Data-Driven Fragility Modeling of Utility Poles: Insights, Limitations, and the Need for Physics-Informed Features

Hanbyeol Shin
Dept. of Civil, Environmental and Geodetic Engineering
The Ohio State University
Columbus, OH, USA
shin.922@osu.edu

Abdollah Shafieezadeh
Dept. of Civil, Environmental and Geodetic Engineering
The Ohio State University
Columbus, OH, USA
shafieezadeh.1@osu.edu

Steven Quiring
Dept. of Geography
The Ohio State University
Columbus, OH, USA
quiring.10@osu.edu

***Abstract*—Data-driven fragility modeling of utility poles offers a powerful alternative to traditional physics-based methods but is hindered by the extremely imbalanced nature of real-world failure data. This paper presents a comprehensive framework to generate continuous fragility functions (failure probabilities) from binary (failure/survival) historical data. We frame this as a probabilistic classification problem and compare three key data-level resampling techniques to manage the extreme class imbalance: Synthetic Minority Over-sampling (SMOTE), Random Under-sampling (RUS), and a hybrid SMOTE-RUS pipeline. We evaluate Logistic Regression (LR) and LightGBM (LGBM) models across 10 independent iterations to ensure statistical robustness. Performance is evaluated on the original, imbalanced test set using both threshold-optimized metrics (e.g., F1-Score) and ranking metrics (ROC AUC), and physical interpretability. The results demonstrate that linear models are insufficient, yielding negligible F1-Scores and learning physically incorrect relationships. While the non-linear LGBM models produced a stable predictive signal, SHAP analysis revealed that all configurations failed to learn the correct physical impact of pole geometry (aspect ratio). This critical finding highlights the limitations of purely data-driven models and underscores the necessity of incorporating additional structural and environmental covariates to achieve physically-consistent fragility curves.**

***Index Terms*—fragility assessment, wind hazard, utility poles, imbalanced dataset**

I. Introduction

Power distribution networks are critical infrastructure systems increasingly vulnerable to extreme wind events, such as hurricanes and severe windstorms. Utility poles, the primary support structures for overhead distribution lines, represent a critical point of failure during these hazard events. The failure triggers widespread power outages, prolonged restoration times, and significant economic losses. Consequently, quantifying the structural vulnerability of utility poles through fragility analysis is essential for enhancing system resilience, supporting everything from pre-event mitigation to post-event resource allocation [1].

Traditional fragility modeling has relied predominantly on physics-based methods, utilizing finite element analysis and structural mechanics to simulate pole behavior under wind loading. For instance, researchers have developed detailed mechanical models that account for pole material properties, geometric characteristics, loading conditions, and failure mechanisms [1]–[3]. While these models provide a rigorous theoretical foundation, they face inherent practical limitations. First, they require highly detailed asset characteristics, such as specific material properties and localized loading conditions, that are often unavailable for existing infrastructure [3], [4]. Second, physics-based models often rely on simplified assumptions about boundary conditions, load distributions, and failure modes that may not fully capture the complexity of real-world scenarios [1]. Finally, the computational demands of high-fidelity physics-based simulations can be prohibitive when analyzing large distribution networks with thousands of poles [2].

Recent advances in machine learning (ML) techniques offer a data-driven alternative, leveraging rich datasets from asset management systems, outage management systems, damage inspection records, and geographic information systems. When combined with historical hazard data, these datasets provide an empirical basis for developing fragility models directly from observed pole performance during past wind events [4], [5]. Data-driven approaches can inherently capture the complex interactions between multiple factors affecting pole vulnerability, including age-related degradation, environmental conditions, maintenance history, and site-specific characteristics that are difficult to model explicitly using physics-based methods. However, applying ML to pole failures presents a formidable challenge: extreme class imbalance. In real-world observational data, failure events are exceptionally rare compared to survival instances, often leading models to prioritize overall accuracy over the detection of critical damage states.

This paper presents a data-driven fragility modeling frame-

This work was partly supported by the Lichtenstein endowment at The Ohio State University. Any opinions, findings, and conclusions expressed are those of the authors and do not necessarily reflect those of the sponsoring organization.

