Proc. of the European Safety and Reliability Conference (ESREL 2026)
Edited by José C. Matos, Paulo B. Lourenço, Michael Beer, Daniel Oliveira,
Edoardo Patelli, Hélder S. Sousa, Rui A. Silva and Monica Santamaria
 Published by Research Publishing, Singapore.
doi: 10.3850/ESREL2026061419_esrel26-p26427-cd

Towards a Framework for Multivariate Reliability Modeling: A Random Survival Forests Interpretability Case Study

Doha Meslem[1], Leon Hoffe[1] and Stefan Bracke[1]
[1]*Chair of Reliability Engineering and Risk Analytics, University of Wuppertal, Germany.*
E-mail: dmeslem.dm@gmail.com, hoffe@uni-wuppertal.de, bracke@uni-wuppertal.de

As technical systems become increasingly complex, several variables must be monitored during field use to ensure technical reliability. This results in multivariate systems where failures cannot be attributed to a single variable, as interactions among variables can play a critical role. Moreover, real-world sensor data often includes units (candidates) that have not yet failed, introducing the challenge of censored data. This paper presents a literature review and conceptual study of machine learning (ML) methods applied to multivariate sensor datasets with and without failures, with a particular focus on the interpretability of Random Survival Forest (RSF) models. It examines existing ML approaches for identifying critical variables and determining their critical ranges, encompassing unsupervised, supervised, and survival modeling approaches. Additionally, it reviews how different modeling frameworks address censored data, feature interdependence, and degradation behavior. Building on this, the study explores RSF. It is highlighted as a promising example of a non-parametric method capable of capturing multivariate interdependencies. The paper discusses the possibility of applying RSF to time-to-failure data, assessing its potential to identify critical variables and determine the critical ranges that lead to failure. RSF is expected to serve as the basis for a general multivariate reliability model, providing interpretations into which parameters accelerate failure; an emerging, highly relevant yet underexplored area in industrial reliability research.The study is based on simulated multivariate time-series sensor data representing light electric vehicle (LEV) fleet applications. These datasets serve as representative examples to illustrate the type of real-world information suitable for survival-based reliability modeling. These findings support proactive maintenance decisions. Overall, the study bridges predictive maintenance and explainable survival modeling with regard to technical complex systems, offering guidance on selecting suitable methods and enhancing understanding of time-to-failure prediction and critical variable identification.

Keywords: Explainable machine learning, Reliability modeling, Random Survival Forests (RSF), Multivariate time series, Automotive case study.

1. Introduction

As industries pursue faster and more reliable solutions, engineering systems have become increasingly complex, generating high-dimensional datasets with strong interdependencies. Multivariate systems increasingly rely on machine-learning-based approaches to handle complex and high-dimensional data and support reliability and maintenance decisions.

In reliability engineering, the main challenge is no longer failure detection but selecting appropriate modeling strategies. Although many methods exist, uncertainty remains about which variables drive failure and within which operating ranges. This paper focuses on interpretability techniques for identifying critical variables and their critical ranges in Random Survival Forest (RSF) models, and outlines a data-driven framework for reliability model selection informed by the conducted literature review.

A case study using a simulated fleet-level e-moped dataset demonstrates the framework through an RSF model. The paper combines a theoretical review of the framework and survival-based reliability modeling with a practical RSF application, illustrating how data-driven characterization supports interpretable reliability analysis beyond purely data-centric approaches.

2. Literature Review

Literature on RSF is well established since the initial work of Ishwaran et al. (2008). Methodological aspects of RSF, including ensemble hazard estimation and limitations of variable-importance-

based explanations under censoring and correlated predictors, have been studied in detail (Langbein et al., 2025; Schmid et al., 2016).

