FEDERATED SURVIVAL ANALYSIS: ENSEMBLE AND NEURAL METHODS FOR DISTRIBUTED TIME-TO-EVENT DATA*

Alberto Archetti
Department of Electronics, Information and Bioengineering
Politecnico di Milano
Milan, Italy
alberto.archetti@polimi.it

ABSTRACT

Survival analysis plays a pivotal role in healthcare by modeling time-to-event outcomes, such as patient mortality or disease progression, and directly influencing clinical decisions and resource management. Despite its central role, the inclusion of advanced machine-learning predictive techniques faces significant barriers: the scarcity of large, high-quality datasets and restrictive privacy regulations that prevent institutions from sharing sensitive patient information. Federated learning addresses these barriers by allowing multiple healthcare entities to collaboratively develop models without centralizing or exchanging confidential data, thus providing access to more extensive and diverse patient cohorts while preserving privacy. This thesis introduces novel methods specifically tailored to federated survival analysis. First, we propose FPBoost, a fully parametric gradient boosting approach designed to flexibly model survival data by directly optimizing the survival likelihood. Second, we investigate polynomial-based interpolation as a post-processing method to enhance the calibration performance of discrete-output neural survival models. Third, building upon ensemble methods, we present FedSurF, a federated version of random survival forests capable of aggregating locally trained models effectively in just one communication round, even under data heterogeneity. Finally, we explore generative models to create realistic, privacy-preserving synthetic datasets, enabling comprehensive validation and improved robustness of survival prediction methods. Collectively, these contributions demonstrate that the use of advanced machine learning techniques significantly improves the discrimination and calibration of survival models, ultimately fostering precise and efficient healthcare decision-making.

***K*eywords** Survival Analysis · Machine Learning · Ensemble Methods · Federated Learning

1 Introduction

Survival analysis, the statistical study of time-to-event processes such as disease progression, patient mortality, or mechanical failures, plays a critical role in healthcare [1, 2]. Accurate predictions about event timing are essential for guiding clinical decisions, allocating resources, scheduling treatments, and ultimately improving patient outcomes. Although machine learning approaches offer significant potential to enhance the predictive capabilities of survival models by capturing complex, nonlinear relationships, their practical integration into clinical practice faces significant obstacles. The main challenges are stringent data privacy regulations and the fragmentation of patient datasets across multiple institutions, hindering the aggregation of sufficient data for robust, generalizable models [3].

Federated learning (FL) has emerged as a strategy to overcome these limitations. By enabling collaborative training of machine learning models without the need to centralize sensitive patient data, FL addresses privacy concerns and

*This document provides a concise summary of the author's doctoral thesis, submitted in fulfillment of the requirements of the ScuDO doctoral school as part of the *National Ph.D. Programme in Artificial Intelligence* (Cycle XXXVII). The research was conducted under the supervision of Prof. Matteo Matteucci (Politecnico di Milano), with Prof. Stefano Di Carlo (Politecnico di Torino) serving as the program chair.

facilitates access to diverse, multi-institutional datasets [4, 5]. In this paradigm, individual institutions train models locally and share only model parameters, which are aggregated centrally to form a robust global model. Thus, FL maintains patient confidentiality while harnessing the collective strength of distributed data sources [6].

Within this context, this document summarizes the methodologies we developed during our doctoral studies designed to advance survival analysis in both centralized and federated scenarios. As the main thesis, this manuscript is split into two parts: the first focusing on single-client survival analysis, and the second extending to federated learning. For a full description of the techniques proposed and the results obtained, please refer to the original document.

Our first contribution is FPBoost, a fully parametric gradient boosting algorithm specifically tailored to survival analysis [7], described in Section 2.1. FPBoost optimizes the survival likelihood directly and utilizes ensembles of parametric distributions as base learners, offering enhanced interpretability and predictive performance relative to existing state-of-the-art approaches. Complementing this, in Section 2.2 we explore polynomial-based interpolation methods as post-processing strategies to improve the calibration of neural network survival models with discrete-time predictions [8, 9].

