FACULTY OF APPLIED SCIENCES
UNIVERSITY
OF WEST BOHEMIA

DEPARTMENT OF
CYBERNETICS

Bachelor's Thesis

Predicting Risk of Multiple Sclerosis Worsening

Marek Hanzl

PILSEN, CZECH REPUBLIC

2024

FACULTY OF APPLIED SCIENCES
UNIVERSITY
OF WEST BOHEMIA

DEPARTMENT OF
CYBERNETICS

Bachelor's Thesis

Predicting Risk of Multiple Sclerosis Worsening

Marek Hanzl

Thesis advisor
Ing. Lukáš Picek, Ph.D.

PILSEN, CZECH REPUBLIC 2024

Citation in the bibliography/reference list:
HANZL, Marek. *Predicting Risk of Multiple Sclerosis Worsening*. Pilsen, Czech Republic, 2024. Bachelor's Thesis. University of West Bohemia, Faculty of Applied Sciences, Department of Cybernetics. Thesis advisor Ing. Lukáš Picek, Ph.D.

ZÁPADOČESKÁ UNIVERZITA V PLZNI
Fakulta aplikovaných věd
Akademický rok: 2023/2024

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

(projektu, uměleckého díla, uměleckého výkonu)

Jméno a příjmení:	**Marek HANZL**
Osobní číslo:	**A21B0372P**
Studijní program:	**B0714A150005 Kybernetika a řídicí technika**
Specializace:	**Umělá inteligence a automatizace**
Téma práce:	**Predikce rizika zhoršení u pacientů s roztroušenou sklerózou**
Zadávající katedra:	**Katedra kybernetiky**

Zásady pro vypracování

1. Prostudujte metody a datové korpusy vhodné k predikci rizika zhoršení stavu pacientů s roztroušenou sklerózou.
2. Navrhněte a realizujte řešení pro automatickou analýzu a rozpoznávání při predikci rizika zhoršení pacientů s roztroušenou sklerózou a predikci kumulativní pravděpodobnosti zhoršení pactientů s roztroušenou sklerózou. Při návrhu zohledněte různé metody tj. klasické a *transformer-based*.
3. Proveďte kvalitativní a kvantitativní analýzu a vyhodnocení přesnosti navrženého systému.

Rozsah bakalářské práce: **30-40 stránek A4**
Rozsah grafických prací:
Forma zpracování bakalářské práce: **tištěná/elektronická**
Jazyk zpracování: **Angličtina**

Seznam doporučené literatury:

- PINTO, Mauro F., et al. Prediction of disease progression and outcomes in multiple sclerosis with machine learning. *Scientific reports*, 2020, 10.1: 21038.
- WANG, Zifeng; SUN, Jimeng. SurvTRACE: Transformers for survival analysis with competing events. In: *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. 2022. p. 1-9
- AIDOS, Helena, et al. iDPP@ CLEF 2023: The Intelligent Disease Progression Prediction Challenge. In: *European Conference on Information Retrieval*. Cham: Springer Nature Switzerland, 2023. p. 491-498

Vedoucí bakalářské práce: **Ing. Lukáš Picek, Ph.D.**
Výzkumný program 1

Datum zadání bakalářské práce: **17. října 2023**
Termín odevzdání bakalářské práce: **20. května 2024**

Západočeská univerzita v Plzni
Fakulta aplikovaných věd
L.S.

Doc. Ing. Miloš Železný, Ph.D.
děkan

Doc. Dr. Ing. Vlasta Radová
vedoucí katedry

V Plzni dne 17. října 2023

Declaration

I hereby declare that this Bachelor's Thesis is completely my own work and that I used only the cited sources, literature, and other resources. This thesis has not been used to obtain another or the same academic degree.

I acknowledge that my thesis is subject to the rights and obligations arising from Act No. 121/2000 Coll., the Copyright Act as amended, in particular the fact that the University of West Bohemia has the right to conclude a licence agreement for the use of this thesis as a school work pursuant to Section 60(1) of the Copyright Act.

In Pilsen , on 13 May 2024

. . . . . .

Marek Hanzl

Abstract

This work focused on designing, testing and developing an automated system for predicting the risk of worsening and cumulative probability of worsening of patients with multiple sclerosis. For this purpose, several datasets, models, and metrics were selected and evaluated. Multiple standard methods, e.g., Random Forest, Gradient Boosting, and even a novel transformer-based method, e.g., SurvTRACE, were used to predict the multiple sclerosis progression. The considerable performance increase was achieved by (i) hyper-parameter fine-tuning, (ii) validation procedure and (iii) data pre-processing. The functionality of the newly proposed system was tested and verified during the iDPP@CLEF challenge, which focused on providing clinicians with AI-based methods for better prediction of multiple sclerosis progression. Participation in the competition provided excellent opportunities to compare achieved results with the other competing teams. The accuracy was evaluated on the validation and test sets within the competition, where the proposed methods achieved two first places in four applied tasks. The methods achieved second and third place in the others. The best method based on the Random Forest algorithm achieved a mean C-Index of 0.834 when predicting the overall risk of worsening and a mean AUROC score of 0.881 when predicting the cumulative probability of worsening.

Keywords

Multiple Sclerosis • Artificial Intelligence • Survival Analysis • Gradient Boosting • Transformers

Abstrakt

Cílem této práce bylo navrhnout, otestovat a vyvinout systém pro automatickou predikci rizika zhoršení a kumulativní pravděpodobnosti zhoršení pacientů s roztroušenou sklerózou. Pro tento účel bylo vybranáno a otestováno několik datových sad, modelů a metrik. Pro predikci vývoje roztroušené sklerózy jsme využili standardních metod, tj. Random Forest, Gradient Boosting, ale i nově navrženého transformeru, tj. SurvTRACE, jež jsme dále významně zpřesnili díky (i) optimalizaci trénovacích hyperparametrů, (ii) zvolení vhodné validační procedury a (iii) předzpracováním dat. Funkčnost nově navrženého systému jsme ověřili v rámci soutěže iDPP@CLEF, zaměřené na pomoc lékařům při predikci vývoje nemoci použitím metod založených na umělé inteligenci. Účast v soutěži poskytla skvělé možnosti pro srovnání dosažených výsledků s dalšími tými, jež se problematikou zabývají. Přesnost jsme vyhodnotili jak na validační, tak na testovací sadě v rámci soutěže, kde jsme dosáhli dvou prvních míst z celkem čtyř úloh, kterých jsme se účastnili. Ve zbylých jsme získali druhé a třetí místo. Jako nejlepší se ukázala metoda založená na algoritmu Random Forest, která dosáhla průměrného C-indexu 0.834 při predikci celkového rizika zhoršení a průměrného skóre AUROC 0.881 při predikci kumulativní pravděpodobnosti zhoršení.

Klíčová slova

Roztroušená skleróza • Umělá inteligence • Analýza přežití • Gradient Boosting • Transformery

Acknowledgement

I would like to express my deep gratitude to my thesis advisor, Ing. Lukáš Pick, Ph.D., for his guidance, patience, and feedback. At the same time, I want to thank my father, who wholeheartedly supported me throughout my studies, and my encouraging friends and family.

Contents

1 Introduction

Multiple sclerosis is a chronic autoimmune disease characterised by progressive impairment of neurological functions, leading to a patient's gradual loss of motor, sensory, visual, or cognitive capabilities [1, 2]. It can lead to various physical and mental symptoms, while the progression and severity vary widely between individuals. The multiple sclerosis is widely spread and affects primarily young adults, especially women. Most people diagnosed are between 20 to 50 years old [3]. To provide context: Over 2.8 million people suffer from multiple sclerosis worldwide, and around 300 people are diagnosed with it daily. The overall costs related to multiple sclerosis treatment were estimated in 2019 at $85.3 billion, only in the US [4]. Heterogeneity of each patient's disease progression immensely increases the difficulty of selecting a proper medical treatment and motivates further clinical research. It urges a need for novel and reliable methods that should assist the clinical decision-making process and help advance the development of effective treatments, leading to better care of patients. This issue is explored, and different approaches are further discussed. An overall indicator of the patient's status has to be defined to provide dynamic predictions depending on a specific time horizon. These tasks can be solved using machine learning methods that are usually data-hungry. Hence, a suitable dataset has to be selected to train various predictive models and to evaluate their performance.

However, there exists an organization that shares the objectives of the thesis, the iDPP@CLEF[1] [5, 6] (Intelligent Disease Progression Prediction) lab, which aims to address this issue by opening international challenges. Proposed challenges aim to provide an evaluation ground for developing new artificial intelligence and machine learning techniques to detect patient complications early, stratify individuals according to their risk levels, and predict disease progression over time. The competition provides a highly curated dataset, presenting many opportunities to maximize predictive performance. Moreover, the competition defines suitable evaluation metrics and provides a diverse ground for comparison of results achieved by different approaches of competing teams. The 2023 edition offered three tasks:

[1] https://brainteaser.health/

1. *Predicting risk of disease worsening.*
2. *Predicting the probability of worsening at different time windows.*
3. *Impact of Exposition to Pollutants* – Amyotrophic Lateral Sclerosis.

The first two tasks share the research goal of this paper and the proposed methods were submitted to these tasks. The first task requires ranking subjects based on the overall risk of worsening, while the second task specifies the required predictions by explicitly assigning the cumulative probability of worsening at different time intervals. Data for both tasks provide pre-computed data of occurrence and time of worsening. These tasks, divided into two sub-tasks: A and B, provide two unique datasets. These differ in the definition of worsening based on the Expanded Disability Status Scale (EDSS) [7].

To address the risk and cumulative probability of multiple sclerosis worsening, several artificial intelligence-based standard survival analysis methods are selected which include Random Survival Forest and Gradient Boosting models [8]. Additionally, a novel transformer-based SurvTRACE model [9] is more thoroughly described and tested as well. Furthermore, an extensive overview of the hyper-parameter fine-tuning of these selected methods is provided to achieve peak performance. In a few cases, these trained models are combined by averaging their predictions to create new ensemble model methods. To allow for robust evaluation of overall performance, multiple metrics are selected based on the competition's proposal, i.e., Harrell's Concordance Index (C-Index) [10] for the first task of evaluating the overall risk and AUROC curve [11] with O/E ratio [12] for the second task of assessing the cumulative predictions. The methods with the highest validation score, i.e., the highest C-Index, were picked to produce the final predictions for the submissions. The achieved results are then extensively described and compared. In addition, a case study is provided to further qualitatively explain the best model's predictions.

Related Work 2

The problem of predicting a disease progression contains multiple specific concepts that are established in the following sections. The first section describes a variety of available datasets in the multiple sclerosis research domain, i.e., their characteristics, the dataset selected for development, and the reasons why it was picked. Compared to other machine learning domains, the topic of multiple sclerosis worsening prediction belongs to the field of survival analysis. The characteristics of this area of research are addressed in the second section, followed by a brief overview of typical machine learning models, e.g., Random Survival Forest or Gradient Boosting methods. Additionally, a deeper dive into an innovative transformer-based model, SurvTRACE [9], is provided. Finally, several key metrics used for the evaluation of predictive performance are defined and described, namely the C-Index, AUROC, and O/E Ratio, which are used to assess the results of newly proposed methods.

2.1 Datasets

Machine-learning-based predictive methods are typically considered data-hun-gry and require relatively large datasets. Thus, it is necessary to obtain a suitable, sufficiently large dataset, which would contain information well-describing the underlying dependencies between the current patient state and the probability of multiple sclerosis worsening. However, this problem domain is pretty unique, and the amount of freely available datasets is sparse (BioGPS[1]) and many of them have different goals. For instance, many datasets are tabular and provide vast statistical information about the disease progression of numerous patients. Conversely, many datasets specialize in multiple sclerosis predictions based on images from MRI scans. Therefore, several considerable datasets are discussed to highlight their advantages and disadvantages.

