Extending the Neural Additive Model for Survival Analysis with EHR Data

Matthew Peroni∗
Operations Research Center, Massachusetts Institute of Technology
Enolink Inc.
Marharyta Kurban
Enolink Inc.
Sun Young Yang
Young Sun Kim
Hae Yeon Kang
Ji Hyun Song
Department of Internal Medicine, Seoul National University Hospital Healthcare System Gangnam Center

Abstract
With increasing interest in applying machine learning to develop healthcare solutions, there is a desire to create interpretable deep learning models for survival analysis. In this paper, we extend the Neural Additive Model (NAM) [1] by incorporating pairwise feature interaction networks and equip these models with loss functions that fit both proportional and non-proportional extensions of the Cox model. We show that within this extended framework, we can construct non-proportional hazard models, which we call TimeNAM, that significantly improve performance over the standard NAM model architecture on benchmark survival datasets. We apply these model architectures to data from the Electronic Health Record (EHR) database of Seoul National University Hospital Gangnam Center (SNUHGC) to build an interpretable neural network survival model for gastric cancer prediction. We demonstrate that on both benchmark survival analysis datasets, as well as on our gastric cancer dataset, our model architectures yield performance that matches, or surpasses, the current state-of-the-art black-box methods.

Keywords: Interpretable ML, Survival Analysis, Neural Additive Model

1. Introduction
Motivated by a series of well-documented failures of machine learning (ML) models being deployed in real-world systems with harmful biases and costly inaccurate predictions [2, 3, 4], and the discovery that post-hoc black-box model explainers are not always truthful to the underlying model [5], there has been significant interest in developing interpretable, white-box machine learning algorithms [6], especially for high-stakes decision making, such as healthcare [7]. While development in this direction has been promising, there still exists a gap in the literature that studies the use of interpretable ML methods, especially deep learning methods, for survival analysis with censored data. Since healthcare datasets are often censored and extremely imbalanced, standard classification and regression models, which is what the majority of recent developments in interpretable ML have been designed for, can be ineffective. A promising approach in this direction is the Neural Additive Model (NAM) introduced in [1], which offers a directly intelli-

© M. Peroni, M. Kurban, S.Y. Yang, Y.S. Kim, H.Y. Kang & J.H. Song.
gable and flexible model that can be easily configured to work with survival data.

In this paper, we extend NAM by incorporating pairwise feature interaction networks and equip these models with loss functions that fit both proportional and non-proportional extensions of the Cox Proportional Hazards (CoxPH) model [8]. We show that not only does this extension allow us to model crucial feature interactions, it also allows us to model the risk contribution of a feature over time. We find that adding this flexibility has a significant impact on model accuracy, while also allowing us to discover temporal patterns within the data. A primary goal of this paper is to demonstrate the ability of interpretable models to achieve accuracy comparable to state-of-the-art black-box models on challenging real-world clinical survival data. In this vein, we apply the extended NAM model architecture to data from the Electronic Health Record (EHR) database of Seoul National University Hospital Gangnam Center (SNUHGC) to build an interpretable survival model for gastric cancer prediction. We demonstrate that on our gastric cancer dataset, the extended NAM model yields performance that matches, or surpasses, the current state-of-the-art black-box survival methods, such as DeepSurv [9] and XGBoost [10].

2. Related Work

This paper builds directly upon the Neural Additive Model architecture given in [1]. To the authors’ knowledge, the only other paper to date that has investigated applying NAM to survival data is [11], which uses a NAM model as a surrogate for a black-box deep learning model that is trained on survival data. Our work differs in that we extend the NAM architecture in order to create interpretable models that can replace black-box models entirely, rather than using them as surrogates. Other interpretable deep learning methods have been developed, such as the Neural Interaction Transparency (NIT) model [12]. The NIT model takes a top-down approach to modeling feature interactions, letting the model determine what interactions to include during the training process. In contrast, our extended NAM approach, outlined in Section 3.1, takes a bottom-up approach, requiring the feature interactions to be identified and specified manually. While the NIT approach is more efficient for discovering useful feature interactions, we find that when working with healthcare survival data, there is a desire to control the feature interactions being modeled. Further, in all experiments we run, only a few interaction pairs are significant enough to include in the final model. Therefore, a bottom-up approach guided by domain experts seems to suit this domain better.

