Deep Neural Networks for Predicting Recurrence and Survival in Patients with Esophageal Cancer After Surgery

Yuhan Zheng[1](✉)[0000−0002−9762−6530], Jessie A Elliott[2], John V Reynolds[2], Sheraz R Markar[3], Bartłomiej W. Papież[1](✉)[0000−0002−8432−2511], and ENSURE study group[4]

[1] Big Data Institute, University of Oxford, Oxford, United Kingdom
yuhan.zheng@univ.ox.ac.uk, bartlomiej.papiez@bdi.ox.ac.uk
[2] Trinity St. James's Cancer Institute, Trinity College Dublin and St. James's Hospital, Dublin, Ireland
[3] Nuffield Department of Surgery, University of Oxford, United Kingdom
[4] Young Investigator Division, European Society for Diseases of the Esophagus

Abstract. Esophageal cancer is a major cause of cancer-related mortality internationally, with high recurrence rates and poor survival even among patients treated with curative-intent surgery. Investigating relevant prognostic factors and predicting prognosis can enhance post-operative clinical decision-making and potentially improve patients' outcomes. In this work, we assessed prognostic factor identification and discriminative performances of three models for Disease-Free Survival (DFS) and Overall Survival (OS) using a large multicenter international dataset from ENSURE study. We first employed Cox Proportional Hazards (CoxPH) model to assess the impact of each feature on outcomes. Subsequently, we utilised CoxPH and two deep neural network (DNN)-based models, DeepSurv and DeepHit, to predict DFS and OS. The significant prognostic factors identified by our models were consistent with clinical literature, with post-operative pathologic features showing higher significance than clinical stage features. DeepSurv and DeepHit demonstrated comparable discriminative accuracy to CoxPH, with DeepSurv slightly outperforming in both DFS and OS prediction tasks, achieving C-index of 0.735 and 0.74, respectively. While these results suggested the potential of DNNs as prognostic tools for improving predictive accuracy and providing personalised guidance with respect to risk stratification, CoxPH still remains an adequately good prediction model, with the data used in this study.

Keywords: Esophageal Cancer · Survival · Recurrence · Deep Neural Networks · Early Intervention · Patient Stratification

1 Introduction

Esophageal cancer is a major cause of cancer-related mortality internationally. The average 5-year Overall Survival (OS) rate is less than 25% [1], ranging

from 10% to 55% depending on the stage of which the disease is detected [2]. While surgical resection, known as esophagectomy, remains the primary treatment for esophageal cancer, the prognosis of post-operative patients remains poor. Despite advancements in cancer management strategy, more than 50% of the patients experience a recurrence within 1-3 years following curative-intent surgery [3], with a median survival time of 24 months [4]. Therefore, identifying prognostic factors associated with a higher risk of recurrence, as well as predicting and stratifying patients based on their recurrence and survival probabilities, are crucial to the delivery of personalised medicine approaches that could potentially improve oncologic outcomes. Current risk stratification methods for patients with esophageal cancer predominantly rely on pathological data, primarily tumor staging [5]. This does not fully leverage all available clinical and patient-level data efficiently, and does not account for individual variations.

To address these issues, some studies have developed models for prognosis prediction. For example, logistic regression models have been employed to predict absolute risks for patients with esophageal cancer [6, 7]. However, these models predict a single-point outcome event without incorporating time-to-event analysis and are limited to one histologic type only. The Cox Proportional Hazards model (CoxPH) [8] is a widely used regression model that allows the study of the relationships between time-to-event outcomes and a set of covariates. Many studies have employed CoxPH to identify prognostic factors for different outcomes [9–11]. However, CoxPH model assumes linear relationships between covariates and that the relative hazard remains constant over time. This hinders its ability to capture higher level interactions between variables and outcomes.

Recent developments in AI have led to increased applications of machine learning (ML) models in oncology to address more complex problems. For example, Zhang et al. [13] explored multiple ML methods for survival prediction in squamous cell carcinoma, and demonstrated that while CoxPH model remains sufficiently good for interpretive studies, ML approaches have the potential to enhance predictive accuracy. Gong et al. [14] explored artificial neural networks (ANNs) in survival prediction, though these did not outperform other traditional ML models such as XGBoost [15]. However, most of these aforementioned studies relied on data collected from a single center, raising questions about their generalisability and robustness when applied to larger multicenter cohorts. Most studies focus on only one type of outcome, and the prediction values on other outcomes remain unknown. Moreover, these studies often utilize a limited number of features. There is a significant clinical interest in incorporating a more comprehensive set of features that take account into, for example, improvements in treatment technologies or surveillance strategies. Gujjuri et al. [12] implemented CoxPH and Random Forest using ENSURE dataset. However, the results showed that Random Forest did not surpass CoxPH in both discrimination and calibration.

