Deep Learning for CT-Based Survival Analysis of Idiopathic Pulmonary Fibrosis Patients

Alexander C. Whitehead *Student Member, IEEE*, Ahmed H. Shahin, An Zhao, Daniel C. Alexander, Joseph Jacob, and David Barber

***Abstract*—Idiopathic pulmonary fibrosis is an interstitial lung disease that causes scarring of the lungs, leading to a decline in lung function and eventually death. Because this disease has a heterogeneous disease progression, predictive models could guide clinicians in making decisions about disease management. Some survival analysis methods, such as Cox, seek to rank participants based on their predicted survivability. However, Cox cannot directly output a survival time. DeepHit is a neural network based survival analysis method which predicts the most likely histogram bin of survival time. A disadvantage of DeepHit is that, when training, an error of one year is equivalent to an error of one hundred years. A common problem encountered is that training data is often censored, where the exact time of death is unknown except that it is past a censoring time. Here, a comparison of neural network approaches utilising five different losses is presented. Compared are; ranking based approaches (such as Cox or Cox with a memory bank of previous predictions) and death distribution based approaches (such as DeepHit and likelihood with a uniform or Gaussian distribution to sample censoring times). The input to each model is a single computed tomography volume (plus optionally clinical features) and the output is a survival time. Improvements over previous work includes; a larger model with a learned downsampling, a parameterised activation (which starts linear and becomes non-linear), a softplus output, orthogonal initialisation, an optimiser integrating weight decay, gradient accumulation, and an annealed learning rate. Evaluations used include; mean and relative absolute error, the concordance index, the Brier score, and a visual analysis of Grad-CAM results. Overall, the likelihood models performed the best, with DeepHit a close second and both Cox models a distant last. The Event Conditional Likelihood model performed marginally better than the alternative.**

This work was funded by Open Source Imaging Consortium (OSIC).

Alexander C. Whitehead, Ahmed H. Shahin, An Zhao, Daniel C. Alexander, and David Barber are with the Department of Computer Science, University College London, London, UK.

Ahmed H. Shahin, An Zhao, Daniel C. Alexander, and Joseph Jacob are with the Centre for Medical Image Computing, University College London, London, UK.

Joseph Jacob is with Lungs for Living Research Centre, University College London, London, UK.

David Barber is with the Centre for Artificial Intelligence, University College London, London, UK.

(contact: alexander.whitehead.18@ucl.ac.uk).

I. Introduction

UNDER the umbrella of Interstitial Lung Diseases (ILDs) , Idiopathic Pulmonary Fibrosis (IPF) is characterised by a buildup of scar tissue in and a stiffening of the lungs of the patient; leading to a reduction in the volume of the lung, resulting in a shortness of breath and eventually death. As with many other ILDs , the progression of the disease is heterogeneous, and prognosis is challenging [1].

Previously, methods to monitor IPF included testing lung function, with spirometric measurement of lung volume [2], or taking Computed Tomography (CT) acquisitions over time [3]. Both approaches are limited by several factors, including, physical limitations of the patient, technical accuracy (spirometer baseline drift bias), and longitudinal data availability. In contrast, measuring how long an IPF patient could survive on the basis of a baseline CT scan can be a useful clinical outcome to measure in order to prioritise resources and plan intervention. For this reason, here the focus is on accurately predicting IPF patient death time based on a baseline CT scan and associated clinical features.

In Cox Proportional Hazards Survival Analysis [4] the death time of one participant is compared to another, and the model ranks patients according to their expected death times. However, a limitation is that the model does not directly output survival times. Recently models have been introduced that attempt to predict more directly the death time of a participant. For instance, DeepHit uses a Convolutional Neural Network (CNN) to perform feature extraction on input CT and outputs the probability of a survival time falling within predetermined bins of a death time histogram [5]. This treats survival analysis as a multi-class classification problem with a class for each time-bin that a patient could die in. A disadvantage is that the bins are not ordinally related and the model is penalised as much for making an error of one month as it is an error of ten years.

Censoring is a significant issue in survival analysis in which the precise death time of a patient is unknown. In the OSIC IPF data set [6], approximately $66\,\%$ of the records are right-censored, meaning that the time of death is above a known value but it is unknown by how much. A simple approach would be to remove censored data, however, this would discard a very significant fraction of training data. Missing data in clinical records is a related issue. Again, with approximately $66\,\%$ of the OSIC IPF data [6] set has some missing clinical information.

Here, a number of survival analysis models that predict death time using a Neural Network (NN) are presented. The inputs of which being a baseline CT scan and associated patient clinical information (such as height, age, etc). The models are able to address censoring and missing clinical information following [7], [8] respectively. Different training losses are used, including ones based on classical Cox based ranking, likelihood, and DeepHit. In the case of the Cox based loss, one with and one without a memory bank of previous predictions is used (to allow the loss to be approximated at all previously seen data points [8]). In the case of the likelihood based loss, one where censoring time is sampled in the classical way and one where censoring time is sampled from a uniform distribution is used [7]. Extensions over previous work include; a change to the NN architecture (larger, with a learnt downsampling, parameterised activation and softplus output, and orthogonal initialisation), a new optimiser, gradient accumulation (as such an increased batch size), and an annealed learning rate.

II. Methods

A. Data Acquisition and Preparation

A total of 550 CT acquisitions were taken from the OSIC data set [6]. Each volume was segmented to remove data outside of the lung and was normalised independently. Where appropriate, clinical features, such as age and sex were also used. If missing clinical features were present, their value was imputed following [8]. Data were split into train and test groups using five fold Cross Validation (CV).

B. Models

Each NN consisted of seven CNN blocks, within which were two convolutions with stride one and one with stride two. Each convolution had a kernel size of three and used an orthogonal activation [9]. Between each layer there was a Parametric Rectified Linear Unit (PReLU) activation [10], initialised with α set to one (meaning the network begins linear and becomes more non-linear as training progresses). At each downsampling step, the number of channels doubled. Global average pooling and flattening layers were used before fully connected layers reduced the number of units until it equals the output size (by halving the number of units at each layer). When clinical features were used, they were concatenated to the output of the flattening layer. A softplus activation was used at the output for numerical stability. The model architecture was selected using the Event Conditional Likelihood loss.

AdamW was used as the optimiser, with weight decay, to improve the convergence rate as well as to penalise against large weights and overfitting [11]. The learning rate started close to zero and increased linearly to the target learning rate over the first one tenth of iterations, before reducing back to close to zero over the next nine tenths. For each loss calculation a batch size of four was used, this is because Cox loss requires a batch size greater than one (for the sake of comparison the same number was used for all losses). This was approximately increased to 32 using gradient accumulation, meaning eight gradients were averaged together at each iteration. Because of the Memory Bank (MB) the effective batch size of Cox MB was greater.

For comparison, five loss functions were trialled:

- **Event Conditional Likelihood** - Maximise likelihood using a Gaussian to model the time of death, where the censoring time was sampled from a uniform distribution from time zero up to the death time [7].
- **Classical Likelihood** - Maximise likelihood using a Gaussian to model the time of death, where the censoring time was sampled from a Gaussian distribution parameterised by the censor time. This is one of the classical ways to handle censoring [5].
- **Cox** - Cox Proportional Hazards [4].
- **Cox MB** - Cox Proportional Hazards with MB [8].
- **DeepHit** - Log-likelihood, with a maximum output value of 105 years and 840 bins [5].

For both likelihood losses a fixed Standard Deviation (STD) equal to one year was used. For both Cox losses the output was converted to survival times using the Breslow estimator [12].

C. Evaluation

For evaluation of the results of the five loss functions the following methods were used; the Mean Absolute Error (MAE) and Relative Absolute Error (RAE) for the uncensored data between the predicted and true value was taken, the concordance index, the Brier score and a visual analysis of Grad-CAM images [13]–[15]. For display of Grad-CAM a slice was selected which displayed fibrosis

TABLE I
A COMPARISON OF MAE, RAE, THE CONCORDANCE INDEX, AND THE BRIER SCORE. THE AVERAGE SURVIVAL TIME WAS APPROXIMATELY 32 MONTHS. HERE C REFERS TO CLASSICAL LIKELIHOOD, EC LIKELIHOOD REFERS TO EVENT CONDITIONAL LIKELIHOOD, AND CF REFERS TO WHEN THE CLINICAL FEATURES WERE INCLUDED IN THE MODEL.

	MAE	RAE	C-Index	Brier
EC Likelihood	22.7 ± 1.51	1.72 ± 0.89	0.77 ± 0.05	0.22 ± 0.07
EC Likelihood CF	21.5 ± 1.32	1.98 ± 0.82	0.80 ± 0.03	0.18 ± 0.05
C Likelihood	28.9 ± 1.96	2.23 ± 0.01	0.76 ± 0.05	0.25 ± 0.01
C Likelihood CF	25.3 ± 1.74	2.04 ± 0.01	0.75 ± 0.04	0.20 ± 0.01
Cox	187 ± 309	17.0 ± 30.7	0.73 ± 0.04	0.61 ± 0.28
Cox CF	233 ± 287	26.4 ± 21.9	0.72 ± 0.03	0.57 ± 0.16
Cox MB	166 ± 267	17.7 ± 28.2	0.74 ± 0.03	0.53 ± 0.31
Cox MB CF	179 ± 294	16.3 ± 22.8	0.73 ± 0.05	0.56 ± 0.24
DeepHit	38.4 ± 14.8	3.99 ± 0.34	0.72 ± 0.03	0.40 ± 0.01
DeepHit CF	31.3 ± 9.19	3.50 ± 0.42	0.71 ± 0.04	0.42 ± 0.01