3. Methods

We cover the extended NAM architecture and detail how it can be equipped for survival analysis. For a brief overview of survival data and analysis, see Appendix A.

3.1. Models

We extend the NAM model introduced in [1] by modeling select feature interactions. The extended NAM (NA2M) belongs to a class of models called Generalized Additive Models plus Interactions (GA2M) [13] which take the form \( g(E[y]) = \sum f_i(x_i) + \sum_{(i,j) \in s} f_{ij}^\theta(x_i, x_j) \) where \( g \) is the link function (e.g. identity, logarithm). The NA2M model architecture learns a linear combination of feature networks \( f_i^\theta \) that attend to a single feature \( i \) and feature pair networks \( f_{ij}^\theta \) that attend a pair of features, parameterized by \( \theta \). We denote by \( s \subseteq \{(i,j) : 1 \leq i < j \leq n\} \) the set of feature pairs used by a given model. The networks are trained jointly using backpropagation. Details of model training are given in Appendix B.1. In this work, we fit the NA2M model as the log-relative risk function in the CoxPH model, such that \( h_r(x; \theta) = \sum f_i^\theta(x_i) + \sum_{(i,j) \in s} f_{ij}^\theta(x_i, x_j) \). Allowing the model to capture second-order
pair interactions can improve accuracy and enhance clinical investigations by illustrating the predictive patterns present in the data for feature pair interactions.

A major benefit we found of extending to the NA2M architecture for survival analysis is that by treating the event time $T$ as a feature, we can create a non-proportional Cox-Time model as defined in [14] by learning a time-varying log-relative risk function as $h_r(x, t; \theta) = \sum f^\theta_i(x_i, t) + \sum f^\theta_{ij}(x_i, x_j)$. We refer to this architecture as TimeNAM, or TimeNA2M if feature pairs are included. That is, while both model formats are based on the NA2M model architecture, TimeNAM is comprised solely of feature-interaction nets where each net considers a feature-time interaction, whereas TimeNA2M permits feature-time interactions and, additionally, feature pair interactions, as in the NA2M model. In both cases, we are modeling the risk contribution of each feature over time. As we will demonstrate in the results (4.2), this added expressiveness can lead to a significant boost in accuracy while preserving the interpretability of the model.

4. Results

4.1. Data

We evaluate our models on four benchmark clinical survival datasets: METABRIC [15, 16], SUPPORT [17], WHAS [18], and Rotterdam & GBSG [19, 20, 21]. We use the versions of these datasets produced by [9] without any additional processing.

We also evaluate our models on comprehensive medical annual check-up data, including endoscopic findings and blood test results, collected from 129K+ patients who visited one of the largest medical screening facilities in South Korea. A cancer diagnosis was determined based on the stomach biopsy results of the samples taken during endoscopy procedures. The final cohort consists of 59,540 negative and 248 positive patients. See Appendix C for more details. We refer to this dataset as SNUHGC.

4.2. Experiment Results

In the following, we compare our extended NAM models against the linear Cox model [8], XGBoost [10], and DeepSurv [9] on our benchmark datasets and SNUHGC data. We evaluate the models using the concordance index (c-index) metric, as defined in [22]. The c-index is the standard metric for survival analysis and is the equivalent of the ROC AUC score for survival data. Hyperparameter tuning for each model was conducted using Bayesian Optimization [23], details can be found in appendix B.2. To identify important features pairs for inclusion in the extended NAM models, we use the FAST algorithm formulated in [24] with minor modifications for public datasets and clinical expertise for the SNUHGC dataset. For each dataset, we include at most five feature interactions in the extended NAM models.

4.2.1. Benchmark Experiments

Our main results are summarized in Table 1. In two instances, Metabric and GBSG, we see that the TimeNAM models actually outperform our comparison models. This is possible because TimeNAM is able to consider time as a feature, whereas the XGBoost and DeepSurv models, as originally implemented, are based on the Cox proportional hazards model. These other models could be modified to fit the Cox-Time framework, but our primary goal in this work is to demonstrate the flexibility of the extended NAM models and their ability to match the performance of pre-existing black-box models.

We also observe that some datasets, such as GBSG, have significant pairwise interactions such that the NA2M model outperforms the baseline NAM model, while other datasets, such as Metabric and Support, have significant time-varying risk factors such that the TimeNAM models outperform the baseline NAM model. In the case of GBSG, we see that there are predictive patterns in both the pairwise interactions and between features and time, such that both NA2M and TimeNAM improve upon the baseline, and the combination of Time-
Table 1: C-Index for models across datasets. Means and standard deviations reported from stratified 5-fold cross validation.

| Model     | METABRIC   | SUPPORT   | GBSG       | WHAS       | SNUHGC     |
|-----------|------------|-----------|------------|------------|------------|
| Linear Cox| 63.45 ± 1.19 | 57.04 ± 0.94 | 66.26 ± 1.32 | 74.49 ± 2.63 | 76.29 ± 3.33 |
| NAM       | 64.11 ± 1.26 | 60.50 ± 1.10 | 65.76 ± 2.72 | 74.87 ± 2.08 | 73.4 ± 5.16  |
| NA2M      | 64.44 ± 1.24 | 60.76 ± 0.78 | 67.65 ± 0.96 | 75.64 ± 2.72 | 74.31 ± 3.80 |
| TimeNAM   | 66.79 ± 2.30 | 61.95 ± 0.84 | 67.76 ± 1.37 | 74.58 ± 1.82 | 76.21 ± 1.95 |
| TimeNA2M  | 66.07 ± 2.26 | 62.21 ± 0.73 | 68.25 ± 1.44 | 75.51 ± 2.09 | 76.62 ± 3.00 |
| DeepSurv  | 64.98 ± 0.80 | 60.97 ± 1.19 | 67.21 ± 0.98 | 80.94 ± 1.15 | 76.26 ± 3.36 |
| XGBoost   | 64.65 ± 1.34 | 61.97 ± 0.98 | 67.47 ± 1.81 | 86.99 ± 0.98 | 76.50 ± 1.67 |

NAM plus feature interactions yields an even better model.

The one outlier in these results is the WHAS dataset, where none of the NAM based models perform much better than the linear Cox model or come close to the black-box models. Upon further investigation we discovered that, as noted in [14], the dataset is a case-control dataset, such that there are duplicate records of patients in the data. Therefore, it is possible that without the constraints placed on the NAM models, the DeepSurv and XGBoost models are more capable of memorizing data that is present in both the training and test sets. Indeed, we can see in Table 3 in Appendix B.2 that the hyperparameter search for XGBoost on the WHAS dataset selected the highest possible value for the number of estimators, suggesting that overfitting to the training data (since boosted trees are trained on the residual) was beneficial in this case.

4.2.2. SNUHGC Experiments

The last column of Table 1 summarizes the model performance on the SNUHGC dataset. The table indicates that a simple NAM model has the lowest performance due to its inability to model feature interactions. Adding feature pairs improved the performance slightly, but the performance was still lower than that of the benchmark models. The addition of the time component along with the feature interactions(TimeNA2M) yielded performance equivalent to DeepSurv and XGBoost.

Figure 1 provides an illustrative example visualization of the TimeNA2M output. It shows the risk contribution of HbA1c and indicates that poor glycaemic control may increase the risk of gastric cancer over time, which is in line with the experimental findings [25].

![HbA1c risk contribution over time for SNUHGC. The Plot depicts the exact output from the feature net that models the risk contribution of HbA1c over time in the TimeNA2M model fit to this dataset.](image)
pairs selected by PIs in the last column: gender - age, age - h. Pylori, h. Pylori - atrophic gastritis, diet sodium - gastritis. It is worth noting that not all plots can be explained easily. For example, the figures show that higher salt intake is associated with protective properties, which contradicts the research [26]. Such a mismatch could be explained by the fact that patients’ sodium intake data were self-reported and, thus, might be unreliable. Another explanation is that patients with gastric conditions tend to adjust their salt intake. Thus, reduced sodium intake could be a proxy for the presence of other conditions. However, the model correctly identifies atrophic gastritis, intestinal metaplasia, and h. Pylori infection as the main contributing factors as illustrated by the feature importance plot (Figure 7) and gets the direction of their effect correct. It also correctly identifies most lifestyle factors such as smoking and drinking. From this starting point, however, we can work with clinicians to understand the data and refine the model.

As the model is solely reflective of the patterns present in this dataset, which is also highly imbalanced, we observe some of the feature (pair) networks produce risk functions that could not be explained by domain expertise, discussed further in Appendix C. However, since these NAM models can be easily modified with the guidance of domain knowledge and the plots (such as 1) are an exact representation of the model, we believe these models provide an advantage over the black-box models.