In this work, we developed models to predict Disease-Free Survival (DFS) and OS for patients with esophageal cancer following curative-intent surgery. The work is divided into two main components. The first component is prognostic

factor identification task, which aims to identify significant prognostic factors that influence outcomes based on their hazard ratios and significance values, thereby providing clinical guidance. The second component is a prediction task, which aims to develop robust models for prognosis prediction on multicenter heterogeneous dataset. This helps stratify patients based on their risks, which could potentially facilitate personalisation of postoperative treatment and surveillance strategies.

Our contributions are threefold. Firstly, we developed models using a large heterogeneous multicenter cohort [16]. Secondly, we incorporated a comprehensive set of easily accessible and readily identifiable features into the models, including several general center-specific features, to explore more broadly prognostic factors. Finally we carried out extensive experiments with deep neural network (DNN)-based models, and compared their predictive performance with CoxPH model.

The remainder of this paper is organised as follows. Section 2 introduces the details of the dataset, preprocessing steps and provides an overview of the final dataset used for this work. The three models employed and the experimental setup which includes training and implementation details, are described in Section 3.1 and Section 3.2, respectively. Section 4.1 and Section 4.2 present results for the prognostic factor identification task and prediction task. Finally, the discussion and conclusion can be found in Section 5.

2 Dataset and Preprocessing

2.1 Dataset

This work is based on data collected from the European iNvestigation of SUrveillance after Resection for Esophageal cancer (ENSURE) study [16], a retrospective non-interventional study taken across 20 European centers. Patients with esophageal or junction cancer undergoing curative intent treatment from June 2009 to June 2015 were all considered for inclusion. In total, there are 4972 patients and over 170 variables. All patients were staged according to the 8th edition of the American Joint Committee on Cancer (AJCC) staging [17].

The use of the dataset and this study has been approved by he Joint Research Ethics Committee of Tallaght University Hospital and St. James's Hospital, Dublin, Ireland (SJH-TUH JREC Ref 2943 Amendment 1).

2.2 Outcome Variable Definition

In this work, DFS is defined as the time from treatment (i.e., surgery) to recurrence or death from any cause [18]. Patients who are lost to follow-up or remain alive without recurrence at the end of the study are recorded as censored events. OS is defined as the period from diagnosis to death from any cause [19]. Patients that are lost to follow-up or still alive at the end of the study are recorded as censored event.

2.3 Patient Inclusion and Variable Selection Criteria

Patient Inclusion. In this work, we removed patients with missing DFS and/or OS outcome, as well as patients with rare histologic type (i.e., non-adenocarcinoma and non-squamous cell carcinoma). We excluded further patients with postoperative death for DFS prediction by definition.

Variable Selection. Variables used in our models were selected by experienced clinicians, based on the literature review and their clinical importance. Variables exhibiting clinically known high correlations with other variables, lacking well-established relationships with outcomes, or variables that were often poorly documented by centers, were excluded from the study. Additionally, while there is no single acceptable threshold for missing rate, the approach to dealing with missingness requires careful consideration. Blindly applying imputation strategies to variables with high missing rate could also impose biases [20]. Therefore, after further assessment by clinicians, a set of variables was additionally removed based on both their rate of missingness and their clinical relevance.

In this work, we did not apply any ML or statistic-based variable selection strategies. Evidence [30] suggests that feature selection prior to model application does not significantly improve model performance, especially that we either adopted regularisers in the model (more details in Section 3.2) or the ML models themselves have internal feature selection capabilities to handle high-dimensional data in this study. As a result of this variable selection processes, 37 variables were selected with a missing rate of less than 30%.

2.4 Missingness and Imputation

In this study, the missingness mechanism was assumed to be Missing At Random (MAR) [21], as whether the data is missing or not depends exclusively on their availability at center during data collection process [22, 23]. This assumption allows us to apply imputation strategies to handle missingness. A flow chart illustrating the overall process, which is going to described below, can be found in Figure 1 in Section A.1.