Addressing federated settings, we develop FedSurF, a federated extension of Random Survival Forests [10], designed to aggregate locally trained decision-tree ensembles efficiently within a single aggregation step [11, 12, 13], as outlined in Section 3.1. FedSurF has proven particularly effective in handling heterogeneous and limited datasets across participating institutions, demonstrating robust predictive capabilities in multi-center scenarios. Finally, recognizing the critical need for training data at scale, in Section 3.2 we investigate generative modeling techniques to create realistic synthetic survival datasets [14]. These privacy-preserving datasets facilitate large-scale training, validation, and benchmarking of survival models without compromising sensitive patient information.

Collectively, these contributions advance the field of federated survival analysis by effectively leveraging modern machine learning methodologies in privacy-sensitive environments. Through improved predictive accuracy, calibration, and generalizability, our work supports more precise clinical decision-making and promotes responsible healthcare resource allocation.

2 Single-Client Survival Analysis

In this section, we focus on modeling time-to-event outcomes when data resides in a single, centralized node. We introduce a novel fully parametric gradient boosting method (FPBoost) and an interpolation-based postprocessing step of neural models to enhance calibration. These techniques set the stage for advanced single-client survival modeling before extending to federated scenarios.

2.1 FPBoost: Fully Parametric Gradient Boosting

The performance of survival analysis models heavily influences downstream clinical decisions and resource allocation in healthcare. In recent years, gradient-boosted and tree-based ensemble methods have emerged as state-of-the-art predictive tools in standard regression and classification tasks. However, within survival analysis, our empirical studies highlighted that bagging-based methods—specifically, Random Survival Forests (RSFs) [10]—frequently outperform existing gradient-boosting techniques [15, 16]. We attribute this discrepancy largely to the use of partial-likelihood losses in conventional survival boosting algorithms, which restrict model flexibility and may limit their practical performance relative to assumption-free approaches such as RSFs [7].

Motivated by this observation, we propose FPBoost (Fully Parametric Gradient Boosting), a novel survival model combining gradient boosting, bagging-like model mixtures, and direct optimization of the full survival likelihood. FPBoost circumvents the proportional hazard assumption by representing hazard functions as weighted sums of differentiable survival distributions, such as Weibull and LogLogistic. Thus, FPBoost defines the hazard function for a sample with covariates $\mathbf{x}$ at time t as

$$h\big(t \mid \Theta(\mathbf{x})\big) \;=\; \sum_{j=1}^{J} w_j(\mathbf{x})\, h_j\big(t \mid \eta_j(\mathbf{x}), k_j(\mathbf{x})\big),$$

where $h_j(\cdot)$ denotes the hazard of the j-th parametric head, parameterized by shape $k_j(\mathbf{x})$, scale $\eta_j(\mathbf{x})$, and weight $w_j(\mathbf{x})$, each modeled via ensembles of decision trees. By optimizing these parameters using gradient boosting to minimize the negative log-likelihood of survival data (paired with ElasticNet regularization), FPBoost achieves a flexible yet interpretable hazard representation capable of capturing diverse risk patterns, including bathtub-shaped hazards commonly observed in survival data.

We validated FPBoost's predictive capabilities through extensive experiments on multiple benchmark datasets. FPBoost consistently ranked first or second across nearly all datasets, significantly outperforming semi-parametric methods (e.g., CoxPH, CoxBoost, DeepSurv) and fully parametric neural models (e.g., Deep Survival Machines, DeepHit). On average, FPBoost provided a concordance improvement of approximately 4.6 points (9% relative gain) over baseline approaches.

2.2 Interpolation Techniques for Neural Models

In addition to exploring tree-based methods, we investigated how time-discrete neural survival models such as Logistic Hazard and DeepHit can be enhanced through interpolation. Our motivation stemmed from the observation that discretizing time into a limited set of anchor points yields coarse, stepwise survival estimates and can result in insufficiently precise long-term predictions. We hypothesized that interpolation, which converts discrete outputs into a continuous survival function, could improve clinical utility without imposing significant computational overhead.

To validate this idea, we evaluated a range of interpolation techniques, including piecewise-constant, linear, piecewise-exponential, monotonic splines, and a Kaplan–Meier–based approach. Each non-stepwise method imposes a simple yet effective inductive bias, enforcing smoother probability transitions between adjacent anchor points. By treating the survival function as a continuous curve, these interpolations correct the abrupt drops or jumps common in purely discrete neural outputs. In addition, we found that interpolation combines well with small output layers, allowing the model to learn a lower number of parameters with higher accuracy.