RSF has been widely applied in clinical and biomedical survival analysis for tasks such as risk stratification and high-dimensional survival prediction (Liao et al., 2024; Wang et al., 2019). Consequently, RSF applications and interpretability studies are predominantly demonstrated in health-related domains, with limited focus on engineering reliability settings. While RSF is frequently used, interpretability remains inconsistent in practice, as many applied studies rely on global importance measures that do not explain variable influence on failure risk directly (Kargar-Sharif-Abad et al., 2024). Recent surveys confirm that survival-based machine learning often achieves strong predictive performance but provides limited, non-standardized explanations beyond importance rankings (Langbein et al., 2025).

Several frameworks address algorithm selection in predictive maintenance and prognostics (Arena et al., 2024; Leohold et al., 2025). However, these approaches primarily operate at an algorithmic level and do not support dataset-driven method selection aimed at identifying critical variables and their associated operating ranges. Related survey and comparison studies in survival and reliability modeling further emphasize the importance of dataset characteristics for model selection (Ramkumar and Oscar, 2024), yet remain limited to survival-analysis-centric perspectives or high-level reviews.

Consequently, despite mature survival models such as RSF and existing taxonomies, there remains limited structured guidance mapping observable dataset characteristics to suitable modeling approaches in reliability engineering. This gap is particularly relevant in practical applications, where datasets vary substantially in censoring and label quality. The present work builds on existing literature by contributing a data-characteristics-driven decision study and demonstrating, through an RSF case study, how interpretability can be extended from identifying critical variables toward identifying critical operating ranges.

3. Background Information and Framework Exploration

A framework is understood as a structured decision aid that maps observable dataset characteristics to suitable modeling families. In reliability analysis of complex engineering systems, the central challenge is not the lack of available machine-learning approaches, but selecting methods that match the available data and support the identification of critical variables and operating ranges associated with failure risk. The choice of modeling approach is therefore primarily driven by the type and quality of available data.

When failure labels are not available, unsupervised learning approaches are typically applied. In such settings, failure is defined implicitly through deviations from normal system behavior rather than explicit failure events. Common unsupervised methods include autoencoders, Isolation Forest, Local Outlier Factor (LOF), and change-point detection techniques, which identify anomalous operating regimes in multivariate sensor data. While effective for anomaly detection, these methods provide limited linkage between anomaly severity, failure timing, and remaining useful life. Interpretability is typically post-hoc, for example by attributing reconstruction error to input features using SHAP-based techniques, as demonstrated by Antwarg et al. (2021) for autoencoder-based anomaly detection. However, even with such explanations, translating anomaly scores into standardized reliability quantities remain challenging (An and Cho, 2015; Sakurada and Yairi, 2014; Wu et al., 2024).

When labeled data are available, supervised learning approaches can be applied. Typical supervised models include logistic regression, Random Forests (RF), and gradient boosting methods such as XGBoost, which are widely used in fleet-level failure classification due to their robustness to nonlinear relationships and strong performance on tabular sensor features. These models generally output class labels or failure probabilities and are evaluated using metrics such as accuracy or AUC (Area Under the Receiver Operating Characteristic Curve, measuring the model's ability to

discriminate between classes across thresholds). Interpretability is commonly addressed through feature importance measures in tree-based models and post-hoc explanation methods such as SHAP, which enable the identification of influential variables and their critical operating ranges via dependence plots (Ribeiro et al., 2016). While effective for distinguishing failed from non-failed samples, standard supervised classification does not explicitly model time-to-failure and typically discards information from censored units.

When time-to-event information is available, survival analysis provides a supervised modeling framework that explicitly accounts for right-censored observations. Common survival approaches include Cox Proportional Hazards (CoxPH), RSF, and neural-network-based models such as DeepSurv and DeepHit. CoxPH offers strong interpretability through hazard ratios but relies on linearity assumptions unless extended. RSF relaxes these assumptions through non-parametric tree ensembles, capturing nonlinear effects and interactions while naturally handling censoring. Deep survival models further increase modeling flexibility but typically require post-hoc attribution methods to achieve interpretability. A key advantage of survival modeling is its direct alignment with reliability objectives, as it enables the identification of variables that accelerate failure risk and supports analysis of how risk evolves across operating ranges. Extracting critical variables and ranges therefore often requires a combination of model-intrinsic interpretability measures and targeted post-analysis (Katzman et al., 2018; Lee et al., 2018).

4. Methodology

This section describes the theoretical foundations of the applied methods and outlines how the RSF approach is used to model failure risk and interpretability in the presented case study.

4.1. *RSF Theory*

RSF is a non-parametric survival analysis method that uses an ensemble of decision trees to estimate time-to-failure and relative risk. Each tree is trained on a bootstrap sample of the dataset, such that approximately 64% of observations are unique, while the remaining out-of-bag samples provide an internal test set for that tree.

At each node, RSF randomly selects a subset of candidate features, and evaluates multiple candidate split points for each feature, typically defined as midpoints between consecutive sorted values. For a given feature x_k, each candidate cut c defines a split of the form

$$x_k \leq c \text{ vs. } x_k > c.$$

Each candidate split is evaluated using the log-rank test statistic, which measures how strongly the resulting left and right child nodes differ in their survival behavior.

At a node, for a candidate split s, RSF maximizes the log-rank statistic

$$\mathrm{LR}(s) = \frac{\left[\sum_j (d_{1j} - e_{1j})\right]^2}{\sum_j v_{1j}} \tag{1}$$

where, at each event time t_j, d_{1j} denotes the observed number of events in the left node, $e_{1j} = \frac{Y_{1j}}{Y_j} d_j$ the expected number of events, Y_{1j} the number of observations at risk in the left node, Y_j the total number at risk, and v_{1j} the corresponding variance term. The split chosen for the node is the feature-cut pair (x_k, c) that yields the maximum log-rank value among all evaluated candidates, corresponding to the strongest separation of survival outcomes between child nodes.

Trees are grown recursively until stopping criteria are met. Once a tree is fully grown, each terminal leaf estimates the survival function using the Kaplan–Meier estimator and the cumulative hazard using the Nelson–Aalen estimator. The Kaplan–Meier survival function for a terminal node is given by:

$$\hat{S}(t) = \prod_{t_j \leq t} \left(1 - \frac{d_j}{Y_j}\right) \tag{2}$$

where t_j is the j-th event time, d_j the number of failures at t_j, and Y_j the number at risk just before t_j. Final predictions are obtained by averaging survival functions or hazard estimates across all trees in the ensemble (Ishwaran et al., 2008; Ishwaran and Kogalur, 2010).

RSF Hyperparameters and Tree Construction

RSF was implemented using the scikit-survival library. An ensemble of 800 trees (`n_estimators` = 800) was used to obtain stable predictions. Node splitting required at least 10 samples (`min_samples_split` = 10), and terminal nodes contained a minimum of 5 observations (`min_samples_leaf` = 5) to ensure reliable Nelson–Aalen estimates. Feature-level randomness was introduced using `max_features = "sqrt"`, with optimal feature–threshold splits selected by the log-rank criterion.

4.2. *Case Study*

The following sections describe the dataset, feature construction, labeling strategy, and the RSF modeling approach.

4.2.1. *Dataset and Feature Construction*

The dataset consists of 84 independent e-moped simulation runs, each representing the operational history of a single unit and containing approximately 31,000 high-frequency time-stamped observations. Recorded sensor signals include battery current (three channels), battery voltage, vehicle speed, applied torque, and traveled distance. To enable simulation-level modeling, each time series was aggregated into a fixed-length representation by computing descriptive statistics per variable, including minimum, maximum, mean, variance, range, and selected percentile (25th, 50th, 75th, and 90th percentiles). In total, 11 features per simulation were constructed, combining original sensor variables with derived statistical descriptors.

4.2.2. *Labeling Strategy and Event Definition*

RSF requires, for each instance, a time-to-event value, failure labels, and features (Ishwaran et al., 2008). In the absence of failure labels, heuristic event labels were assigned by defining a binary indicator ($0 =$ non-failure-like, $1 =$ failure-like) and an observation time corresponding to either the onset of failure-like behavior or censoring. Feature selection was guided by physical considerations, targeting electrical stress, load variability, and mechanical demand. Accordingly, the 90th percentile of battery current was selected to capture sustained electrical load and associated I^2R losses, current variance to reflect electrical instability and dynamic loading, and torque variance to serve as a proxy for cyclic mechanical stress linked to fatigue, which are widely recognized contributors to degradation in electro-mechanical systems. Maximum values were treated cautiously due to their sensitivity to isolated spikes, whereas variance- and percentile-based measures provide a more robust characterization of sustained operating stress. Based on this physical rationale, current variance, the 90th percentile of current, and torque variance were selected to define failure-like behavior within this context. Simulations exceeding the upper 60% thresholds in all these three features were labeled as failures, yielding approximately 25% failed units across all simulations, a common percentage within such approaches.

5. Results and Discussion

As the concept of interpretability and explainability is a central focus of this paper, related gaps in the application of commonly used algorithms were first identified. To address these gaps, RSF was applied to the processed dataset as a case study, serving as an investigation of interpretability-oriented analysis. This study does not focus solely on survival curves, as in most RSF research, but also examines critical variables, operating ranges, multivariate relationships within the data, and model explainability, as demonstrated in the following sections. The evaluation metrics are divided into model performance, risk scores and ranking, survival meaning of risk, critical variables, and ranges. The analysis therefore begins by establishing the predictive validity of the model before any interpretative conclusions are drawn, ensuring that all subsequent explanations are based on a model with meaningful predictive capability.

5.1. *Model Performance*

The C-index evaluates how well an RSF ranks survival risk (Schmid et al., 2016). A value of 0.92 means the model correctly orders which unit fails first in 92% of comparable cases.

5.2. *Risk Score and Ranking*

After validating the model, RSF risk scores are used to rank simulations by relative failure risk. Higher scores indicate more stressed units. Comparing the ranking with the failure labels indicates that a risk threshold of approximately 3.8 correctly identifies about 85% of the failure cases, demonstrating that higher risk scores are strongly associated with observed failures.

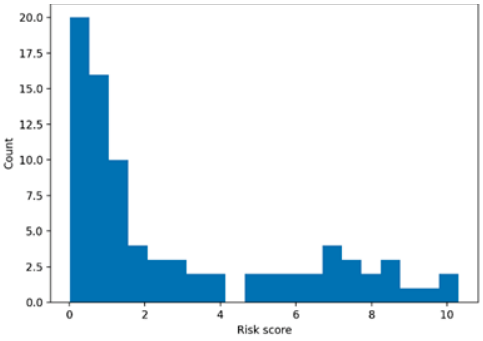

Fig. 1. Risk score distribution among samples

Figure 1 shows the distribution of RSF risk scores across all simulations. The distribution is strongly right-skewed, with approximately 54% of simulations exhibiting low risk scores up to around 1.6, corresponding to units predominantly labeled as healthy. Beyond this range, risk scores spread more gradually across higher values, indicating a continuous increase in failure risk rather than a sharp binary separation. This gradual spread reflects multiple intermediate risk levels and suggests that degradation develops progressively across the fleet, rather than occurring abruptly.

5.3. *Survival Function*

Survival function, a standard metric by RSF, represents the probability that a simulation survives beyond a given time point and is evaluated for test samples over the observed time range.

Figure 2 Predicted survival curves for the first 10 test samples. The horizontal axis represents time steps (sampling index), corresponding to the sequential position in the recorded time series;

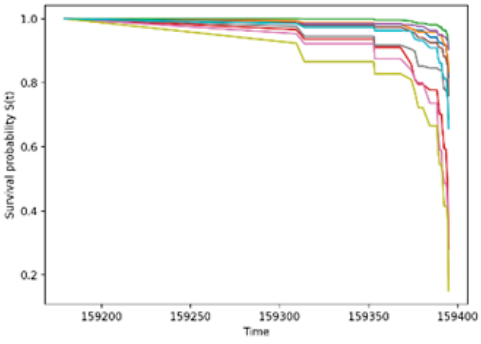

Fig. 2. Predicted survival curves against time

each step corresponds to 0.5 s (sampling rate: 2 Hz). The vertical axis shows the predicted survival probability. Survival probabilities remain similar in early stages and diverge mainly after approximately 80% of the runtime, reflecting increasing differences in degradation behavior.

5.4. *Reliability Interpretation and Maintenance Implications*

The predicted survival curves provide time-dependent reliability estimates for each vehicle, where the survival probability indicates the likelihood that a unit remains operational beyond time stated. For practical maintenance decision-making, a fixed reliability threshold of 80% was adopted as an intervention boundary. For each vehicle, the survival curve is evaluated up to this threshold, and once the predicted survival probability falls below 80%, the unit is considered insufficiently reliable and flagged for preventive maintenance, even if no failure has yet occurred due to censoring. As observed in the survival curves, this threshold is reached earlier for some vehicles and later for others, reflecting degradation behavior across the fleet. This approach enables reliability-based maintenance planning rather than deterministic failure time prediction and is naturally supported by RSF through its ability to model censored data. The choice of the prediction horizon and reliability threshold remains application-specific and can be aligned with company-defined maintenance intervals or risk tolerance levels, thereby supporting the gen-

eralizability of the proposed approach across different operational settings.

5.5. *Critical variables*

RSF identified risk patterns consistent with the rule-based labeling adopted in this study. The output is evaluated through a comparison of mean predicted risk across different combinations of satisfied conditions, as shown in Figure 3.

As illustrated, when none or only one of the top 60% conditions is satisfied, mean risk remains low, with moderate increases when two conditions are met. In contrast, when all three conditions are simultaneously satisfied, mean predicted risk increases sharply, indicating a strong interaction effect beyond individual variables.

To further validate this behavior, a binary indicator (rule_mask) is introduced, where rule_mask = 1 denotes that all predefined conditions are simultaneously satisfied, and rule_mask = 0 otherwise. Figure 3 shows that risk scores remain moderate when rule_mask = 0, but increase substantially when rule_mask = 1, reflecting extreme risk states.

This behavior, therefore, demonstrates strong agreement between heuristic labels and learned survival risk.

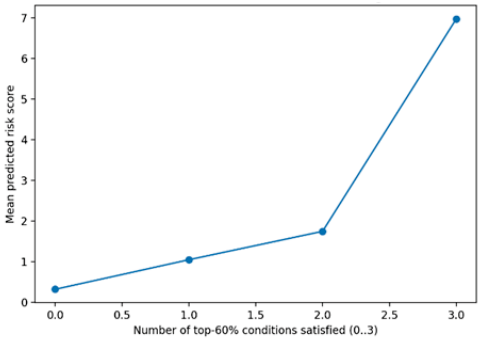

Fig. 3. Increase in mean predicted risk with the number of satisfied high-risk conditions

5.6. *Feature Importance*

After establishing multivariate relationships, feature importance analysis was used to identify the variables primarily responsible for increased failure risk. As expected, the results indicate that current variance, torque variance, and the upper 90th percentile of current are the most influential variables, as permuting these features leads to the largest degradation in model performance. Other variables exhibit only minor contributions in comparison.

5.7. *Critical Range*

The initial analysis illustrates how the model responds to a feature in general, independent of the empirical study. Initially, one feature is varied gradually while all other features are held fixed at their median values, and the resulting change in predicted risk is observed. This shows what the model has learned would happen to risk if only this feature increased.

For torque variance, as shown in Figure 4 (left), the predicted risk remains nearly constant up to approximately 280–290 %², beyond which a sharp increase is observed. In this region, risk rises rapidly from around 2.2 to values above 3.5, indicating a clearly defined critical operating range.

In contrast, the current range, shown in Figure 4 (right) does not exhibit a distinct threshold behavior. Instead, predicted risk changes gradually across the feature range, suggesting a weaker and more spread influence on failure risk.

For current variance, a notable increase in predicted risk occurs around 1.15–1.2 × 10^7 mA^2, where risk rises sharply from approximately 1.7 to above 4.5, indicating a clear critical range. The 90th percentile of current shows an even more pronounced threshold behavior, with a steep rise in risk beyond roughly 1950–2000 mA, reaching values above 7 and marking a distinct critical operating range. Variables not included in the failure labeling do not exhibit comparably clear threshold-like ranges in this experiment.

The relationship between individual features and predicted risk was analyzed using the simulation data. High-risk values concentrate within specific ranges for key variables (e.g., current variance, q90 current, and torque variance), consistent with previously identified critical thresholds, while other variables show more diffuse patterns

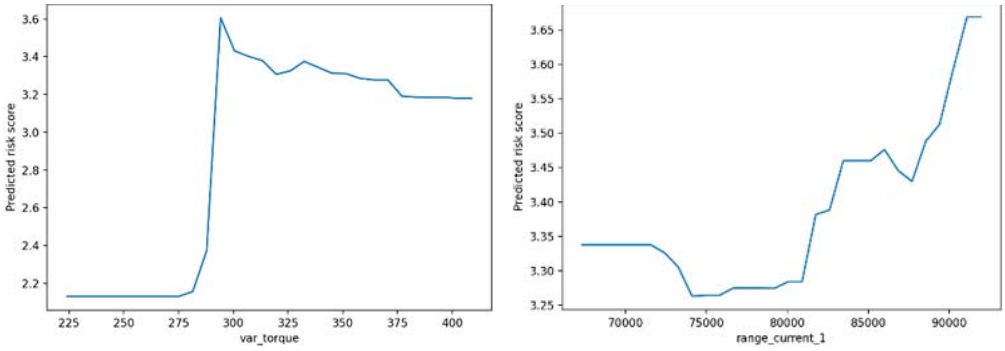

Fig. 4. Predicted risk score against torque variance (left) and current range (right)

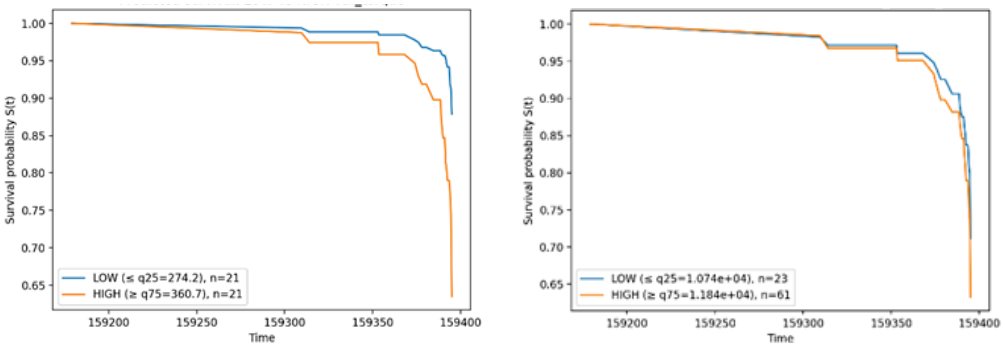

Fig. 5. Survival probability against time, low vs high-torque variance (left) and maximum current (right)

without clear risk transitions.

To assess whether individual variables meaningfully affect failure time, survival curves were compared between simulations with low and high feature values. For each variable, two survival curves are plotted; a clear separation between these curves indicates a strong influence on survival behavior.

Simulations with high torque variance exhibit lower survival probabilities over time compared to the low group, indicating increased failure risk under elevated mechanical variability. As shown in Figure 5 (left), higher torque variance leads to earlier and steeper survival decline, confirming its role as a critical stress driver. In contrast, variables such as maximum current show only limited separation between low- and high-value groups, as shown in Figure 5 (right), suggesting a weaker influence on survival. Overall, clear survival-curve separation corresponds to variables with strong impact on failure timing, whereas overlapping curves indicate limited standalone relevance.

6. Conclusion and Outlook

This study demonstrated the use of RSF for reliability analysis on a simulated fleet-level e-moped dataset, highlighting their ability to capture nonlinear risk behavior, multivariate interactions, and censored observations. The results provided interpretable insights into critical variables and operating ranges relevant for reliability-driven maintenance decisions. By emphasizing explainability, the analysis shows how survival-based machine learning can move beyond predictive performance toward actionable reliability insights in complex engineering systems.

Future work would extend this study by incorporating real-world data to validate the findings

under practical operating conditions. In addition, neural network–based survival models, such as DeepSurv, could be considered to compare predictive performance and interpretability against RSF. Additionally, further development of the proposed framework would follow. Further analysis of reliability thresholds, including the selected 80% intervention boundary, could assess their influence on maintenance timing, risk tolerance, and operational cost. Finally, applying the framework to additional domains would further validate its generalizability for reliability-oriented model selection.

Acknowledgement

This study was conducted as part of the research project 'Development of a Multivariate Reliability Model for Representing Degradation Behavior, Determining the Remaining Useful Life, and Predicting Failure Events (DEREPRO) of Technically Complex Products Using Multivariate Usage Profiles' funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 557903399.

References

An, J. and S. Cho (2015). Variational autoencoder based anomaly detection using reconstruction probability. *Special Lecture on IE*.

Antwarg, L., R. M. Miller, B. Shapira, and L. Rokach (2021). Explaining anomalies detected by autoencoders using shapley values. *Expert Systems with Applications 186*, 115736.

Arena, S., E. Florian, F. Sgarbossa, and I. Zennaro (2024). A conceptual framework for machine learning algorithm selection in predictive maintenance. *Engineering Applications of Artificial Intelligence 133*, 108340.

Ishwaran, H. and U. B. Kogalur (2010). Consistency of random survival forests. *Statistics & Probability Letters 80*(13–14), 1056–1064.

Ishwaran, H., U. B. Kogalur, E. H. Blackstone, and M. S. Lauer (2008). Random survival forests. *The Annals of Applied Statistics 2*(3), 841–860.

Kargar-Sharif-Abad, H., K. Böhm, and M. Reischl (2024). Shap-driven explainability in survival analysis. In *Proceedings of the Explainable AI in Healthcare Workshop*, Volume 3765 of *CEUR Workshop Proceedings*.

Katzman, J. L., U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger (2018). Deepsurv: Personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Medical Research Methodology 18*(1), 24.

Langbein, S. H., M. Krzyzinski, H. Baniecki, P. Biecek, and M. N. Wright (2025). Interpretable machine learning for survival analysis. *Biometrical Journal 67*(6).

Lee, C., W. R. Zame, J. Yoon, and M. van der Schaar (2018). Deephit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 32.

Leohold, S., H. Engbers, and M. Freitag (2025). Enhancing condition monitoring: A semi-automatic framework using meta-learning for algorithm selection. *IFAC-PapersOnLine*.

Liao, T., T. Su, Y. Lu, L. Huang, W.-Y. Wei, and L.-H. Feng (2024). Random survival forest algorithm for risk stratification and survival prediction in gastric neuroendocrine neoplasms. *Scientific Reports 14*, 26969.

Ramkumar, P. and M. Oscar (2024). Comparison of cox proportional hazards and random survival forests for engineering reliability. *Reliability Engineering & System Safety 241*, 109437.

Ribeiro, M. T., S. Singh, and C. Guestrin (2016). Why should i trust you? explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD*.

Sakurada, M. and T. Yairi (2014). Anomaly detection using autoencoders with nonlinear dimensionality reduction. *Proceedings of the MLSDA Workshop*.

Schmid, M., M. N. Wright, and A. Ziegler (2016). On the use of harrell's c for clinical risk prediction via random survival forests. *Expert Systems with Applications 63*, 450–459.

Wang, P., Y. Li, and C. K. Reddy (2019). Machine learning for survival analysis: A survey. *ACM Computing Surveys 51*(6).

Wu, D., S. Gundimeda, S. Mou, and C. J. Quinn (2024). Unsupervised change point detection in multivariate time series. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Volume 238 of *Proceedings of Machine Learning Research*. PMLR.