On a Survival Gradient Boosting, Neural Network and Cox PH based Approach to Predicting Dementia Diagnosis Risk on ADNI

Henry Musto
Department of Computing
Goldsmiths College, University of London
United Kingdom
hthom018@gold.ac.uk

Daniel Stamate
Department of Computing
Goldsmiths College, University of London,
and School of Health Sciences
The University of Manchester
United Kingdom

Doina Logofatu
Faculty of Computer Science and Engineering
Frankfurt University of Applied Sciences
Germany

Lahcen Ouarbya
Department of Computing
Goldsmiths College, University of London
United Kingdom

***Abstract*—In recent years, attention within the clinical prediction community has turned to the use of survival machine learning as a tool for predicting the risk of developing a disease as a function of time. The current work seeks to contribute to existing literature which demonstrates the utility of these methods when applied to a dementia prediction context. We use the Alzheimer's Disease Neuroimaging Initiative ADNI dataset and model deterioration within two distinct groups, those deemed cognitively normal and those with a formal diagnosis of Mild Cognitive Impairment. In agreement with existing literature we find that survival machine learning outperforms standard survival analysis methods such as Cox PH model, and has very good predictive ability. We propose an innovative approach to predicting dementia diagnosis risk on ADNI, which explores the use of survival neural network and survival extreme gradient boosting techniques that have hitherto seldom been applied to this context. The stability of our models was investigated within a Monte Carlo simulation framework.**

***Index Terms*—Dementia Risk Prediction, Survival Analysis, Survival Machine Learning, Gradient Boosting, Neural Networks, Clinical Prediction Modelling, Monte Carlo**

I. Introduction

By 2050 it is estimated that around 132 million people will be living with dementia worldwide. Currently, between 60-70% of all dementias are of Alzheimer's Disease type (AD) [1]. Although clinical prediction modelling using Machine Learning (ML) methodologies has the potential to alleviate suffering by predicting those who are likely to develop the disease, thus far, none of the published ML models that seek to address dementia prediction have been successfully adopted into clinical practice [2]. One of the challenges that prevent this successful integration concerns the information that can be gleaned from classic ML methods. These techniques often deliver a binary or multinomial prediction indicating the likelihood of the development of disease. However, in prognostic modelling, it is usually more valuable to model the risk of developing the disease as a function of time. This limitation of classical ML techniques restricts clinicians' ability to accurately track and communicate the risk of disease occurrence over time with the patient [3]. In the research literature, many AD studies that use ML methods do indeed employ this classification approach, whereby the outcome to be predicted is either binomial or multinomial within a specific timeframe [4] [5]. Conversely, the datasets used in clinical research are often derived from longitudinal studies, whereby clinical marker data is collected from participants over months and years [6]. This data is inherently dependent on time, something that standard classification approaches cannot account for, as they cannot consider the predictive power of time in conjunction with other predictors. Furthermore, classification models cannot handle drop-outs, where participants are lost to follow up, something which is common in longitudinal studies. As a result, an emerging field of exploration seeks to build on classical time-dependent models, such as survival analysis, to develop machine learning models which can predict the time-dependent risk of developing AD and thus move beyond simple classification. Survival analysis is a statistical method that predicts a hazard score indicating the risk of an event's occurrence, as a function of time. A common aspect of survival analysis is the presence and handling of censored data, indicating that the event of interest has not occurred while the subject was part of the study. Censored data requires the use of specialised techniques, of which, the Cox proportional hazards (Cox PH) model [7] has historically been the most widely used. However, the Cox model typically works best with small datasets and does not scale well to complex multidimensional data

[8]. ML techniques inherently handle multidimensional data and have therefore been adapted to handle censored data, allowing ML to offer a more flexible alternative for analysing high-dimensional, censored, heterogeneous data [8]. A further strength of survival-based techniques is that the models can provide not only a binary or multinomial outcome but also the risk of such outcomes occurring at different timepoints, allowing more information for clinicians, researchers, and participants.

This work has several aims. First, it aims to build upon existing literature demonstrating the utility of survival-based ML techniques in predicting the risk of deterioration at different time points in AD using the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset. Secondly, it aims to explore the utility of employing survival neural network and survival extreme gradient boosting techniques and compare them with Cox PH model in dementia prediction, and to propose an uniform approach to predicting dementia diagnosis risk based on these ML and statistical methods. In particular, these survival ML modelling techniques have hitherto seldom been explored in this context and may provide better predictive performance than existing statistical survival analysis models tested in this setting. The rest of the paper will be ordered as follows. First, it will review existing literature on survival-based ML as applied to clinical questions in general and AD prediction in particular. Next, the problem of interest will be defined. Then the proposed methodology will be introduced. Before the results are presented, the study design of the dataset will be described, including predictors and diagnostic criteria. A discussion of the implications of these results will then follow.