Our experiments on several benchmark datasets, including WHAS, GBSG2, and METABRIC, showed that even a basic method such as linear interpolation can consistently improve integrated metrics like the Brier score (IBS). Kaplan–Meier interpolation sometimes achieved marginally higher gains, particularly in larger datasets where censoring plays a stronger role. Moreover, our studies revealed that networks with fewer output nodes outperformed those with larger layers once interpolation is applied, suggesting an optimal trade-off between neural capacity and subsequent smoothing.

3 Federated Learning for Survival Analysis

In this section, we provide an overview of the studies carried on survival analysis across multiple institutions. We propose FedSurF, which aggregates locally trained random survival forests in a single communication round, and explore generative strategies for synthesizing realistic data while preserving privacy.

3.1 FedSurF: Federated Survival Forests

While previous studies address survival analysis under a single centralized repository, this contribution focuses on adapting tree-based ensembles to decentralized settings through federated learning (FL). Motivated by the difficulty of sharing sensitive clinical data and the intrinsic distribution shifts across medical centers, we developed FedSurF [12, 13], a federated adaptation of Random Survival Forests (RSFs). In essence, each institution trains a local RSF and returns only its most performant trees, as measured by a survival metric such as the concordance index, to a central server. The server then constructs a single global forest by aggregating the trees received in a single communication round, thus avoiding iterative parameter updates common to most FL algorithms.

FedSurF addresses two major issues in federated settings. First, traditional gradient-based methods often require many global rounds, consuming significant bandwidth and computational resources. In contrast, the one-shot procedure of FedSurF procedure allows each participant to build its local forest independently, sharing only a set of trees at the end. Second, data heterogeneity among clients arises naturally when populations differ by region, disease subtype, or treatment protocols, impeding convergence for traditional methods. FedSurF instead, has proven particularly resilient to heterogeneity, by assigning higher weights to trees that prove discriminative in each site domain.

Empirical evaluations on real-world datasets confirm the effectiveness of FedSurF. When clients distribute data in a label-skewed fashion (i.e., different survival patterns across sites) [11], FedSurF consistently attains competitive performance relative to neural approaches such as DeepSurv or DeepHit trained via FedAvg and FedProx.

Taken together, these observations indicate that one-round tree aggregation can be a viable option for survival analysis in decentralized systems where frequent communication is impractical. By combining the advantages of ensemble learners with federated training, FedSurF offers a practical and robust solution for heterogeneous healthcare domains.

3.2 Generative Data in Federated Learning and Survival Analysis

Building upon federated and survival frameworks, these contributions explore the role of generative modeling in both enhancing privacy-preserving collaborative training and mitigating data scarcity. Generative techniques can create synthetic datasets that retain essential statistical properties of real data while concealing sensitive details, thus providing a flexible and secure alternative to conventional federated learning pipelines. Our investigations center on two main contributions.

First, we introduce the Secure Generative Data Exchange (SGDE) protocol [14], which enables cross-silo federations to exchange models that generate synthetic samples rather than sharing gradients or parameters. Each institution independently trains a privacy-preserving generator and pushes it to a central server, which subsequently redistributes these generators to federation members. Because only the final model parameters (rather than continuous gradient exchanges) traverse the network, SGDE markedly reduces exposure to adversarial threats such as membership inference and data poisoning. Experiments with both tabular and image datasets highlight that models trained with synthesized data from SGDE can surpass the performance of standard FedAvg approaches. This demonstrates that exchanging generative models is a viable strategy to protect confidentiality while maintaining robust accuracy.

Second, we propose a filtering pipeline tailored for survival analysis, wherein a survival model trained on real data evaluates newly generated samples and discards the low-quality or unrepresentative ones. This selective filtering approach, which leverages approximate likelihood metrics, refines the synthetic dataset to match or even exceed the performance attained by using real data only. The resulting gains show a clear benefit to employing generative pipelines in survival analysis: they permit scaling datasets to sizes well beyond their original scope, while improving model discrimination.

Overall, these findings illustrate the promise of synthetic data generation in overcoming the dual issues of limited data availability and stringent confidentiality constraints, offering a versatile and practical solution for next-generation survival modeling in federated environments.