[1] http://biogps.org/dataset/tag/multiple%20sclerosis/

2.1.1 Tabular Datasets

iDPP Dataset. The data provided through the iDPP competition [5, 6] consists of tabular data with static and dynamic features describing the patient state from different perspectives. They are split between two datasets based on the sub-tasks A and B, which differ in definitions of worsening corresponding and EDSS definitions (Expanded Disability Status Scale) [7, 13], which provides an opportunity to test the difference and impact on predictive performance. These datasets include the medical history of 1,192 patients from two clinical institutions located in Pavia and Turin, Italy. The dynamic data span over 2.5 years and are split into different subsets. These comprise information on relapses, EDSS, Evoked Potentials, MRIs, and the multiple sclerosis course. The ground-truth data, i.e., patient outcomes, include the actual occurrence of worsening and the relative time of this occurrence. The provided datasets for both sub-tasks (A, B) were split into training and test sub-sets in approximately 80/20 ratio. The test dataset labels, i.e., outcomes, were kept secret until the submission deadline.

The sub-task A provides data about 440 patients for training and 110 for testing. In the case of sub-task B, information about 510 patients is available for training and 128 for testing. Even though the number of patients is relatively high, only a fraction of them include all types of medical records. For reference, there are only 103 (23.4%) patients in the training dataset A and 155 (30.4%) in the dataset B with medical records from all dataset sub-sets. Each medical record refers to a single entry (row) in the dataset. The number of unique patients and the number of their medical records for the dataset subsets are listed in Table 2.1. The amount of data present differs widely between patients, as indicated in Table 2.1. This significant data imbalance poses an issue in the classification stage, as standard classifiers face challenges when dealing with the class imbalance and often prioritize the larger classes and disregard the smaller ones, reaching a sub-optimal solution [14]. However, the imbalance is likely inherent to the problem. Furthermore, severe time gaps between clinical visits of many patients are present, leaving out important information about the time progression of the disease.

	Dataset	Static	Outcomes	EDSS	EP	Relapses	MRI	MS Type
Patients	A	440	440	439	153	259	279	210
Records		440	440	2,661	1,211	481	960	310
Patients	B	510	510	510	183	284	303	218
Records		510	510	3,069	1,522	553	966	325

Table 2.1: iDPP train dataset's characteristics. Counts of unique patients and medical records in data sub-sets, included in the provided A and B datasets.

The iDPP dataset provides numerous advantages. To name a few, it is freely accessible via the competition. It is rich in volume and diverse in the number of features. There are two very similar variants which provide compelling research opportunities. Furthermore, there is the possibility to compare different approaches of other competitors. Consequently, the dataset proved to be the best option and was later selected for development.

CHUC Dataset. Another suitable tabular dataset was proposed by Olivera et al. [2]. The dataset contains information about 187 distinct patients from a database of the Neurology Department of Centro Hospitalar e Universitário de Coimbra (CHUC). The general characteristics are provided in Table 2.2. In addition to having similar features as mentioned in the iDPP dataset, e.g., static information, EDSS, MRI, relapses, lesions, etc., the used database contained the whole medical history of the patients, including all physical exams and family history. This means that the provided features would be very versatile and could give a broader understanding of the disease mechanism. However, the number of unique patients is more than six times lower when compared to the iDPP dataset. Furthermore, the dataset is private and proprietary, and thus, it is unsuitable in the terms of this work.