II. Literature Review

Spooner et al. [8] systematically compared the performance and stability of ML algorithms and feature selection methods suitable for high-dimensional, heterogeneous, censored clinical data, in the context of cognitive ageing and AD, by predicting the risk of AD over time [8]. The authors assessed ten survival-based machine-learning techniques alongside the standard Cox proportional hazard model. The Sydney Memory and Aging Study (MAS) dataset and Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset were utilised. All algorithms evaluated performed well on both data sets and outperformed the standard Cox proportional hazards model. Another paper that explores the clinical utility of survival modelling within the domain of AD research comes from [10], which looked at the interaction between socioeconomic features and polygenic hazard scores on the timing of Alzheimer's diagnosis using Cox proportional hazard survival analysis. Only the standard Cox PH technique was used. The authors could demonstrate the clinical utility of using socioeconomic markers and the presence of the APOE4 gene expression to predict the time to AD diagnosis. Although a small study focusing on only one model, this work demonstrated the utility of survival-based models in AD prediction. However, more work was needed to build upon these results using ML methods. This was achieved in [11] using ML survival-based methods to predict the risk of developing AD in the English Longitudinal Study of Aging (ELSA) dataset. This work again found that Survival ML outperformed Cox methods. On the other hand, [12] found the standard Cox regression and two ML models (Survival Random Forest and Extreme Gradient Boosting) had comparable predictive accuracy across three different performance metrics, when applied to the Prospective Registry For Persons with Memory Symptoms (PROMPT) dataset [13]. The authors concluded that survival ML did not perform better than standard survival methods. In comparison, [14] found that multi-modal survival-based deep learning methods produced good results when applied to the ADNI dataset, comparable to [8]. However, more recent work by [15] found that a neural network-based survival model did not outperform Survival Random Forest, but did outperform a standard Cox Proportional hazard model. Recent examples have shown promise in attempting to outperform the classic Cox proportional hazard model, using survival ML and survival neural networks/ deep learning on clinical datasets. This supports the continued exploration of survival ML as a predictive tool for clinical risk problems [11].

III. Study Overview

This study uses survival-based ML methods to predict the risk of deterioration, defined as receiving a worse diagnosis at their final visit to the data collection centre before leaving the study, compared to baseline diagnosis. Furthermore, the study aims to build models to predict the risk of receiving a worse diagnosis within the data collection period using survival-based ML. These models will then be tested for stability, and two estimations of the general test error will be calculated based on C-Index and Calibration scores [15] [16].

IV. Methodology

A. Alzheimer's Disease Neuroimaging Initiative

The data used in this paper was derived from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database [6]. This longitudinal case-control study was initiated in 2004 by the National Institute of Aging (NIA), The National Institute of Biomedical Imaging and Bioengineering (NIBIB), The Food and Drug Administration (FDA), as well as elements of the private and non-profit sectors. The initial protocol, ADNI1, was conducted over six years, recruiting 400 subjects diagnosed with Mild Cognitive Impairment (MCI), 200 subjects with Alzheimer's (AD), and 200 healthy controls (CN). The initial goal of the ADNI study was to test whether repeated collections of neuroimaging, biomarker, genetic, and clinical and neuropsychological data could be combined to contribute in an impactful way to research dementia [6]. Data for the present paper was downloaded on the 1st of October 2023 through the ADNIMERGE package in R. This package combines predictors from the different ADNI protocols. The final combined dataset contains 115 variables and 15,157 observations, which included multiple observations per participant.

These observations represent data collection events where participants made up to 23 visits to study sites. The data used for this work is a subset of the full dataset, containing only information from the original ADNI2 study. After some initial cleaning, the resulting data contained 607 observations and 52 variables consisting of 50 input attributes, 1 time attribute (defined as the time in months until the participant visited the data collection centre for the last time), and 1 outcome attribute. The outcome attribute consisted of three diagnostic classes received at their final visit to the data collection centre: those who received a diagnosis of Cognitively Normal (CN), those who received a diagnosis of Mild Cognitive Impairment (MCI), and those who received a diagnosis of Alzheimer's Disease (AD) [4]. Building on [15] it was decided to exclude CSF-derived biomarkers from the modelling process.