4 Conclusion

This thesis tackled the challenge of modeling time-to-event outcomes under stringent privacy constraints, bridging modern survival analysis with federated learning. We introduced (i) FPBoost, a fully parametric gradient boosting algorithm that flexibly models continuous-time hazards; (ii) demonstrated how interpolation enhances calibration in discrete neural survival models; (iii) proposed FedSurF, a one-round federated random survival forest suitable for heterogeneous clinical data; and (iv) we explored generative strategies for synthetic survival data, mitigating both data scarcity and privacy risks. Going forward, efforts will focus on extensions to multimodal or competing-risk scenarios, as well as fine-tuning generative pipelines for censored data. By uniting privacy-aware design with predictive performance, our contributions aim to foster more accurate, fair, and eficient decision support in healthcare.

Acknowledgments

The author expresses his sincere gratitude to his supervisor, Prof. Matteo Matteucci, for his guidance and support. He also acknowledges Professors Giacomo Boracchi, Danilo Ardagna, and Massimiliano Pierobon for their valuable discussions and collaboration. Special thanks are extended to the members of the Ph.D. examination committee: Prof. Luigi De Russis, president of the committee; Professors George Chen and Kevin Xu, thesis reviewers; and Professors Pietro Ducange and Sanjay Purushotham, committee members. The author further thanks his colleagues Eugenio, Diego, and Francesco for their friendship and support. Finally, this research was partially funded by the European Commission under the H2020 grant No. 101016577 (AI-SPRINT), and supported by the FAIR project (Future Artificial Intelligence Research), funded through the NextGenerationEU program (PNRR-PE-AI, M4C2, Investment 1.3).

References

[1] John P Klein, Melvin L Moeschberger, et al. *Survival analysis: techniques for censored and truncated data*, volume 1230. Springer, 2003.

[2] Ping Wang, Yan Li, and Chandan K. Reddy. Machine learning for survival analysis: A survey. *ACM Comput. Surv.*, 51(6), 2 2019.

[3] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):1–7, 2020.

[4] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020.

[5] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2):1–210, 2021.

[6] Lingjuan Lyu, Han Yu, and Qiang Yang. Threats to Federated Learning: A Survey. *arXiv:2003.02133 [cs, stat]*, March 2020. arXiv: 2003.02133.

[7] Alberto Archetti, Eugenio Lomurno, Diego Piccinotti, and Matteo Matteucci. Fpboost: Fully parametric gradient boosting for survival analysis. *arXiv preprint arXiv:2409.13363*, 2024.

[8] Alberto Archetti, Francesco Stranieri, and Matteo Matteucci. Deep survival analysis for healthcare: An empirical study on post-processing techniques. In Francesco Calimeri, Mauro Dragoni, and Fabio Stella, editors, *Proceedings of the 2nd AIxIA Workshop on Artificial Intelligence For Healthcare (HC@AIxIA 2023)*, volume 3578 of *CEUR Workshop Proceedings*, pages 99–121, Rome, Italy, 2023. CEUR-WS.org.

[9] Alberto Archetti, Francesco Stranieri, and Matteo Matteucci. Bridging the gap: improve neural survival models with interpolation techniques. *Progress in Artificial Intelligence*, pages 1–16, 2024.

[10] Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3):841 – 860, 2008.

[11] Alberto Archetti, Eugenio Lomurno, Francesco Lattari, André Martin, and Matteo Matteucci. Heterogeneous datasets for federated survival analysis simulation. In *Companion of the 2023 ACM/SPEC International Conference on Performance Engineering*, pages 173–180, 2023.

[12] Alberto Archetti and Matteo Matteucci. Federated survival forests. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9, 2023.

[13] Alberto Archetti, Francesca Ieva, and Matteo Matteucci. Scaling survival analysis in healthcare with federated survival forests: A comparative study on heart failure and breast cancer genomics. *Future Generation Computer Systems*, 149:343–358, 2023.

[14] Eugenio Lomurno, Alberto Archetti, Lorenzo Cazzella, Stefano Samele, Leonardo Di Perna, and Matteo Matteucci. SGDE: Secure generative data exchange for cross-silo federated learning. In *AIPR 2022, International Conference on Artificial Intelligence and Pattern Recognition*, 2022.

[15] Greg Ridgeway. The state of boosting. *Computing science and statistics*, pages 172–181, 1999.

[16] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery.