (Cites the role of adhesive softening, bond degradation, and need for protection.)

[2] Firmo, J. P., Correia, J. R., & Bisby, L. A. (2012). *Fire behavior of FRP-strengthened concrete structural elements: a state-of-the-art review.* (Anchorage protection importance; debonding mechanisms and mitigation.)

[3] Zhang, L., *et al.* (2018). *Thermo-mechanical finite element modelling of FRP-strengthened RC beams in fire.* (High-fidelity FE approach; sensitivity to material assumptions.)

[4] Dong, Y., *et al.* (2016). *Numerical simulation of insulated FRP-strengthened concrete members under standard fire exposure.* (Parametric insulation scenarios; qualitative alignment with our model trends.)

[5] Bhatt, P., Kodur, V., & Naser, M. (2021). *Machine learning-based prediction of fire resistance time for FRP-strengthened RC beams.* (Duration-focused ML; limited interpretability and no mode classification.)

[6] Bhatt, P. (2023). *Fire resistance of FRP-strengthened beams* (Version 6) [Dataset]. Mendeley Data. doi:10.17632/3c2szhbdn5.6.