Other Datasets. A dataset worth considering for this topic is *Motor evoked potentials for multiple sclerosis, a multiyear follow-up dataset* [15]. It consists of substantial tabular data of 5,586 visits from about 963 patients monitored throughout 6 years. According to the paper, it consists of static data, e.g., patient's sex, date of birth, etc. and dynamic data, which consists of only evoked potentials and EDSS scores. Unfortunately, it seems the dataset is not openly available at the time of writing this paper. Moreover, the dataset is comparatively less diverse than the iDPP dataset as it provides fewer unique features. Another promising dataset would likely be *MSDA-Core Dataset* [16]. However, it seems that the dataset has not been gathered yet at the time of writing this paper. Therefore, it is not possible to evaluate the potential usability of the dataset.

Patients	SP developed	Patient's characteristics	Visits in first 5 years
187	21 patients (11%)	Gender: 51 men (27%) Onset age: 31.10±10.54 Tracked years: 11.01±8.18 Annotated years 13.22±4.87	1st year: 1.57±0.93 2nd year: 1.25±1.27 3rd year: 1.14±1.06 4th year: 1.19±1.03 5th year: 1.28±0.91

Table 2.2: The CHUC dataset's characteristics [2]. SP stands for secondary progressive and describes a specific multiple sclerosis course.

2.1.2 Image Segmentation Datasets

Many available studies related to multiple sclerosis research aim at lesion image segmentation. If they openly publish their datasets, these are usually smaller and provide segmented images from MRI scans, with supplementary metadata describing general patient information. This is insufficient for this work's goal as it is aimed to provide a more general tool. For instance, these multiple sclerosis studies include:

- Brain MRI dataset of multiple sclerosis with consensus manual lesion segmentation and patient meta information [17].
- Multiple sclerosis lesions segmentation from multiple experts: The MICCAI 2016 challenge dataset [18].
- Statistical mapping analysis of lesion location and neurological disability in multiple sclerosis: application to 452 patient data sets [19].

2.2 Methods

The prediction of a disease progression, such as the risk of multiple sclerosis worsening, naturally belongs to the statistical sub-field of survival analysis. The following section provides a broad overview of machine learning methods used in this problem domain. Firstly, a description of survival analysis's main concepts and challenges is provided. Secondly, we will take a look into multiple machine learning methods used in the field, e.g., random-forest-tree-based or gradient-boosting-based methods. Finally, a novel deep-learning approach using a transformer-based method to solve the survival analysis task is extensively described.

2.2.1 Survival Analysis

Survival analysis is a branch of statistics which focuses on analyzing and understanding the data where the target is the time until an event of interest occurs - the survival time. One of the main challenges in this context are instances where event outcomes become unobservable after a certain amount of time or when, in some instances, no event occurs during the monitoring period at all [20]. This phenomenon is called censoring and, in general, can be produced by the following reasons: (i) a patient has not experienced the relevant outcome, such as relapse or death, by the end of the study; (ii) a patient is lost due to follow-up during the study period; (iii) a patient experiences a different kind of event that makes further follow-up impossible [21]. In this case, the assumed occurrence of an event happens after the end of the study. This is called right censoring, and it is typically the most common in survival data. We could visualize such an example of transferring real-world data