Different imputation strategies were applied to the prognostic factor identification task and prediction task that were mentioned in Section 1. Multiple Imputation by Chained Equations (MICE) [24] was used for prognostic factor identification task, with 10 iterations per imputation set. Multiple imputation (MI), which takes the uncertainty of imputation into account and fills different multiple plausible values, is important to reduce bias and chance of false-positive and false-negative conclusions particularly in interpretation tasks [25]. The multiple imputed datasets were passed into models, optimised and analysed separately, and final results were combined using Rubin's rule [26]. For prediction task, where the impact of imputation uncertainty is generally less critical, we used single-point multivariate imputation by chained equations, which is typically sufficient for predictive modeling purposes.

When performing imputation, outcome variables, including the binary event indicator and the time-to-event variable, were also included in the prediction matrix to prevent bias [27]. The time-to-event variable was transformed to its cumulative hazard function with the non-parametric Nelson-Aalen estimator [28] as suggested in [29]. The imputation was conducted within the cross-validation (CV) loop during training [30] to prevent any information leakage from the validation set into the training process.

Prior to imputation, the nominal categorical variables were dummy coded. It is important to note that during the imputation process, continuous values were generated for all dummy-coded binary variables, and these values were not rounded to the nearest integer, as recommended based on the findings in [31]. Additionally, continuous numerical variables were scaled by zero-score standardisation to bring all variables to approximately similar dynamic ranges to improve numerical stability during training.

Table 1: Summary of the dataset used for model development. DFS task (n=3921), OS task (n=4077).

Outcome	No. of Variables	No. of Patients	No. of Observed Events	Min. (months)	Max. (months)	Median (months)	Mean (months)
DFS	34	3921	2308	0	173	29.7	36.1
OS	34	4077	2173	0.2	176.7	37.47	41.92

Furthermore, after standardisation, three variables that had Pearson correlation coefficients higher than 70% were removed. While there is no definitive threshold for exclusion, we set this threshold based on the interpretations provided in [32] and common practices in the field. As a result, 34 variables were ultimately selected for model development.

2.5 Data Overview

Table 1 summarises the statistics for the dataset used in DFS and OS tasks, respectively.

3 Methods and Experiments

3.1 Models

In this work, three models were employed to predict DFS and OS: a regression model CoxPH [8] and two neural network-based models named DeepSurv [33] and DeepHit [34]. CoxPH is a semi-parametric regression model that takes the form $h_0(t)exp(\sum_i x_i \cdot \beta_i)$, where $h_0(t)$ is baseline hazard function, x_i is covariate

and β_i is coefficient. The model assumes that the effect of a factor is constant over time and there is a linear relationship between predictors and log-hazards. DeepSurv is a DNN-based extension of CoxPH model. It models the hazard function as $h_0(t)exp(f_{\boldsymbol{\theta}}(\boldsymbol{x}))$, where $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ is a neural network that takes covariates as input and outputs a scalar. This allows DeepSurv to capture high-level interactions among features. DeepHit, on the other hand, employs an end-to-end DNN that learns the distribution of survival times directly, without making any assumptions about the underlying stochastic process.

In this work, CoxPH model was employed for the prognostic factor identification task. For the prediction task, all three models were used, with CoxPH serving as a baseline for comparison with neural network-based methods. These models were chosen to leverage their respective strengths in handling different aspects of survival analysis, from traditional regression assumptions to capturing complex interactions and learning distributions directly from data.

3.2 Experimental Setup

Dataset Splitting Strategy. The dataset was split into two parts: 80% for training and 20% as held-out testing dataset. For the prognostic factor identification task, the training set was further split into 85% for training and 15% for validation. Stratified bootstrapping was performed on the validation set to select the best set of hyperparameters. For the prediction task, a stratified 5-fold CV was performed on the 80% training set for hyperparameter selection. The imputation and standardisation were performed within the CV loop to avoid information leakage, as mentioned in Section 2.4. A graphical illustration of the splitting strategy can be found in Figure 2 in Appendix Section A.2.

Hyperparameter Tuning. Hyperparameter selection was conducted in a grid-search manner. A detailed list of the optimal set of hyperparameters for each model and task can be found in Table 4 in Appendix Section A.3. Elastic net regularisation (i.e., L1 (Lasso) and L2 (Ridge) regularisation penalties) was applied to CoxPH. CoxPH with Elastic net [36] was generally found to outperform standard CoxPH during training.

Performance Evaluation. Three metrics were used to evaluate the discrminative performances of the models: concordance index (C-index), Integrated Brier Score (IBS), and time-dependent AUC (tAUC, also known as dynamic AUC).

Implementation. All the models and analyses were implemented using Python 3.10.5. Survival models were implemented with lifelines 0.28.0 and pycox 0.2.3. The CoxPH was trained on a CPU with a memory of 15.2GB. DeepSurv and DeepHit were trained on NVIDIA GPUs with 40GB of RAM.

Table 2: Multivariate CoxPH analysis results for DFS and OS. Relative hazard ratio was calculated for nominal categorical variables with one category as reference (indicated as 'ref' in the table). $P < 0.05$ was considered as significant. Only significant variables are listed here. NA: neoadjuvant; CRT: chemoradiation therapy. Definitions and staging criteria of the features can be found in [17].

	DFS		**OS**	
Variable	**HR (95% CI)**	**P-value**	**HR (95% CI)**	**P-value**
Sex				
Female	ref			
Male	1.200 (1.199-1.200)	**0.007**	1.134 (1.130-1.138)	0.050
Age (Years)	1.017 (1.001-1.033)	0.553	1.115 (1.114-1.115)	**<0.001**
Clinical N stage				
cN0	ref			
cN2	1.155 (1.147-1.163)	0.081	1.239 (1.238-1.240)	**0.010**
Tumor Site				
Junctional	ref			
Lower	1.207 (1.206-1.207)	**0.006**	1.135 (1.132-1.138)	**0.042**
Middle	1.309 (1.306-1.311)	**0.019**	1.266 (1.262-1.269)	**0.027**
Proximal margin positive	1.994 (1.994-1.994)	**<0.001**	1.292 (1.267-1.318)	0.121
Radial margin positive	1.549 (1.549-1.549)	**<0.001**	1.426 (1.426-1.426)	**<0.001**
Pathologic T stage				
T0	ref			
T3	1.583 (1.583-1.583)	**<0.001**	1.490 (1.490-1.490)	**0.001**
T4	1.672 (1.671-1.672)	**0.002**	1.752 (1.751-1.752)	**0.002**
Pathologic N stage				
N0	ref			
N1	1.484 (1.484-1.484)	**<0.001**	1.245 (1.243-1.246)	**0.013**
N2	1.664 (1.664-1.664)	**<0.001**	1.528 (1.528-1.528)	**<0.001**
N3	3.087 (3.087-3.087)	**<0.001**	2.991 (2.991-2.991)	**<0.001**
Pathologic M stage				
M0	ref			
M1	1.707 (1.707-1.707)	**<0.001**	1.919 (1.919-1.919)	**<0.001**
Differentiation				
Gx, cannot be assessed	ref			
Poorly differentiated	1.379 (1.378-1.379)	**0.004**	1.447 (1.446-1.447)	**0.003**
Lymphatic invasion	1.055 (1.000-1.113)	0.573	1.316 (1.316-1.316)	**<0.001**
Venous invasion	1.292 (1.291-1.292)	**<0.001**	1.086 (1.060-1.112)	0.306
Perineural invasion	1.161 (1.158-1.164)	**0.038**	1.193 (1.193, 1.194)	**0.006**
Number of nodes analyzed	0.894 (0.894, 0.894)	**0.002**	0.883 (0.884, 0.884)	**<0.001**
Treatment protocol				
Surgery only	ref			
NA CRT then surgery	1.326 (1.326-1.326)	**0.003**	1.198 (1.195-1.200)	**0.025**
Clavien-Dindo Grade	1.060 (1.059-1.060)	**<0.001**	1.169 (1.169-1.169)	**<0.001**
Length of stay (Days)	1.077 (1.076-1.077)	**0.010**	1.052 (1.050-1.053)	**0.044**
Cancer cases per year	0.919 (0.919-0.919)	**0.008**	0.925 (0.924-0.926)	**0.024**

4 Results

4.1 Prognostic Factor Identification Task

Table 2 summarises the multivariate analysis results of CoxPH with the significant variables (P-value < 0.05) being listed only, along with their hazard ratio (HR), and 95% confidence interval (CI).

4.2 Prediction Task

Table 3 summarises the discriminative prediction performance of three models for DFS and OS respectively. Comparing all three metrics reveals that DeepSurv demonstrates comparable performances to CoxPH, while DeepHit demonstrates slightly inferior performance in terms of IBS. Figure 3 in Appendix Section A.4 provides examples of predicted OS curves obtained from the three models for the same random set of five patients. Notably, while CoxPH and DeepSurv exhibit similar shapes and distributions, DeepHit shows a completely different profile, with minimal variation among the five prediction curves. Despite DeepHit generally ordering patients consistently in terms of survival probabilities compared to the other two models, this profile suggests poorer calibration performance.

Table 3: Summary of model performances. IBS: Integrated Brier Socre; tAUC: time-dependent AUC.

	C-index (95% CI) ↑	IBS (95% CI) ↓	tAUC (95% CI) ↑
DFS			
CoxPH	0.733 (0.710, 0.755)	**0.174 (0.160, 0.187)**	0.720 (0.682, 0.799)
DeepSurv	**0.735 (0.714, 0.758)**	0.176 (0.163, 0.193)	**0.749 (0.727, 0.801)**
DeepHit	0.729 (0.707, 0.752)	0.249 (0.243, 0.263)	0.729(0.693, 0.797)
OS			
CoxPH	0.734 (0.710, 0.758)	**0.164 (0.153, 0.181)**	**0.783 (0.738, 0.818)**
DeepSurv	**0.740 (0.716, 0.764)**	0.169 (0.152, 0.192)	0.781 (0.734, 0.827)
DeepHit	0.739 (0.716, 0.762)	0.214 (0.201, 0.233)	0.776 (0.707, 0.827)

5 Discussion and Conclusion

In this work, we analysed a heterogeneous multicenter dataset to investigate the contribution of covariates to and predictive performance of three models on DFS and OS in patients with esophageal cancer. The significant prognostic factors identified aligned well with clinical literature and experiences. For example, pathologic tumor staging features appear to be strong prognostic factors, and are generally more significant than clinical staging [35]. The more advanced the pathologic stage of the tumor is, the higher the hazard ratio. In terms of

prediction, DeepSurv consistently outperformed CoxPH in both DFS and OS tasks, with C-index of 0.735 and 0.740, respectively, when C-index serving as the primary metric. Overall, the two DNN-based models demonstrated comparable discriminative performance to CoxPH; though DeepHit was found to exhibit poorer calibration performance compared to the other two models. The use of a multicenter international dataset, which includes patients with either adenocarcinoma or squamous cell carcinoma, suggested broader applicability of these findings across diverse cohort in various clinical settings. In general, despite their ability to model more complex interactions, DNN-based models did not greatly outperform the CoxPH. The CoxPH, which is interpretable and computationally efficient, still remains a sufficiently good prediction model with tabular data.

While all three models demonstrated good discriminative performance, it is inferred that these results likely represent the upper bound achievable with tabular data. It is worth noting that some significant features, for example, Clinical N stage, are derived from radiologic assessment scans (Computed Tomography (CT), Positron Emission Tomography (PET)) [17]. Therefore, incorporating imaging-derived features such as radiomics could provide more detailed information and potentially enhance model performance [42]. It should be noted that, among the three models, DeepHit posed particular challenges during training, showing large fluctuations in performance and high sensitivity to hyperparameters. This difficulty can be attributed to its end-to-end neural network architecture, which involves a multitude of hyperparameters. More advanced hyperparameter selection techniques such as Bayesian optimisation could therefore be explored [39] during the training. Graphical convolutional neural work (GNN) [40], which has been an emerging model in survival analysis, could also be explored in the future. In addition, it could be observed that DFS and OS share some common significant prognostic factors. This suggest the possibility of multi-task learning of these two prediction tasks [41]. Furthermore, introducing additional calibration techniques [37] could improve the alignment of predictions with ground truth data. Methods like SHAP [38] could also enhance the interpretability of neural networks by identifying crucial features in predictive models.

Acknowledgments. This work is funded by Cancer Research UK (CRUK). The authors acknowledge the contributions of Sinead King, St. James's Hospital, Dublin, Ireland; Hannah Adams, Gloucestershire Hospitals NHS Foundation Trust, England; and Masaru Hayami, CLINTEC, Karolinska Institutet, Stockholm, Sweden. The authors acknowledge the contributions to and previous works on the ENSURE study by Elliott J.A. et al. [16] and Gujjuri R.R. et al. [12].

The authors would like to thank the Oxford Biomedical Research Computing (BMRC) facility for providing the computing resources. Special thanks to Lav Radosavljevic from University of Oxford for his professional advice on statistical analysis.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

A Appendix

A.1 Imputation Procedure

Figure 1 illustrates the overall process of variable preprocessing and imputation, as described in Section 2.4.

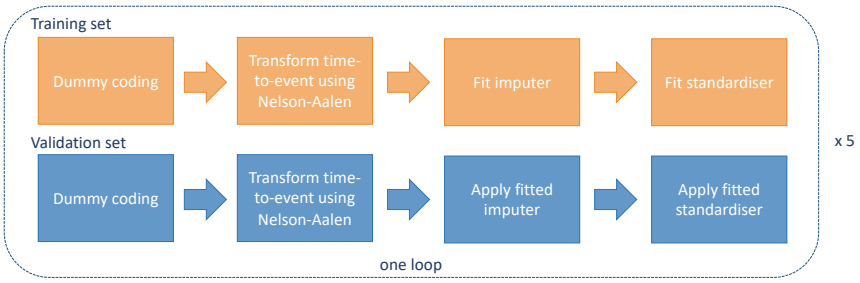


Fig. 1: Flowchart illustrating the preprocessing and imputation process. The process is performed for each loop and repeated across all loops within the CV.

A.2 Splitting Strategy

Figure 2 illustrates the splitting strategy during training for the two tasks.

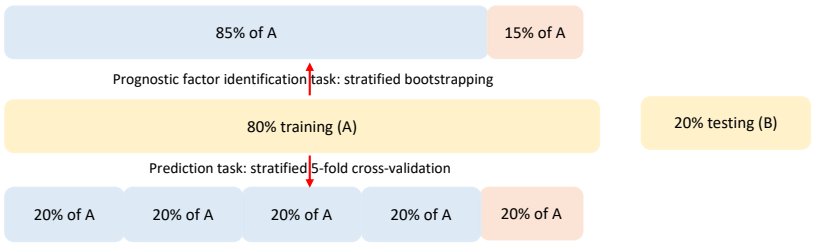


Fig. 2: Graphical illustration of the splitting strategy during training for prognostic factor identification task and prediction task, where blue and orange color represents the training and validation set for each task, respectively.

A.3 Hyperparameter Tuning

Table 4 summarises the optimal set of hyperparameters of the three models. For DeepSurv and DeepHit, various network structures were explored, along with different number of epochs, batch sizes, optimiser schedulers, and learning rates.

Table 4: Summary of the optimal hyperparameter set of the three models for DFS and OS.

(a) Best hyperparameter set of CoxPH.

	L1 penalty	**L2 penalty**
DFS	0.008	0.001
OS	0.006	0.002

(b) Best hyperparameter set of DeepHit. lr: learning rate; w decay: weight decay.

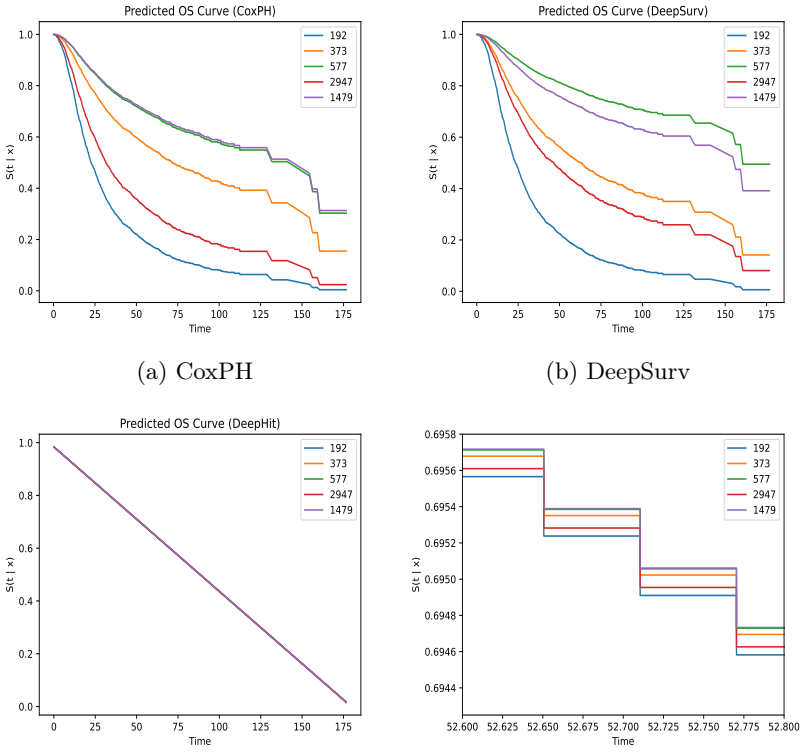
	network	**dropout**	**epochs**	**batch size**
DFS	[64, 64]	0.1	75	64
OS	[64, 128, 64]	0.1	70	256
	optimiser	**initial lr**	**scheduler**	**w decay**
DFS	Adam	0.1	Exp.LR, γ=0.7	0.05
OS	Adam	0.1	Exp.LR, γ=0.7	0.05

(c) Best hyperparameter set of DeepHit. lr: learning rate; w decay: weight decay.

	network	**dropout**	**epochs**	**batch size**	**optimiser**
DFS	[64, 128, 64]	0.1	25	64	Adam
OS	[64, 128, 64]	0.1	100	64	Adam
	initial lr	**scheduler**	**w decay**	**no. of output**	**output interp. no.**
DFS	0.005	Exp.LR, γ=0.7	0.05	60	50
OS	0.005	Exp.LR, γ=0.7	0.05	60	50

A.4 Predicted Survival Curves

Figure 3 shows the predicted survival curves by three models of the same five patients.



(a) CoxPH (b) DeepSurv

(c) DeepHit (d) DeepHit zoomed-in

Fig. 3: Predicted OS curves for the same random set of five patients by three models respectively. The legend shows the patient ID.

References

1. Siegel, R., Giaquinto, A.N., Jemal, A.: Cancer statistics, 2024. CA Cancer J Clin. **74**(1), 12–49 (2024)
2. Mariette, C., et al.: Factors predictive of complete resection of oerable esophageal cancer: review of 746 patients. Gastroenterol Clin Biol. **26**(5), 454–462 (2002)
3. Boerner, T., et al.: Incidence and management of esophageal cancer recurrence to regional lymph nodes after curative esophagectomy. Int. J. Cancer. **152**(10), 2109–2122 (2023)
4. Kunisaki, C., et al.: Surgical Outcomes in Esophageal Cancer Patients with Tumor Recurrence After Curative Esophagectomy. J Gastrointest Surg **12**(5), 802–10 (2008)
5. Barbar, L., et al.: Prognostic immune markers for recurrence and survival in locally advanced esophageal adenocarcinoma. Oncotarget. **10**(44), 4546–4555 (2019)
6. Wang, Q., Lagegren, J., Xie, S.: Prediction of individuals at high absolute risk of esophageal squamous cell carcinoma. Gastrointest Endosc. **89**(4), 726–732 (2019)