B. Input Variables

- Baselines Demographics: age, gender, ethnicity, race, marital status, and education level were included in the original dataset.
- Neuropsychological test results, including those from the Functional Activities Questionnaire (FAQ), the Mini-Mental State Exam (MMSE), and Rey's Auditory Verbal Learning Test (RAVLT), were included in the data. This numeric data is well-validated as a tool for identifying cognitive impairment in general and AD-related cognitive impairment in particular. Full details of the tests included can be found in [17].
- Positron Emission Tomography (PET) measurements (FDG, PIB, AV45) are indirect measures of brain function using the Positron Emission Tomography neuroimaging modality.
- Magnetic Resonance Imaging (MRI) measurements (Hippocampus, intracranial volume (ICV), MidTemp, Fusiform, Ventricles, Entorhinal and WholeBrain) are structural measurements of a participant's brain derived from the Magnetic Resonance Imaging neuroimaging modality.
- APOE4 is an integer measurement representing the appearance of the epsilon4 allele of the APOE gene. This allele has been implicated as a risk factor for AD [18].
- Last Visit is defined for this work as the number of months from baseline data collection to the subject's last visit at a data collection centre. This variable was added to explicitly define a time predictor for the survival modelling approach presented in this work.

C. Data Preprocessing

Boolean variables were created, indicating the location of missing data for each predictor. Variables with missingness at 90% or greater of the total rows for that predictor were removed. All nominal predictors were dummy-coded. The data was split into two groups to predict deterioration using survival-based ML and Cox PH model. The first group contained only those diagnosed as cognitively normal (CN) on their first visit to the data collection centre. The second group contained only those diagnosed with Mild Cognitive Impairment (MCI) on their first visit to the data collection centre. Deterioration was defined as receiving a worse diagnosis on their final visit to the data collection centre. Full detais of outcome definitions can be seen in Table I and Table II.

The resultant two datasets had 285 and 322 observations respectively and 92 variables (See Table III).

TABLE I
THOSE WHO RECEIVED A COGNITIVELY NORMAL (CN) DIAGNOSIS AT BASELINE WERE THE ONLY GROUP INCLUDED. THE MODELS PREDICTED THE DIAGNOSES THESE PARTICIPANTS RECEIVED AT THE FINAL VISIT, DEFINED HERE.

Outcome	Definition
CN	Those diagnosed with CN at baseline who received the same diagnosis at final visit.
MCI/AD	Those having received a diagnosis of CN at baseline either received a diagnosis of MCI or AD at final visit.

TABLE II
THOSE WHO RECEIVED A MILD COGNITIVE IMPAIRMENT (MCI) DIAGNOSIS AT BASELINE WERE THE ONLY GROUP INCLUDED. THE MODELS PREDICTED THE DIAGNOSES THESE PARTICIPANTS RECEIVED AT THE FINAL VISIT, DEFINED HERE.

Outcome	Definition
CN/MCI	Those diagnosed with MCI at baseline and who received the same diagnosis at their last visit or a diagnosis of CN.
AD	Those diagnosed with MCI at baseline and who received a diagnosis of AD at their last visit.

D. Models

Model development, evaluation, and validation were carried out according to methodological guidelines outlined by [19]; results were reported according to the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) guidelines [20]. This paper explored three algorithms:

Cox Proportional Hazard Model (Cox PH) - The Cox model is expressed by the hazard function, which is the risk of an event occurring at time t as follows:

$$h(t) \ = \ h_0(t) \ * \ exp(\beta_1 X_1 + \beta_2 X_2 + \beta_p X_p) \qquad (1)$$

where t represents the survival time, $h(t)$ is the hazard function acting upon survival time t, $X_1, X_2, ...X_p$ are the values of the p covariate, $\beta_1, \beta_2, ...\beta_p$ are the coefficients that measure the effect of the covariates on the survival time, and $h_0(t)$ is the baseline hazard function. The coefficients are estimated by maximising the partial likelihood and so the model does not require tuning.

Survival XGBoost (SXGB) - Extreme Gradient Boosting is a tree-based ensemble method that grows trees sequentially, by adhering to a gradient descent procedure informed by a loss function, often the negative log liklihood, defined as:

$$\min_{\theta} \sum_{y} -\log(p(y;\theta)) \tag{2}$$

where the model seeks to minimise the negative log of difference between the true outcome y, observed in the training data, and the outcome predicted by the model θ.

This function informs a step function, which calculates the most appropriate adjustments to the model parameters in order to converge on a solution. In the case of the SXGB the negative log-liklihood is used to calculate steps towards a solution that finds the risk score derived from a Cox Proportional Hazard technique. Thus, the SXGB model seeks to find a risk score that will most closely reflect the true risk for that participant [28].

Hyperparameter tuning was performed for model optimisation with the following values. Number of trees was between 500-1500, max depth was between 1-10, min child weight was between 0.0001-0.001, eta was between 0.1-1 and alpha was between 0.01-1. The best performance for the CN group was found when alpha was 0.01, eta was 0.05, min child weight was 0.0001, max depth was 3, and number of rounds was 1000. For the MCI group the best performance was found when alpha was 0.1, min child weight was 0.0001, max depth was 5, and number of rounds was 1500.

Survival DeepHit Neural Networks (SNN) - DeepHit is a multi-task neural network comprising a shared sub-network and K cause-specific sub-networks. The architecture differs from a conventional multi-task neural network in three ways. First, it utilises a single softmax layer as the output layer of DeepHit to ensure that the network learns the joint distribution of K possible outcomes, not the marginal distributions of each outcome. Second, it maintains a residual connection from the input covariates into the input of each cause-specific sub-network. Third, it uses a combination of the negative log likelihood loss function and a ranking loss function, similar to the concordance index, in order to define a differentiable loss function. This function is defined as:

$$Loss = L1 + L2 \tag{3}$$

L1 is defined as:

$$L1 = -\sum_{i=1}^{N}[1(k^{(i)} \neq \theta) \cdot log(y^{(i)}_{k^{(i)},s^{(i)}})]$$
$$+1(k^{(i)} = \theta) \cdot log(1 - \sum_{k=1}^{k} \hat{F}_k\,(s^{(i)}|x^{(i)})) \tag{4}$$

where $1(\cdot)$ is an indicator function. The first term captures the information provided by uncensored patients; the second term captures the censoring bias by exploiting the knowledge that they are alive (in usual survival analysis terms) at the censoring time, so that the first hitting event will occur among one of the K causes after the given censoring time [21].

L2 incorporates estimated Cumulative Incidence Functions (CIF) calculated at different times in order to finetune the network to each cause-specific estimated CIF. In order to do so the loss function employs a ranking function similar to the concordance index. This index states that a patient who dies at time s should have a higher risk at time s than a patient who survived longer than s. For L2 the ranking loss function is defined as:

$$L2 = \sum_{k=1}^{K} \alpha_k \cdot \sum_{i \neq j} A_{k,i,j \cdot n}(\hat{F}_k(s^{(i)}|x^{(i)}), \hat{F}_k(s^{(i)}|x^{(j)})) \tag{5}$$

where the coefficients α_k are chosen to trade off ranking losses of the k-th competing event, and $n(x, y)$ is a convex loss function. The full technical description of this model can be found in [21].

For the SNN modelling process the hyperparameter search space was as follows. Number of nodes between 16-96, activation function either relu or leakyrelu, number of epochs between 100-15, batch size between 32 and 48, leanrning rate between 0.0001, 0.01, lambda between 0-1, and alpha between 0-0.5. For the CN group, the best performance was found when the number of nodes was 16, activation was relu, epochs was 500, batch size was 32, learning rate was 0.001, lambda was 0.3, and alpha was 0.5. For the MCI group, the best performance was found when the number of nodes was 32, activation was relu, number of epochs was 500, batch size was 48, learning rate was 0.001, lambda was 0.1, and alpha was 0.

TABLE III
THE FINAL DIMENSIONS OF THE TWO DATASETS AFTER PREPROCESSING.

Dataset	Variables	Observations
CN at baseline	92	285
MCI at baseline	92	322

E. Nested Cross-Validation and Monte-Carlo Simulation

A Nested Cross-Validation procedure was implemented to tune and evaluate the models so precise estimates of the model's performance of unseen cases (internal validation) could be gathered [4]. Nested Cross-Validation consisted of an outer 5-fold CV (model assessment) and an inner 5-fold CV (model tuning). We conducted a Monte Carlo procedure of 100 repetitions of the nested CV using different random splits per model to assess the models' stability. Performance statistics were recorded for each model produced by each iteration. Each performance statistic's mean and standard deviation across all iterations were recorded when the Monte Carlo was complete. To ensure the representativeness of training and test samples in both procedures, the data splitting was stratified based on the AD cases variable.

F. Performance Metrics

To assess model performance, two statistics were recorded. Discrimination was assessed using the Concordance index or C-index [19]. This metric, also called Harrel's C-index, provides a global assessment of the model and can be considered a more general form of the AUCROC measure typically used in binary classification tasks. The C-index computes the percentage of comparable pairs within the dataset whose risk score was correctly identified by the model. Comparable pairs are defined as a selection of two observations, which can be compared in terms of survival time predicted by the model. If both are censored, then they are not included in the computation for this metric. A pair is considered concordant if the observation who experiences the earlier event is identified as having greater risk and discordant otherwise. Thus the total concordance score for a model is the ratio of concordant pairs within the dataset divided by the total number of comparable pairs [16]. More formally, the concordance index is defined as:

$$C = \frac{\sum concordant\ pairs}{\sum comparable\ pairs} \tag{6}$$

Secondly, calibration was assessed using Van Houwelingen's Alpha Survival Measure of non-proportional hazards models [15]. This metric is defined as:

$$\alpha\ =\ \sum \delta / \sum H_i(t_i) \tag{7}$$

where δ is the true censoring indicator observed from the test data, H_i is the cumulative hazard predicted by the model and t_i is the observed survival time. The model is well calibrated if the estimated α is equal or close to 1. Calibration is a formal comparison between the probability distribution and resultant survival instances observed in the test data and the probability distribution and resultant survival predictions generated by the model. A full exploration of this metric can be found in [22].

G. Software and Hardware

The data analysis was conducted using the R language [23]. Initial data cleaning was performed using base R functions and the Tidyverse R package [24]. The creation of dummy variables was performed using the Caret R package [25]. The nested crossvalidation procedure, including training, tuning and evaluation, was performed on the Cox PH, SXGB, and SNN models using the mlr3 R package [26]. The hardware consisted of 1 server running Ubuntu with a 16-core Ryzen processor, 128 GB of RAM and a 4090 RTX 24GB GPU.

V. Analysis Results

The nested cross-validation C-index and Calibration performance for each model type is detailed below.

The best performing model for both groups was the Survival XGBoost, followed by the Survival DeepHit model. Both Survival ML models outperformed the Cox Proportional Hazards model in terms of Calibration and C-Index performance metrics. All models performed better on the MCI data than the CN data.

TABLE IV
C-Index and Calibration scores for the models applied to the CN and MCI groups.

Model	C-Index - CN / MCI	Calibration - CN / MCI
Cox PH	0.59/0.78	0.01/0.25
SXGB	0.84/0.86	0.9/0.8
SNN	0.70/0.77	0.60/0.91

The Results of the Monte-Carlo Simulation are detailed below.

TABLE V
Monte Carlo simulation of 100 iterations for the models applied to the CN group.

Model	Mean (SD) C-Index - CN	Mean (SD) Calibration - CN
Cox PH	0.59(0.06)	0.03(0.02)
SXGB	0.82(0.03)	0.9 (0.1)
SNN	0.70(0.06)	0.60(0.03)

TABLE VI
Monte Carlo simulation of 100 iterations for the models applied to the MCI group.

Model	Mean (SD) C-Index - MCI	Mean (SD) Calibration - MCI
Cox PH	0.78(0.02)	0.33(0.08)
SXGB	0.86(0.01)	0.81 (0.07)
SNN	0.77(0.02)	0.91(0.1)

When considering the Monte-Carlo simulation, for both groups the survival XGBoost model proved the best performing in terms of the mean C-index and Calibration scores over 100 iterations of the nested cross-validation procedure.

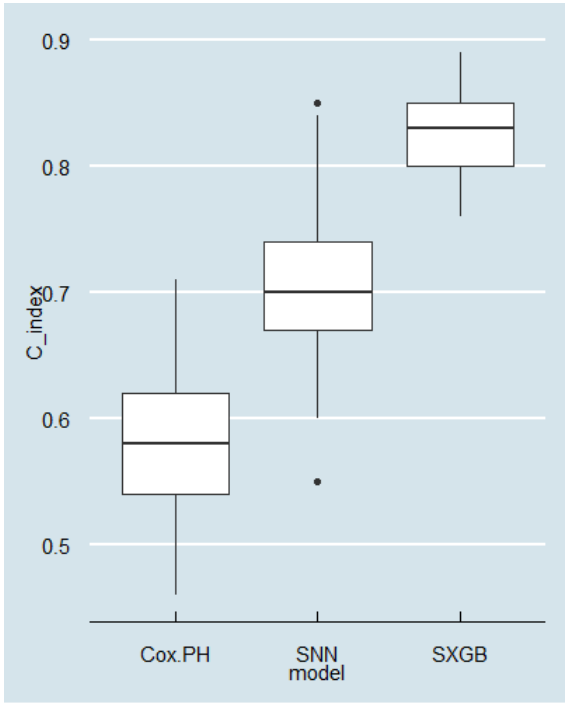

Fig. 1. Boxplots of C-Index performances for models applied to the CN group, obtained in the Monte Carlo simulation of 100 iterations.

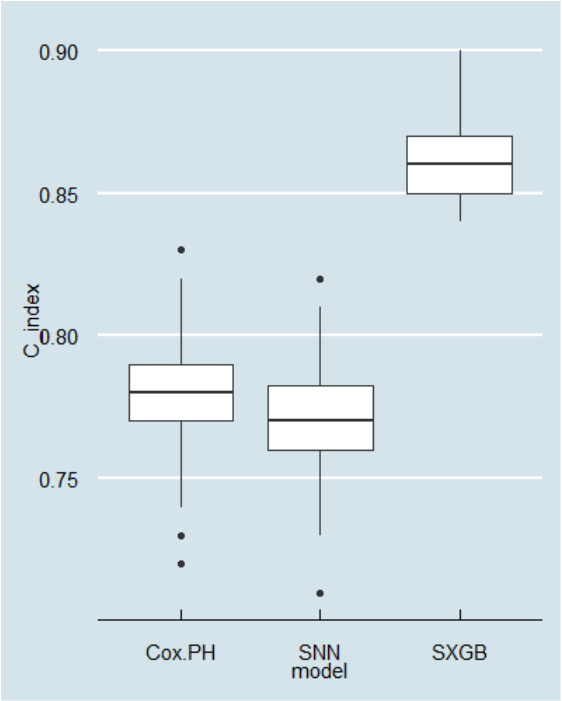

Fig. 2. Boxplots of C-Index performances for models applied to the MCI group, obtained in the Monte Carlo simulation of 100 iterations.

VI. Discussion

This study aimed to further explore the potential of survival-based ML as a tool for predicting time to AD diagnosis. This paper demonstrates the clear utility of such methods when predicting on the ADNI dataset. This supports the work of [8] [11] [14] [15] and thus provides further support for the utility of these survival based techniques in the context of dementia prediction modelling.

In line with [15] we found that a tree-based method proved superior to a survival neural network model, in terms of discrimination and calibration, when applied to the ADNI dataset.

This paper is one of only two known papers to explore Calibration in the context of survival ML in dementia. Calibration allows the quantification of a models' ability to closely match the probability distribution of the true outcomes in the dataset at hand. A poorly calibrated model may provide predictions that overestimate or underestimate the density of outcomes within the sample. In the context presented here, poorly calibrated models may lead to clinicians being misinformed as to the true number of patients who are likely to deteriorate resulting in a worse diagnosis. Conversely a well calibrated model allows for greater accuracy in decision making regarding clinical outcomes and the clinical interventions that may be implemented in response to those outcomes.

To our knowledge, this is one of only two papers to use survival XGBoost to model dementia progression. This paper found that XGBoost provided superior predictive ability to the standard Cox PH model. This is in contrast with [12] which found no difference between tree-based ensemble methods and the standard Cox model. This may be due to the way in which the trees were created with different variable selection and splitting criteria [15]. XGBoost is a powerful machine learning technique that uses an ensemble tree based method to build trees that allow progress through a gradient descent method. Gradient descent requires a loss function be defined, and in this instance the loss function was the negative log likelihood. However, as has been noted elsewhere [29] the relationship between the negative log-likelihood and the C-Index, the most commonly used metric in survival analysis, has not been fully proved. This means that we may be able to optimise this model to produce better predictive power if we use a loss function more aligned with the performance metric of interest. However, the DeepHit model used here, does indeed use a loss function that is statistically related to the C-Index, but this model did not produce superior performance, as compared to SXGB. On the other hand, Neural Networks usually perform better on image and audio classification rather than tabular data, such as the dataset used in this study [27] and this may explain the performance of this model in the current context.

This study agrees with the growing body of evidence showing that survival-base ML can and does outperform standard Cox PH models, when used to predict risk of deterioration in dementia. Despite these results being broadly in line with previous work [8] [11] [14] [15] there are a number of limitations to the current work that should be noted when interpreting the results found here. Firstly, this paper used a relatively small sample for the modelling process. Future work will need to validate the current results on a much larger dataset in order to establish the robustness of our findings. Secondly, the ADNI dataset is approximately 80% white and wholly based in the USA. Future work should therefore seek to validate these results on a more diverse dataset, with a particular focus on data from non-white participants, who are under-represented in the most commonly used datasets in this field. Future work should also aim to improve upon the explainability of the models developed here. In the clinical setting it is essential that health workers are afforded the opportunity to understand how diagnoses are reached so they are confident that they are correct. Therefore, it is essential that any clinical prediction model has the capabilities to allow clinicians to see how predictions are made. Extensions of the current paper should therefore explore the use of explainable AI methods for this purpose.

VII. Conclusion

This paper proposed an innovative survival ML and statistical based methodology and approach to predicting the time to Alzheimer's Disease diagnosis using the Alzheimer's Disease Neuroimaging Initiative ADNI dataset. In particular we explored survival ML models based on neural networks -

SNN, and on gradient boosting - SXGB, which were compared with the Cox PH model. The models were applied to predicting deterioration within two distinct groups of patients, those deemed cognitively normal and those with a formal diagnosis of Mild Cognitive Impairment at baseline. The stability of our models was investigated within a Monte Carlo simulation framework. Overall, the survival ML outperformed the statistical Cox PH model, with the best performances achieved by the gradient boosting based model SXGB.

REFERENCES

[1] 'Dementia Statistics Hub — Alzheimer's Research UK', Dementia Statistics Hub. https://www.dementiastatistics.org/ (accessed Oct. 04, 2023).

[2] T. Rittman, 'Neurological update: neuroimaging in dementia', J. Neurol., vol. 267, no. 11, pp. 3429–3435, Nov. 2020, doi: 10.1007/s00415-020-10040-0.

[3] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, 'Key challenges for delivering clinical impact with artificial intelligence', BMC Med., vol. 17, no. 1, p. 195, Oct. 2019, doi: 10.1186/s12916-019-1426-2.

[4] H. Musto, D. Stamate, I. Pu, and D. Stahl, 'A Machine Learning Approach for Predicting Deterioration in Alzheimer's Disease', in 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA), Pasadena, CA, USA, Dec. 2021, pp. 1443–1448. doi: 10.1109/ICMLA52953.2021.00232.

[5] D. Stamate et al., 'Applying Deep Learning to Predicting Dementia and Mild Cognitive Impairment', in Artificial Intelligence Applications and Innovations, vol. 584, I. Maglogiannis, L. Iliadis, and E. Pimenidis, Eds. Cham: Springer International Publishing, 2020, pp. 308–319. doi:10.1007/978-3-030-49186-4_26.

[6] 'ADNI — About'. https://adni.loni.usc.edu/about/ (accessed Feb. 19, 2023)

[7] D. R. Cox, 'Regression Models and Life-Tables', J. R. Stat. Soc. Ser. B Methodol., vol. 34, no. 2, pp. 187–202, 1972, doi: 10.1111/j.2517-6161.1972.tb00899.x.

[8] A. Spooner et al., 'A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction', Sci. Rep., vol. 10, no. 1, p. 20410, Dec. 2020, doi: 10.1038/s41598-020-77220-w.

[9] D. Stamate et al., 'A metabolite-based machine learning approach to diagnose Alzheimer-type dementia in blood: Results from the European Medical Information Framework for Alzheimer disease biomarker discovery cohort', Alzheimers Dement. Transl. Res. Clin. Interv., vol. 5, no. 1, pp. 933– 938, Jan. 2019, doi: 10.1016/j.trci.2019.11.001.

[10] O. Ajnakina, D. Cadar, and A. Steptoe, 'Interplay between Socioeconomic Markers and Polygenic Predisposition on Timing of Dementia Diagnosis', J. Am. Geriatr. Soc., vol. 68, no. 7, pp. 1529– 1536, Jul. 2020, doi: 10.1111/jgs.16406.

[11] D. Stamate, H. Musto, O. Ajnakina, and D. Stahl, 'Predicting Risk of Dementia with Survival Machine Learning and Statistical Methods: Results on the English Longitudinal Study of Ageing Cohort', in Artificial Intelligence Applications and Innovations. AIAI 2022 IFIP WG 12.5 International Workshops, Cham, 2022, pp. 436–447. doi: 10.1007/978-3-031-08341-9_35.