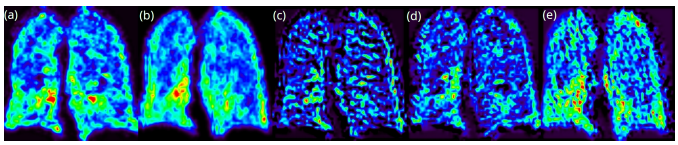

Fig. 1. From left to right a slice through a fibrotic region of a Grad-CAM image, taken from a middle convolution, of a 65 year old patient with a survival time of 30 months For; (a) Death Conditional Likelihood, (b) Classical Likelihood, (c) Cox, (d) Cox MB, and (e) DeepHit. All colour maps are consistent for all images.

in the base of both lungs. Only the model without clinical features was used for Grad-CAM extraction.

III. RESULTS

From TABLE I it can be seen that the the MAE and RAE are often lower for both likelihood based models than for all other models. The MAE and RAE for the DeepHit model is lower than that of the Cox based model, while for the Cox based models not only are their errors high but they also have a substantially higher variance. The MAE and RAE of the Event Conditional Likelihood model is often lower than that of the Classical Likelihood model. The Brier score results also back up this assertion, however it is difficult to draw conclusions from the concordance index results. The results also seem to indicate that there is some benefit to including the clinical features, although from these results alone it is not possible to say if this is due to the added information from the clinical features or the increase in model size. From Fig. 1 it can be seen that both likelihood based model and DeepHit produced updates which identified the fibrosis in both lungs, the Cox based model only seemed to detect fibrosis in the left lung. Both likelihood models seemed to extract updates which are less noisy than the DeepHit update.

IV. DISCUSSION AND CONCLUSION

From a comparison of errors and a visual analysis it appears that the likelihood based models provide the best results most often.

The model used for the DeepHit model had more parameters than the model used for all other methods (due to the output being larger), thus it may not be an entirely fair comparison. However, while using a larger model the method does not provide results significantly better than the likelihood models.

When clinical features were used it seems to improve results, although not significantly. For the increase in complexity it may not be worth including.

What is not factored into the results is computation time. The Cox loss without MB is the fastest to compute, the likelihood losses are not much longer. The DeepHit loss takes slightly longer than both previous methods while the Cox MB loss takes magnitudes longer (approximately six hours vs four days).

REFERENCES

[1] T. E. King *et al.*, "Idiopathic pulmonary fibrosis," in *The Lancet*, vol. 378, Dec. 2011, pp. 1949–1961.

[2] L. C. Watters *et al.*, "A clinical, radiographic, and physiologic scoring system for the longitudinal assessment of patients with idiopathic pulmonary fibrosis," *American Review of Respiratory Disease*, vol. 133, no. 1, pp. 97–103, May 1986.

[3] D. A. Lynch *et al.*, "Diagnostic criteria for idiopathic pulmonary fibrosis: a Fleischner Society White Paper," *The Lancet*, vol. 6, no. 2, pp. 138–153, Feb. 2018.

[4] D. R. Cox, "Regression Models and Life-Tables," *Series B*, vol. 34, no. 2, pp. 187–202, Jan. 1972.

[5] C. Lee *et al.*, "DeepHit: A deep learning approach to survival analysis with competing risks," in *AAAI Conference on AI*, vol. 32, Apr. 2018, pp. 2314–2321.

[6] OSIC, *OSIC Data Repository*.

[7] A. H. Shahin *et al.*, "Deep Learning for Accurate Survival Analysis," *Medical Image Analysis (in review)*, 2023.

[8] A. H. Shahin *et al.*, "Survival Analysis for Idiopathic Pulmonary Fibrosis using CT Images and Incomplete Clinical Data," *arXiv*, Mar. 2022.

[9] W. Hu *et al.*, "Provable Benefit of Orthogonal Initialization in Optimizing Deep Linear Networks," *arXiv*, Jan. 2020.

[10] K. He *et al.*, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *IEEE Conference on CV*, vol. 2015 Inter, 2015, pp. 1026–1034.

[11] I. Loshchilov *et al.*, "Decoupled weight decay regularization," in *Conference on Learning Representations*, Nov. 2019.

[12] N. Breslow, "Covariance Analysis of Censored Survival Data," *Biometrics*, vol. 30, no. 1, p. 89, Mar. 1974.

[13] V. C. Raykar *et al.*, "On ranking in survival analysis: Bounds on the concordance index," in *Advances in Neural Information Processing Systems*, vol. 20, 2008.

[14] T. A. Gerds *et al.*, "Consistent estimation of the expected brier score in general survival models with right-censored event times," *Biometrical Journal*, vol. 48, no. 6, pp. 1029–1040, Dec. 2006.

[15] R. R. Selvaraju *et al.*, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, 2020.