7. Chen, W., et al.: Selection of high-risk individuals for esophageal cancer screening: A prediction model of esophageal squamous cell carcinoma based on a multicenter screening cohort in rural China. Int J Cancer. **148**(2), 329–339 (2021)
8. Cox, D.R.: Regression Models and Life-Tables. Journal of the Royal Statistical Society: Series B (Methodological) **34**(2), 187–202 (1972)
9. Wang, W., et al.: A novel molecular and clinical staging model to predict survival for patients with esophageal squamous cell carcinoma. Oncotarget. **7**(39), 63526–63536 (2016)
10. Gabriel, E., et al.: Novel Calculator to Estimate Overall Survival Benefit from Neoadjuvant Chemoradiation in Patients with Esophageal Adenocarcinoma. Journal of the American College of Surgeons **224**(5), 884–894 (2017)
11. Shapiro, J., et al.: Prediction of survival in patients with oesophageal or junctional cancer receiving neoadjuvant chemoradiotherapy and surgery. Br J Surg. **103**(8), 1039–47 (2016)
12. Gujjuri, R.R.: Predicting long-term survival and time-to-recurrence after esophagectomy in patients with esophageal cancer - Development and validation of a multivariate prediction model. Annals of Surgery. **13**, 971–978 (2023)
13. Zhang, K., et al.: Machine learning-based prediction of survival prognosis in esophageal squamous cell carcinoma. Scientific Reports volume. **13**, (2023)
14. Gong, X., et al.: Application of machine learning approaches to predict the 5-year survival status of patients with esophageal cance. JTD. **3**(11), 6240–6251 (2021)
15. 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, pp. 785–794. (2016)
16. Elliott, J.A., et al.: An International Multicenter Study Exploring Whether Surveillance After Esophageal Cancer Surgery Impacts Oncological and Quality of Life Outcomes (ENSURE). Annals of Surgery **277**(5), 1035–1044 (2023)
17. Rice, T.W., Patil, D.T., Blackstone, E.H.: 8th edition AJCC/UICC staging of cancers of the esophagus and esophagogastric junction: application to clinical practice. Annals of Cardiothoracic Surgery **6**(2), 119–130 (2017)
18. SI, G, et al.: Progression-Free Survival: What Does It Mean for Psychological Well-Being or Quality of Life? Agency for Healthcare Research and Quality (US) (2013)
19. Lebwohl, D., et al.: Progression-free survival: gaining on overall survival as a gold standard and accelerating drug development. Cancer J. **15**(5), 386–94 (2009)
20. Dong, Y., Peng, C.J.: Principled missing data methods for researchers. SpringerPlus. **222**(2),(2013)
21. Mack, C., Su, Z., Westreich, D.: Managing Missing Data in Patient Registries: Addendum to Registries for Evaluating Patient Outcomes: A User's Guide, Third Edition [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US.) **17**(18), (2018)
22. García-Laencina , P.J., et al.: Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values. Computers in Biology and Medicine.(59), 125–133 (2015)
23. Jerez, J.M., et al.: Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artificial Intelligence in Medicine **50**(2), 105–115 (2010)
24. Azur, M.J., et al.: Multiple imputation by chained equations: what is it and how does it work? nt. J. Methods Psychiatr. **20**(1), 40–49 (2011)
25. Li, P., Stuart, E.A., Allison, D.B.:Multiple Imputation: A Flexible Tool for Handling Missing Data. JAMA. **314**(18), 1966–7 (2015)

26. Rubin, D.B.: Flexible Imputation of Missing Data. 2nd edn. Chapman and Hall/CRC. Multiple imputation (2018)
27. Austin, P.C., et al.: Missing Data in Clinical Research: A Tutorial on Multiple Imputation. Canadian Journal of Cardiology **37**(9), 1322–1331 (2021)
28. Colosimo, E., et al.: Empirical comparisons between Kaplan-Meier and Nelson-Aalen survival function estimators. J. Statist. Comput. Simul. **72**(4), 299–308 (2002)
29. White, I.R., Royston, P.: Imputing missing covariate values for the Cox model. Statist. Med. **28**(15), 1982–1998 (2009)
30. Spooner, A., et al.: A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. Sci Rep. **20410**(10),(2020)
31. Ake, C.F., et al.: Rounding After Multiple Imputation With Non-binary Categorical Covariates. 112–30 (2005)
32. Akoglu, H.: User's guide to correlation coefficients. Turkish Journal of Emergency Medicine **18**(3), 91–93 (2018)
33. Katzman, J.L., et al.: DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med Res Methodol **18**(24), (2018)
34. Lee, C., et al.: DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 2314–2321. (2018)
35. Smyth, E.C., et al.: Oesophageal cancer. Nat Rev Dis Primers **3**, (2017)
36. Zou, H., Hastie, Trevor: Regularization and variable selection via the elastic net. J. R. Statist. Soc. B **67**(2), 301–320 (2005)
37. Goldstein, M., et al.: X-CAL: Explicit Calibration for Survival Analysis. Adv Neural Inf Process Syst. **67**(2), 18296–18307 (2020)
38. Lundberg, S.M.,Lee, S.: A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 4768–4777. New York, USA (2017)
39. Kaur, P., Singh, A., Chana, I.: BSense: A parallel Bayesian hyperparameter optimized Stacked ensemble model for breast cancer survival prediction. Journal of Computational Science **60**,(2022)
40. Hou, W., et al.: Hybrid Graph Convolutional Network With Online Masked Autoencoder for Robust Multimodal Cancer Survival Prediction. IEEE Transactions on Medical Imaging **42**(8), 2462–2473 (2023)
41. Yun, S., Du, B., Mao, Y.: Robust Deep Multi-task Learning Framework for Cancer Survival Analysis. In: 2021 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. Shenzhen, China (2021)
42. Furukawa, M., et al.: Prediction of recurrence free survival of head and neck cancer using PET/CT radiomics and clinical information. (2024) https://arxiv.org/abs/2402.18417