4.2.3. METABRIC CASE STUDY

To further illustrate the interpretability of these extended NAM models, we provide the TimeNA2M model visualization for METABRIC in Figure 3. The METABRIC dataset consists of nine features total, including age, four gene and protein expression features (MKI67, EGFR, PGR, ERBB2), and four relevant clinical indicator variables (hormone treatment, radiotherapy, chemotherapy, and ER positive) for patients with breast cancer. There are 1,980 patients in the dataset, of which 57.72 percent died during the study. As discussed in section 4.2, there are predictive patterns in the relationship between features and time that boost the performance of the TimeNAM models. We can see this reflected in the feature function plots, as the relative risk contribution of EGFR, ERBB2, radiotherapy, chemotherapy, and being ER-Positive all clearly vary over time. The included feature pairs do not significantly contribute to the performance of the model, but we do observe some clear interaction between radiotherapy and chemotherapy. In this case, the modeled interaction tells us that to the model, a patient undergoing chemotherapy treatments without radiotherapy indicates the patient is at higher risk of death than if they were treated with only radiotherapy or with both treatments. While the model has no notion of causality, this interpretability can be very useful for collaborations with clinicians with domain knowledge. As mentioned in 4.2, the plots in Figure 3 exactly represent the model, so the model itself can be easily modified by essentially modifying these plots.

5. Discussion and Future Work

In this work, we demonstrated that interpretable, white-box models are capable of achieving performance comparable to black-box models for survival analysis. The modularity, flexibility, and ability to make precise modifications to these extended NAM models enable meaningful collaboration with domain experts where all parties can understand the model and suggest pinpoint modifications. Going forward we plan to continue our work by: (1) Comparing the extended NAM model visualizations to SHAP explanations on the same clinical datasets (we begin to explore this in Appendix C.1 and D). (2) Exploring other modifications to NAM, such as the recently published Neural Basis Model [27]. (3) Evaluating these models on additional datasets.
Figure 2: TimeNA2M feature network outputs for SNUHGC. Each plot corresponds to an individual feature or feature pair network.

Figure 3: TimeNA2M feature and feature pair network outputs for METABRIC.
References

[1] Agarwal, R. et al. Neural Additive Models: Interpretable Machine Learning with Neural Nets. in Advances in Neural Information Processing Systems (eds. Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. S. & Vaughan, J. W.) vol. 34 4699–4711 (Curran Associates, Inc., 2021).

[2] Wexler, R. When a computer program keeps you in jail: how computers are harming criminal justice. New York Times (13 June 2017); https://www.nytimes.com/2017/06/13/opinion/how-computers-are-harming-criminal-justice.html

[3] McGough, M. How bad is Sacramento’s air, exactly? Google results appear at odds with reality, some say. Sacramento Bee (7 August 2018)

[4] Varshney, K. R. & Alemzadeh, H. On the safety of machine learning: cyberphysical systems, decision sciences and data products. Big Data 10, 5 (2016).

[5] Kumar, I.E., Venkatasubramanian, S., Scheidegger, C. & Friedler, S. (2020). Problems with Shapley-value-based explanations as feature importance measures. Proceedings of the 37th International Conference on Machine Learning, in Proceedings of Machine Learning Research 119:5491-5500 Available from https://proceedings.mlr.press/v119/kumar20e.html.

[6] Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1, 206–215 (2019).

[7] Lengerich, B. J., Caruana, R., Nunnally, M. E. & Kellis, M. Death by Round Numbers and Sharp Thresholds: How to Avoid Dangerous AI EHR Recommendations. (2022) doi:10.1101/2022.04.30.22274520.

[8] Cox, D. R. Regression Models and Life-Tables. Journal of the Royal Statistical Society. Series B (Methodological) 34, 187–220 (1972).

[9] Katzman, J. L. et al. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Medical Research Methodology 18, 24 (2018).

[10] Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 785–794 (Association for Computing Machinery, 2016). doi:10.1145/2939672.2939785.

[11] Utkin LV, Satyukov ED, Konstantinov AV. SurvNAM: The machine learning survival model explanation. Neural Netw. 2022 Mar;147:81-102. doi: 10.1016/j.neunet.2021.12.015. Epub 2021 Dec 29. PMID: 34995952.

[12] Tsang, M., Liu, H., Purushotham, S., Murali, P.& Liu, Y. Neural Interaction Transparency (NIT): Disentangling Learned Interactions for Improved Interpretability. in Advances in Neural Information Processing Systems (eds. Bengio, S. et al.) vol. 31 (Curran Associates, Inc., 2018).

[13] Lou, Y., Caruana, R., Gehrke, J. & Hooker, G. Accurate intelligible models with pairwise interactions. in Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining 623–631 (Association for Computing Machinery, 2013). doi:10.1145/2487575.2487579.

[14] Kvamme, H., Borgan, Ø. & Scheel, I. Time-to-Event Prediction with Neural Networks and Cox Regression. Journal of Machine Learning Research 20, 1–30 (2019).

[15] Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, Speed D,
Lynch AG, Samarajiwa S, Yuan Y, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012;486(7403):346–52.

[16] Lakhanpal R, Sestak I, Shadbolt B, Bennett GM, Brown M, Phillips T, Zhang Y, Bullman A, Rezo A. IHC4 score plus clinical treatment score predicts locoregional recurrence in early breast cancer. The Breast. 2016;29:147–52.

[17] Knaus WA, Harrell FE, Lynn J, Goldman L, Phillips RS, Connors AF, Dawson NV, Fulkerson WJ, Califf RM, Desbiens N, et al. The support prognostic model: objective estimates of survival for seriously ill hospitalized adults. Ann Intern Med. 1995;122(3):191–203.

[18] Hosmer DW, Lemeshow S, May S. Applied Survival Analysis: Regression Modeling of Time to Event Data. 2nd ed. New York: Wiley-Interscience; 2008.

[19] Foekens JA, Peters HA, Look MP, Portengen H, Schmitt M, Kramer MD, Brünner N, Jänicke F, Meijer-van Gelder ME, Henzen-Logmans SC, et al. The urokinase system of plasminogen activation and prognosis in 2780 breast cancer patients. Cancer Res. 2000;60(3):636–43.

[20] Schumacher M, Bastert G, Bojar H, Huehner K, Olschewski M, Sauerbrei W, Schmoor C, Beyerle C, Neumann R, Rauschecker H. Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. German breast cancer study group. J Clin Oncol. 1994;12(10):2086–93.

[21] Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med. 2000;19(4):453–73.

[22] Frank E, Harrell Jr, Robert M, Califf, David B. Pryor, Kerry L. Lee, and Robert A. Rosati. Evaluating the yield of medical tests. Journal of the American Medical Association, 247 (18):2543–2546, 1982.

[23] Snoek, J., Larochelle, H. & Adams, R. P. Practical Bayesian Optimization of Machine Learning Algorithms. http://arxiv.org/abs/1206.2944 (2012) doi:10.48550/arXiv.1206.2944.

[24] Lou, Y., Caruana, R., Gehrke, J. & Hooker, G. Accurate intelligible models with pairwise interactions. in Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining 623–631 (Association for Computing Machinery, 2013). doi:10.1145/2487575.2487579.

[25] Zheng, Jiaojiao, et al. ”Haemoglobin A1c and serum glucose levels and risk of gastric cancer: a systematic review and meta-analysis." British journal of cancer 126.7 (2022): 1100-1107.

[26] Yusefi, Ali Reza, et al. "Risk factors for gastric cancer: a systematic review." Asian Pacific journal of cancer prevention: APJCP 19.3 (2018): 591.

[27] Radenovic, F., Dubey, A. & Mahajan, D. Neural Basis Models for Interpretability. (2022) doi:10.48550/arXiv.2205.14120.

[28] John P. Klein and Melvin L. Moeschberger. Survival Analysis: Techniques for Censored and Truncated Data. Springer, New York, 2. edition, 2003.

[29] Davidson-Pilon, (2019). lifelines: survival analysis in Python. Journal of Open Source Software, 4(40), 1317, https://doi.org/10.21105/joss.01317

[30] Lundberg, Scott M., and Su-In Lee. A unified approach to interpreting model predictions. Advances in neural information processing systems 30 (2017).
Appendix A. A Brief Summary of Relevant Survival Analysis

A.1. Survival Data

We consider survival data of the form \((x, T, E)\) where \(x = (x_1, x_2, \ldots, x_K)\) is the input covariates with \(K\) features, \(T\) is a failure event time, and \(E\) is an event indicator. When \(E = 1\), this indicates that an event (e.g. death of patient, development of disease, etc.) has occurred at time \(T\) from the beginning of the study. When \(E = 0\), this indicates that no event has occurred by the last recorded time \(T\), which can be the end of the study, or the last contact with the patient. The latter is considered right-censored data, and is generally considered missing data when applying standard regression or classification techniques. Given the prevalence of censored data in real-world clinical datasets, ignoring this data can significantly bias the model. Therefore, we want to use methods that incorporate this censored data, which forms the basis of survival analysis.

A.2. Survival Analysis

We give a brief description of the essential components of survival analysis for our work, see [28] for a complete overview. The two essential components of survival analysis are the survival function, \(S(t) = Pr(T > t)\), representing the probability that an individual will 'survive' beyond time \(t\), and the hazard function \(h(t)\), given by

\[
h(t) = \lim_{\delta \to 0} \frac{Pr(t \leq T < t + \delta | T \geq t)}{\delta}
\]

which represents the probability an individual will not 'survive' beyond time \(T\), given that they have already 'survived' up until that time. It is known that we can relate the cumulative hazard function, \(H(t) = \int_0^t h(s)ds\) to the survival function \(S(t) = \exp[-H(t)]\). We base our work on two models of hazard function: (1) the Cox Proportional Hazards Model (CoxPH) [8], which represents the hazard function as

\[
h(t|\mathbf{x}) = h_0(t) \exp[h_r(\mathbf{x})]
\]

where \(h_0(t)\) is a non-parametric baseline hazard function, and \(\exp[h_r(\mathbf{x})]\) is the relative risk function. (2) The non-proportional Cox Time Model (Cox-Time) [14] which represents the hazard function as

\[
h(t|\mathbf{x}) = h_0(t) \exp[h_r(\mathbf{x}, t)]
\]

where the relative risk function can depend on time.

Appendix B. Supplementary Materials for Model Evaluation

B.1. Model Training

We train our models on datasets of the form \(\{x^{(i)}, E^{(i)}, T^{(i)}\}_{i=1}^N\) with \(N\) individuals. Suppose we have \(M\) total feature and feature-pair networks in the model. We train the model using an extension of the NAM loss function derived in [1] given by

\[
\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{x}, t, e \sim D}[l(\mathbf{x}, t, e; \theta) + \lambda_1 \eta(\mathbf{x}; \theta) + \lambda_2 \gamma(\theta)]
\]

where

\[
\eta(\mathbf{x}; \theta) = \frac{1}{M} \sum_x \left( \sum_k (f_k^\theta(x_k))^2 + \sum_{(i,j) \in s} (f_{ij}^\theta(x_i, x_j))^2 \right)
\]

is the output regularization and \(\gamma(\theta)\) is the weight decay. In our work, we let \(l(\cdot; \theta)\) be either the loss function derived in [14] for the CoxPH model

\[
\frac{1}{n} \sum_{i: E^{(i)} = 1} \log (\sum_{j \in \tilde{R}^{(i)}} \exp[h_r(\mathbf{x}^{(i)}; \theta) - h_r(\mathbf{x}^{(j)}; \theta)])
\]

where \(h_r(\mathbf{x}; \theta) = \sum f_i^\theta(x_i) + \sum f_{ij}^\theta(x_i, x_j)\) and \(n\) is the number of individuals with a recorded
event, or the loss function for the Cox-Time model

\[ l(;\theta) = \frac{1}{n} \sum_{i: E(i) = 1} \log(\sum_{j \in R(i)} \exp[h_r(x^{(i)}, T^{(i)}; \theta) - h_r(x^{(i)}, T^{(i)}; \theta)]) \]

where \( h_r(t, x; \theta) = \sum f^0_i(x_i, t) + \sum f^0_{ij}(x_i, x_j) \).

It is also worth briefly discussing the relative training time of this model compared to our benchmark models. The standard linear Cox proportional hazards model is the most efficient model, and can solve with \( n = O(1000) \) records and \( p = O(10) \) features in \( O(1s) \). Similarly, XGBoost is very efficient, and can converge in \( O(1s) \) for the public datasets considered in this work, and scales well as \( n \) and \( p \) increase. The neural network models considered do take longer to train, taking between \( O(1s) \) to \( O(1min) \) to train on the public datasets we used, depending on the hyperparameters and size of the problem. It is also important to note that for the non-proportional time models (TimeNAM and TimeNA2M), predictions are more computationally expensive than in the proportional case. In our experience however, the prediction run-time for TimeNAM is still fast (\( O(10ms) \)). For most applications in health care, training time on the order of minutes is still acceptable, and we also note that the NAM-based models are easily modifiable after training, whereas other models such as XGBoost would require making modifications to the training data or hyperparameters and retraining the model.

**B.2. Hyperparameters**

**Extended NAM.** We trained all the extended NAM models with a batch size of 256. We identify one set of parameters for each dataset, and use these parameters across variations of the extended NAM architecture. We tune the dropout in the set \( \{0, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5\} \), feature dropout in the set \( \{0, 0.05, 0.1, 0.2, 0.3\} \), weight decay in the interval \( [1e-6, 1e-4] \), output regularization in the interval \( [0.001, 0.1] \), and learning rate in the interval \( [0.001, 0.1] \). We tune the hidden sizes (for each feature or feature pair network) in the set \( \{[16, 64, 128], [256], [512], [64, 32], [32, 16, 8]\} \) and activation functions in the set \{relu,elu,sigmoid\}. Table 2 gives the hyperparameters used in our model evaluation.

**DeepSurv.** We use the open-source implementation of DeepSurv and use the parameter values identified in [9] for the benchmark datasets.

**XGBoost** We use XGBoost with the objective parameter "survival:cox" that fits the model such that the output is the log-relative risk function \( h_r \). We tune the \( eta \) parameter in the set \( \{0.001, 0.005, 0.01\} \), the \( max depth \) parameter in the set \( \{3, 4, 5\} \), the \( subsample \) parameter in the set \( \{0.5, 0.7, 0.9\} \), the \( num estimators \) parameter in the set \( \{100, 200, 300, 400, 500\} \), the \( min child weight \) parameter in the set \( \{1, 5, 10\} \), the \( gamma \) parameter in the set \( \{0, 0.5, 1\} \), the \( colsample bytree \) parameter in the set \( \{0.5, 0.7, 0.9\} \), and the \( alpha \) and \( lambda \) parameters in the set \( \{0, 0.5, 1\} \). The \( tree method \) is set to either 'auto' or 'hist'. Table 3 gives the hyperparameters used in our model evaluation.

**Linear Cox** We use the open source implementation of the Cox linear regression model provided by the lifelines python package [29] without modification.

**Appendix C. Gastric Cancer Case Study**

Our data was extracted from the Electronic Health Record (EHR) database of Seoul National University Hospital Gangnam Center (SNUHGC), which is one of the largest medical check-up centers in South Korea. Patients in our study cohort were selected from 129,223 predominantly healthy individuals who visited the center for a regular annual check-up between 2007 and 2020. Patients underwent var-
ious clinical tests such as physical examination, blood tests, gastroscopy, and, in some cases, biopsy. All patients completed a self-administered structured questionnaire about their smoking history, alcohol intake, eating habits, family history, and past medical history. A cancer diagnosis was determined based on the stomach biopsy results of the samples taken during endoscopy procedures.

Our cohort selection process includes several steps. First, we filtered out patients who had fewer than two hospital visits. Then, we excluded patients with diagnoses determined by PIs: total gastrectomy, squamous cell carcinoma, neuroendocrine tumor, GIST, and lymphoma/MALToma. Lastly, we excluded patients outside the age range of interest (18-90 years.) The final cohort consists of 59,540 negative and 248 positive patients.

### C.1. SNUHGC SHAP Experiments

Many post-hoc black-box model explanation tools have been developed recently, and SHAP (SHapley Additive exPlanation) is one of the most common approaches [30]. In this section, we use the SHAP library to approximate SHAP values for the interpretable TimeNA2M model we trained on SNUHGC data and produce visualizations. Since we know the ‘ground-truth’ for how this model makes predictions, we can

---

**Table 2: Parameters used for extended NAM models**

| Hyperparameter/Dataset | WHAS | SUPPORT | Metabric | GBSG | SNUHGC |
|------------------------|------|---------|----------|------|--------|
| Dropout                | 0    | 0       | 0.05     | 0    | 0.2    |
| Feature Dropout        | 0    | 0.3     | 0        | 0    | 0      |
| Weight Decay ($\lambda_2$) | 1e-6 | 3.82e-6 | 1e-6     | 1e-4 | 1e-6   |
| Output Reg. ($\lambda_1$) | 0.001 | 0.0466  | 0.001    | 0.001 | 0.0297 |
| Activation             | relu | relu    | relu     | relu | relu   |
| Hidden Size            | 512  | 16      | 64.32    | 128  | 32.32  |
| Learning Rate          | 0.001| 0.001   | 0.001    | 0.001| 0.022  |

**Table 3: Parameters used for XGBoost models**

| Hyperparameter/Dataset | WHAS | SUPPORT | Metabric | GBSG | SNUHGC |
|------------------------|------|---------|----------|------|--------|
| Eta                    | 0.01 | 0.05    | 0.01     | 0.001| 0.01   |
| Max Depth              | 5    | 5       | 4        | 5    | 5      |
| Subsample              | 0.9  | 0.7     | 0.7      | 0.9  | 0.5    |
| Num. Estimators        | 500  | 200     | 300      | 200  | 200    |
| Min Child Weight       | 1    | 10      | 10       | 10   | 10     |
| Gamma                  | 0    | 0       | 1        | 0    | 0      |
| Col. Sample By Tree    | 0.9  | 0.5     | 0.7      | 0.5  | 0.5    |
| Alpha                  | 0    | 1       | 0        | 0    | 0      |
| Lambda                 | 0    | 0.5     | 0        | 0    | 1      |
| Tree Method            | auto | hist    | hist     | auto | auto   |
assess the quality of the SHAP feature attributions.

SHAP explains the prediction for any instance as a sum of contributions from its features. The approach was primarily inspired by game theory. The Shapley value of a specific feature is its contribution to the model prediction, weighted and summed over all possible feature value combinations:

$$
\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[ f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_{S}(x_{S}) \right]
$$

where $S$ represents all possible subsets of the full feature set $F$ that do not contain feature $i$.

As the formula indicates, the exact calculation of SHAP values is computationally challenging. Lundberg and Lee (2017) [30] describe two approximation approaches, Shapley sampling values and Kernel SHAP, which are based on the LIME approach.

SHAP summary plot combines feature importance with feature effects. Each point on the summary plot is a Shapley value for a feature and an instance. The position on the $y$-axis is determined by the feature and on the $x$-axis by the corresponding Shapley value. The resulting plot for the TimeNA2M model created for the SNUHGC data is given in Figure 4.

The plot indicates that 4 out of 5 top features are correctly identified by SHAP. The direction of the effect is also the same for most features. However, by introducing the time component, we can better stratify and analyze the feature effect, such as in the example of HbA1c.

The dependence plots, shown in Figure 5, require further interpretation. For example, the effect of the gender seems to flip around the mean age, which is not observed in the ground truth of the underlying model. This example demonstrates that interpretability and post hoc explainability can diverge, highlighting the importance of white box methods.

**Appendix D. METABRIC Case Study SHAP Experiments**

We tried using SHAP to generate post-hoc explanations of the TimeNA2M model trained on the METABRIC dataset. We used the Permutation SHAP explainer, which is a model agnostic explainer that approximates the Shapley values by iterating over random permutations of the given data. This method was used for efficiency, as the exact SHAP explainer took unreasonably long to fit. The summary plot for the resulting explainer can be seen in Figure 6. We can observe that the explainer identifies duration as the most significant feature by an order of magnitude. While the explainer accurately captures the effect that as duration increases, a patient’s risk increases, this is a fairly uninformative result, as it matches general intuition. Further, it seems to mask the contributions of other features. If one did not have an intuition regarding the meaning of the given features, or how the features ought to impact predictions, it is easy to imagine a scenario where one would conclude that the duration feature is the only significant feature in determining risk. The explainer also seems to indicate that, for example, chemotherapy does not have much impact on the risk prediction under this model. In Figure 3, where we see the ground truth regarding how the model uses the chemotherapy feature, we can see a more complicated story, where the model identifies a patient receiving chemotherapy soon after their diagnosis as increasing their risk, while it seems to decrease their risk in later stages. This illustrates how post-hoc explainers can obscure the actual patterns learned by a model.
Figure 4: SHAP summary plot for TimeNA2M on SNUHGC dataset.

Figure 5: SHAP dependence plots for TimeNA2M on SNUHGC dataset.
Figure 6: SHAP Summary for TimeNA2M Model on METABRIC dataset.

Figure 7: TimeNA2M feature importance for SNUHGC.