979-8-3315-8138-1

work designed to address this extreme class imbalance. We conduct a comparative analysis of linear (Logistic Regression) and non-linear (LightGBM) models combined with three resampling strategies: Synthetic Minority Over-sampling (SMOTE), Random Under-sampling (RUS), and a hybrid pipeline. Beyond predictive performance, we evaluate the models' physical consistency using fitted coefficients and SHAP feature attribution. Our results demonstrate that while non-linear models provide a stable risk-ranking signal, they still struggle to learn physically-plausible relationships for key engineering features. This study establishes a validated baseline and highlights the critical need of integrating additional structural covariates to develop robust, interpretable fragility models.

II. Dataset and Preprocessing

This study leverages multiple data sources to construct a comprehensive dataset for data-driven fragility modeling of utility poles under extreme wind conditions. The dataset integrates utility asset records, historical storm events, and high-resolution meteorological reanalysis products spanning a five-year period from 2010 to 2015 in a Western U.S. coastal region.

A. Data Sources

Three primary categories of raw data were acquired for this study.

- *Utility Asset and Outage Data:* A utility partner provided: (1) a master pole inventory of approximately 1.7 million distribution poles, containing asset characteristics such as geographic coordinates, installation/removal dates, physical dimensions (height and circumference), and materials; (2) equipment replacement records documenting 2,328 pole failures; and (3) operational outage logs.
- *Historical Storm Events:* Historical storm episodes were obtained from the National Centers for Environmental Information (NCEI) Storm Events Database [6]. All recorded events within the utility service territory during the 2010-2015 period were extracted, providing a structured database of storm start times, end times, and geographic extents.
- *Wind Hazard Intensity Data:* Three state-of-the-art meteorological datasets were evaluated as potential sources for hazard intensity: the 0.25° ERA5 Reanalysis [7], the 0.1° ERA5-Land Reanalysis [8], and the 3-km High-Resolution Rapid Refresh (HRRR) [9].

B. Data Integration and Labeling

To create a unified, labeled dataset, a multi-step integration process was implemented. First, the pole replacement records were merged with the master pole inventory using unique pole identifiers to associate each failure event with its detailed asset characteristics. Next, a spatio-temporal matching algorithm was implemented to link these pole replacements to specific storm events from the NCEI database and service disruptions from the utility's outage logs. A replacement was attributed to a storm or outage if its recorded removal date fell within 28 days of the respective event's initiation.

This timeframe is essential because the dataset contains administrative replaced dates (reflecting work-order completion in the asset management system) rather than instantaneous physical failure times. Operationally, this window bridges the transition from the initial emergency restoration phase to the permanent repair phase, during which compromised structures are formally replaced. This 28-day threshold is further justified by the physical reality of major storm restoration; in hardest-hit areas, restoration and the associated data collection can span three to four weeks due to the scale of infrastructure destruction in the case of major storms [10].

C. Hazard Selection and Feature Engineering

A trade-off analysis was conducted to select the optimal meteorological dataset. The HRRR dataset provides the highest spatial resolution (3-km), making it ideal for capturing localized high-wind events. However, its data records are only available beginning in July 2014, which severely limited the number of historical storm events available for model training within our 2010-2015 study period. Conversely, the standard ERA5 Reanalysis provides complete temporal coverage but suffers from a coarse 0.25° ($\approx$28 km) spatial resolution, which is likely to miss localized wind gusts.

The ERA5-Land Reanalysis was ultimately selected as the best compromise. At 0.1° ($\approx$11 km) resolution, it provides a significant improvement over standard ERA5 while still offering the complete temporal coverage that the HRRR dataset lacked. The maximum wind speed from the ERA5-Land dataset during each storm event was subsequently assigned to individual poles through spatial interpolation at each pole's precise geographic coordinates.

With the hazard intensity assigned, the dataset was filtered to include only the five primary wood pole types with known designated fiber strengths: Douglas Fir, Western Cedar, Southern Pine, Western Pine, and Western Larch. Two physics-informed features were then engineered: (1) a fiber strength feature, which maps each wood type to its designated strength in ksi (e.g., 8.0 for Douglas Fir) [11], and (2) an aspect ratio (slenderness), calculated as Height/Diameter.

D. Final Analysis Population

For each storm event, the final analysis population comprises two groups: (1) *failed poles*, identified through the matched replacement records, and (2) *surviving poles*, defined as all other poles from the filtered master inventory that were in service during the storm. This spatio-temporal matching resulted in a dataset with a very large number of pole-storm observations (in the order of $N_{poles} \times N_{storms}$).

The resulting dataset, used for all subsequent modeling, contains a binary failure indicator (failure or survival) and the four engineered input features: aspect ratio, fiber strength, pole age, and wind speed. This final feature set, prior to model training, is extremely imbalanced, containing 79 million survival observations ($y = 0$) and only 260 failure observations

979-8-3315-8138-1

($y = 1$). This extreme imbalance necessitates the specialized modeling framework described in the next section.

III. Data-driven Fragility Modeling Framework

The development of a robust data-driven fragility model requires a framework that can (1) address the inherent challenges of real-world failure data and (2) leverage machine learning algorithms to learn the complex relationship between pole characteristics, hazard intensity, and failure probability. This section details the methodology employed, focusing on the techniques used to handle extreme class imbalance and the suite of classification models evaluated for fragility estimation.

A. Addressing Extreme Class Imbalance

A primary challenge in this analysis is the extreme class imbalance in the data. As established in Section II-D, the dataset is extremely imbalanced, with only 260 failure events (the minority class, $y = 1$) contrasted against over 79 million survival observations (the majority class, $y = 0$). Training on this raw data would cause models to achieve high but misleading accuracy by simply predicting survival in all cases, rendering them not useful for failure risk estimation.

To overcome this, data-level resampling techniques are applied to the training set only to create more balanced data for the models to learn from. While necessary, these techniques present trade-offs. Over-sampling with Synthetic Minority Over-sampling Technique (SMOTE) [12] can be computationally expensive and may create blurry decision boundaries by interpolating samples in a high-dimensional space without regard to physical constraints. Conversely, Random Under-Sampling (RUS) is fast but discards a vast amount of majority-class data. This reliance on a small, random sample of survival data means that a model's performance could be sensitive to the specific random selection. For all of these methods, the fundamental challenge remains: learning a precise, non-linear decision boundary from an extremely sparse set of minority-class data points while maintaining physical plausibility of the learned relationships.

This study compares three distinct strategies to navigate these challenges.

1) SMOTE: The first approach, SMOTE is an over-sampling technique that creates new, synthetic samples of the minority (failure) class. It generates new samples by interpolating between existing minority instances in the feature space, repeating the process until the training data reaches a 1:1 balance.

2) RUS: The second approach, RUS, is an under-sampling technique that reduces the size of the majority (survival) class. This method randomly removes observations from the majority class until a target ratio (e.g., 10:1 survivors to failures) is achieved.

3) Hybrid Resampling (SMOTE-RUS): The third approach is a two-stage hybrid pipeline that combines the benefits of both methods. First, RUS is applied to reduce the majority class to a manageable ratio (20:1). Second, SMOTE is applied to this new, smaller dataset to over-sample the minority class until a perfect 1:1 balance is achieved.

B. Classification Algorithms for Fragility Modeling

The objective of fragility modeling is to estimate a continuous probability of failure, P_f, given a set of input features X (e.g., pole characteristics and hazard intensity). This continuous probability function, $P_f = f(X)$, is the fragility.

Although the models are trained on binary "failure" (1) or "survival" (0) labels from the historical data, the machine learning framework is explicitly designed to output a continuous probabilistic value from 0.0 to 1.0, not a discrete 0/1 prediction. This is achieved through the specific architectures and output functions of the models. We compare two representative methods, each of which derives this continuous probability in a different manner.

1) Logistic Regression (LR): A generalized linear model that serves as a robust baseline. It calculates a linear combination of input features, which is then passed through a sigmoid (or logit) function. This function squashes the output to a continuous value between 0.0 and 1.0, directly representing the probability of failure. Input features are standardized for this model.

2) Light Gradient Boosting Machine (LGBM): A highly efficient Gradient Boosting implementation [13] based on decision trees. The raw output, which is a sum of the predictions from all trees, is passed through a sigmoid function to convert the final log-odds score into a continuous 0.0 to 1.0 probability. This model is trained on the unscaled features.

C. Model Evaluation and Interpretation

All models are evaluated on the same original, imbalanced test set to reflect real-world conditions. Given the extreme sparsity of failure events (0.0003%), the choice of evaluation metrics is critical for both statistical validity and practical utility. Standard "Accuracy" is inadequate in this context:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{1}$$

where TP, FP, TN, and FN represent True Positives, False Positives, True Negatives, and False Negatives, respectively, as a model could achieve 99.99% accuracy by simply predicting that no failures. We focus on metrics that evaluate the model's performance on the rare minority (failure) class.

1) Threshold-Optimized Metrics: Precision, Recall, and F1-Score require a discrete binary prediction. To properly evaluate the model's best-case practical performance, we identify the optimal probability threshold by maximizing the F1-Score for the minority (failure) class.

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

While standard, these metrics are highly sensitive to class distribution. In cases of extreme imbalance, achieving a high F1-Score is mathematically difficult as even a small number of false positives (FP) can lead to a precision near zero. Thus, while we report the F1-Score for completeness, it may not

979-8-3315-8138-1

fully capture a model's ability to differentiate risk in a large asset inventory.

2) Area Under the ROC Curve (ROC AUC): To address these limitations, we prioritize the Receiver Operating Characteristic Area Under the Curve (ROC AUC). The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at all classification thresholds:

$$\text{TPR (Recall)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (4)$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \quad (5)$$

ROC AUC represents the probability that a randomly chosen failure event is assigned a higher risk score than a randomly chosen survival event. This ranking is critical for utilities to prioritize inspection deciles even when absolute failure probabilities are low.

3) Model Interpretability and Feature Attribution: The ultimate framework output is the multivariate fragility function, $P_f = f(X)$. To interpret the drivers of this function, we analyze trained coefficients for LR models and employ SHAP (SHapley Additive exPlanations) for non-linear LGBM models to quantify the contribution of each covariate to the prediction [14].

IV. Results and Analysis

A. Model Evaluation

The six configurations are evaluated over 10 independent iterations with varied random seeds to ensure robustness. The performance of each model, quantified by the mean and standard deviation of metrics relevant to the minority (failure) class, is presented in Table I.

TABLE I
Model Performance on Imbalanced Test Set (10 Iterations)

Algorithm	Re-sampling	Precision	Recall	F1-Score	ROC AUC
LR	SMOTE	0.0002 (0.0002)	0.0282 (0.0269)	0.0003 (0.0004)	0.6134 (0.0158)
	RUS	0.0010 (0.0016)	0.0244 (0.0260)	0.0016 (0.0023)	0.6152 (0.0194)
	Hybrid	0.0004 (0.0006)	0.0282 (0.0255)	0.0008 (0.0011)	0.6145 (0.0163)
LGBM	SMOTE	0.1151 (0.1991)	**0.0462 (0.0303)**	**0.0363 (0.0199)**	0.6014 (0.0403)
	RUS	0.0662 (0.1533)	0.0359 (0.0208)	0.0213 (0.0154)	**0.7198 (0.0146)**
	Hybrid	**0.2674 (0.4138)**	0.0333 (0.0211)	0.0273 (0.0146)	0.6380 (0.0312)

Values represent Mean (Standard Deviation) over 10 iterations.

The results demonstrate a clear distinction between linear and non-linear modeling capabilities for this task. The linear baseline failed to produce a meaningful predictive signal, with F1-Scores and ROC-AUC values remaining near-random across all iterations. This confirms that a simple linear combination of the current features is insufficient to capture the rare failure events.

In contrast, the non-linear LGBM models captured a stable and measurable signal. While the absolute F1-Scores remain low —a mathematical inevitability in datasets with a 0.0003% failure rate where even minimal false positives penalize precision—the stable ROC AUC of 0.72 ($\sigma = 0.0146$) reveals significant operational utility.

For decision-making, the model serves as an effective priority filter rather than a binary classifier. By ranking the asset inventory, utility operators can concentrate limited inspection and hardening budgets on the highest-risk deciles. Even with low absolute probabilities, these prioritized sub-populations identify where the relative likelihood of failure is ten-fold higher than the network average, providing a robust baseline for risk-informed infrastructure management.

B. Model Interpretation

The trained LR models are defined by the log-odds (logit) function:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \quad (6)$$

where p is the probability of failure and $x_1 ... x_4$ represent the standardized features for pole aspect ratio, pole fiber strength, pole age, and wind speed at the pole location, respectively.

TABLE II
Trained Logistic Regression Coefficients

Re-sampling	β_0	β_1 (AR)	β_2 (Fiber Str.)	β_3 (Age)	β_4 (Wind)
SMOTE	0.0025	-0.0960	-0.2347	0.1866	0.3367
RUS	-2.3548	-0.1798	-0.1680	0.1736	0.3061
Hybrid	0.0016	-0.0882	-0.2468	0.1436	0.2775

The trained coefficients, shown in Table II, reveal a critical flaw. While all models correctly identified pole age (β_3) and wind speed (β_4) as risk-increasing factors (positive coefficients), and fiber strength (β_2) as a risk-decreasing factor (negative coefficient), they failed to learn the correct physical relationship for pole geometry. Physically, a higher aspect ratio (a more slender pole) should increase risk. All three LR models incorrectly assigned a negative coefficient to this feature (β_1), contradicting known engineering principles. This interpretive failure provides conclusive evidence that a linear model is insufficient for capturing the underlying fragility relationships.

As shown in Fig. 1, LGBM models have a critical and universal limitation. While the non-linear models correctly learned the impact of wind speed, pole age, and fiber strength, all three LGBM configurations failed to learn the correct physical relationship for aspect ratio. As seen in Fig. 1, all models (a, b, and c) incorrectly associate a high aspect ratio (a slender pole, shown as red dots) with a decrease in failure probability (a negative SHAP value). This consistent, physically-implausible finding suggests that the current feature

979-8-3315-8138-1

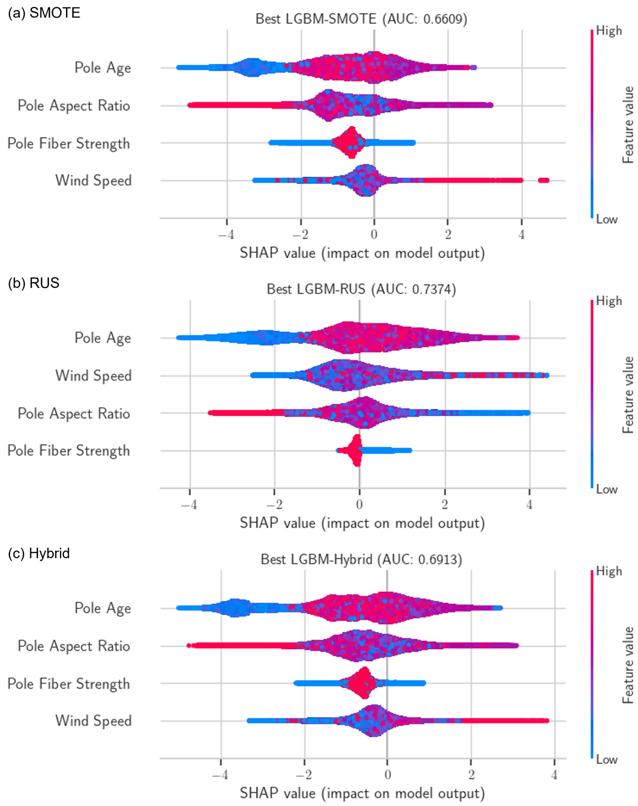

Fig. 1. SHAP summary plot for the LGBM models: (a) SMOTE, (b) RUS, (c) Hybrid

set, while informative, is insufficient to fully describe the complex failure mechanisms and likely reflects the influence of unobserved confounding variables.

V. Conclusion and Future Work

This paper presented a comprehensive data-driven framework for utility pole fragility modeling using extremely imbalanced real-world data, comparing linear and non-linear models across three resampling strategies. Results show that linear models are insufficient, failing both predictive and physical interpretability tests. While non-linear LGBM models produces a usable signal, a trade-off emerged: LGBM (RUS) offered superior risk-ranking (ROC AUC), whereas LGBM (SMOTE) provided the best classification performance (F1-Score).

Critically, SHAP analysis revealed that all configurations failed to learn the correct, physically-plausible relationship for pole geometry (aspect ratio). This universal failure suggests that while observational data captures hazard and age impacts, the current feature set cannot fully resolve the complex, interacting factors of pole failure. This work establishes a validated baseline, but the failure in physical interpretability indicates that critical information is missing. Future research may focus on enriching the model by integrating additional, high-impact covariates. This includes topological and loading characteristics (e.g., span lengths, number and type of attachments like transformers, wind direction relative to the span), geotechnical properties (e.g., soil classification, soil moisture), and site-specific environmental factors (e.g., terrain exposure). The integration of these variables is a necessary next step to develop robust, physically-sound, and high-fidelity fragility curves for utility resilience planning and operational decision support.

References

[1] A. Shafieezadeh, U. P. Onyewuchi, M. M. Begovic, and R. DesRoches, "Age-dependent fragility models of utility wood poles in power distribution networks against extreme wind hazards," *IEEE Transactions on Power Delivery*, vol. 29, no. 1, pp. 131–139, 2014.

[2] A. M. Salman, Y. Li, and M. G. Stewart, "Evaluating system reliability and targeted hardening strategies of power distribution systems subjected to hurricanes," *Reliability Engineering & System Safety*, vol. 144, pp. 319–333, 2019.

[3] J. He, M. J. Anderson, A. Shnidman, B. S. England, E. C. Piesciorovsky, R. A. Davidson, and R. Nevius, "Modeling wooden pole strength degradation using inspection data," *Sustainable Cities and Society*, vol. 62, p. 102587, 2020.

[4] R. Nateghi, S. D. Guikema, and S. M. Quiring, "Power outage estimation for tropical cyclones: Improved accuracy with simpler models," *Risk Analysis*, vol. 34, no. 6, pp. 1069–1078, 2014, referenced for data-driven approaches.

[5] X. Ma *et al.*, "Data-driven approaches for power system resilience," in *2024 IEEE Power & Energy Society General Meeting (PESGM)*, 2024, pp. 1–5, placeholder entry - please update with actual paper details.

[6] National Centers for Environmental Information, "Storm events database," https://www.ncdc.noaa.gov/stormevents/, accessed:Oct. 2025.

[7] H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers *et al.*, "The ERA5 global reanalysis," *Quarterly Journal of the Royal Meteorological Society*, vol. 146, no. 730, pp. 1999–2049, 2020.

[8] J. Muñoz-Sabater, E. Dutra, A. Agust-Panareda, C. Albergel, G. Arduini, G. Balsamo, S. Boussetta, M. Choulga, S. Harrigan, H. Hersbach *et al.*, "ERA5-Land: A state-of-the-art global reanalysis dataset for land applications," *Earth System Science Data*, vol. 13, no. 9, pp. 4349–4383, 2021.

[9] S. G. Benjamin, S. S. Weygandt, J. M. Brown, M. Hu, C. R. Alexander, T. G. Smirnova, J. B. Olson, E. P. James, D. C. Dowell, G. A. Grell *et al.*, "A north american hourly assimilation and model forecast cycle: The rapid refresh," Monthly Weather Review, Tech. Rep. 4, 2016.

[10] S. A. Shield, S. M. Quiring, J. V. Pino, and K. Buckstaff, "Major impacts of weather events on the electrical power delivery system in the united states," *Energy*, vol. 218, p. 119434, 2021.

[11] N. A. W. P. C. (NAWPC), "Modulus of elasticity (moe) for wood poles," https://woodpoles.org/wp-content/uploads/TB_Pole_MOE.pdf, 2024, accessed: Oct. 2025.

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research (JAIR)*, vol. 16, pp. 321–357, 2002.

[13] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2017, pp. 3146–3154.

[14] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

979-8-3315-8138-1