[12] M. Wang et al., 'Dementia risk prediction in individuals with mild cognitive impairment: a comparison of Cox regression and machine learning models', BMC Med. Res. Methodol., vol. 22, no. 1, p. 284, Nov. 2022, doi: 10.1186/s12874-022-01754-y.

[13] F. Sheikh et al., 'Prevalence of mild behavioral impairment in mild cognitive impairment and subjective cognitive decline, and its association with caregiver burden', Int. Psychogeriatr., vol. 30, no. 2, pp. 233–244, Feb. 2018, doi: 10.1017/S104161021700151X.

[14] G. Mirabnahrazam et al., 'Predicting Time-to-conversion for Dementia of Alzheimer's Type using Multi-modal Deep Survival Analysis', Neurobiol. Aging, vol. 121, pp. 139–156, Jan. 2023, doi: 10.1016/j.neurobiolaging.2022.10.005.

[15] Musto, H., Stamate, D., Pu, I., Stahl, D. (2023). Predicting Alzheimer's Disease Diagnosis Risk Over Time with Survival Machine Learning on the ADNI Cohort. In: Nguyen, N.T., et al. Computational Collective Intelligence. ICCCI 2023. Lecture Notes in Computer Science(), vol 14162. Springer, Cham. https://doi.org/10.1007/978-3-031-41456-5_53

[16] E. Longato, M. Vettoretti, and B. Di Camillo, 'A practical perspective on the concordance index for the evaluation and selection of prognostic time-to-event models', J. Biomed. Inform., vol. 108, p. 103496, Aug. 2020, doi: 10.1016/j.jbi.2020.103496.

[17] 'Key ADNI tables merged into one table — adnimerge'. https://adni.bitbucket.io/reference/adnimerge.html (accessed Oct. 04, 2023).

[18] L. G. Apostolova et al., 'ApoE4 effects on automated diagnostic classifiers for mild cognitive impairment and Alzheimer's disease', NeuroImage Clin., vol. 4, pp. 461–472, Jan. 2014, doi: 10.1016/j.nicl.2013.12.012.

[19] E. W. Steyerberg, Clinical Prediction Models, 2nd ed. New York, NY: Springer New York, 2019. doi: 10.1007/978-0-387-77244-8.

[20] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons, 'Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD Statement', Br. J. Surg., vol. 102, no. 3, pp. 148–158, Feb. 2015, doi: 10.1002/bjs.9736.

[21] C. Lee, W. Zame, J. Yoon, and M. van der Schaar, 'DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks', Proc. AAAI Conf. Artif. Intell., vol. 32, no. 1, Art. no. 1, Apr. 2018, doi: 10.1609/aaai.v32i1.11842.

[22] H. C. van Houwelingen, 'Validation, calibration, revision and combination of prognostic survival models', Stat. Med., vol. 19, no. 24, pp. 3401–3415, 2000, doi: 10.1002/1097-0258(20001230)19:24¡3401::AID-SIM554¿3.0.CO;2-2.

[23] 'R: The R Foundation'. https://www.r-project.org/foundation/ (accessed Oct. 04, 2023).

[24] 'Tidyverse packages'. https://www.tidyverse.org/packages/ (accessed Oct. 04, 2023).

[25] M. Kuhn, The caret Package. Accessed: Oct. 04, 2023. [Online]. Available: https://topepo.github.io/caret/

[26] 'Machine Learning in R - Next Generation'. https://mlr3.mlr-org.com/ (accessed Oct. 04, 2023).

[27] V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk, and G. Kasneci, 'Deep Neural Networks and Tabular Data: A Survey', IEEE Trans. Neural Netw. Learn. Syst., pp. 1–21, 2022, doi: 10.1109/TNNLS.2022.3229161.

[28] Avinash Barnwal, Hyunsu Cho & Toby Hocking (2022) Survival Regression with Accelerated Failure Time Model in XGBoost, Journal of Computational and Graphical Statistics, 31:4, 1292-1302, DOI: 10.108010618600.2022.2067548

[29] Elgui, K., Nowak, A., Robin, G. (2023). A Statistical Learning Take on the Concordance Index for Survival Analysis. Proceedings of The 26th International Conference on Artificial Intelligence and Statistics, in Proceedings of Machine Learning Research 206:4712-4731