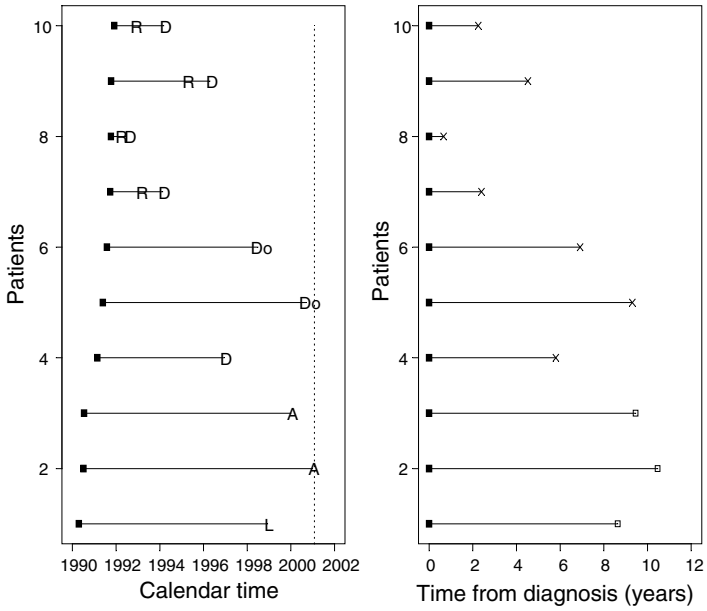

Figure 2.1: An example of censored data. (Left) – graph displays real-world patient data with different outcomes (*D* – death from cancer, *Do* – death from other cause, *A* – alive, *L* – did not attend, and *R* – relapse). The dashed vertical line represents the end of the study. (Right) – data converted in survival analysis format where the dates are normalized to the relative time since diagnosis and outcomes are encoded as × – death and □ – censored. Image taken from [21].

into survival format in Figure 2.1. Left censoring occurs when we observe the event, but it is unknown when it started. In other words, censoring occurs when complete information about the time-to-event is not available for some subjects before the start or by the end of an observation period.

The main goal of the survival analysis can be expressed as estimating the survival and hazard functions. The survival function S describes the probability that the target event does not occur before a specific time t and can be formulated as:

$$S(t) = P(T \geq t)\,. \tag{2.1}$$

Starting at 1, the function monotonically lowers to 0 [20]. The hazard function expresses the likelihood that the target event will occur at the time t if it has not occurred before. It can be defined as in the [9]:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}\,. \tag{2.2}$$

The integral from the expression represents the cumulative hazard function [20], and it is used to evaluate the cumulative probability of worsening.

To summarize, survival analysis is a regression problem with a twist – the data are censored. The problem is then addressed with different methods, which can be roughly divided into two main categories: statistical and machine-learning-based methods. The shared goal is to predict the survival time and to estimate the probability of survival at specific time points [20].

2.2.2 Traditional Approaches

Compared to statistical methods, machine learning techniques experience high popularity in a wide range of tasks due to the ability to model non-linear relationships and subsequently deliver accurate predictions [20]. Compared to regression models, the survival analysis models differ in data (time-to-event, censoring), the output (probability of an event over time), and model assumptions (proportional hazards or specific distributions of event times). Multiple frequently used methods, i.e., their survival analysis counterparts, are described below to provide context for the proposed solution to the multiple sclerosis prediction problem. Furthermore, many teams in the iDPP competition used such approaches with positive results.

Survival Trees. Survival trees are classification and regression trees designed to handle censored data. They operate on the principle of recursively partitioning data according to a specified splitting criterion, where similar objects are grouped within the same node. The key difference compared to the standard decision trees lies in the selection of splitting criterion that can, however, widely vary [20].

Random Survival Forest. The random survival forest is a specialized extension of random forests. It leverages ensemble learning principles and combines decision trees with survival analysis concepts [22]. The model fits several survival trees on various sub-samples of the dataset and averages between them to improve the predictive performance and to control the over-fitting [8]. The randomization reduces the correlation among the trees, improving the performance [20].

Gradient Boosting. The gradient-boosting does not refer to one particular model but is a versatile framework to optimize many loss functions [8]. It creates multiple base learners and combines their predictions to form a powerful overall model. Boosting is built from a series of simple models, weak learners, usually decision trees that are iteratively trained. Each learner fits the data, and the weights of the samples are updated based on the model's performance [22]. In each iteration, the weights of wrong or misclassified samples are increased, while the weights of correct predictions are decreased. The predictions obtained from all these weak learners are weighted and averaged to produce the final predictions.

Survival Support Vector Machine. The survival SVM extends the standard Support Vector Machine by handling the right-censored time-to-event data. They use the kernel trick to model complex, non-linear relationships between features and survival chance [8]. Survival support vector machines are viewed as a very successful supervised learning approach [20] that applies to a wide range of data.

2.2.3 SurvTRACE Transformer

Transformers are popular neural networks that introduce a so-called attention mechanism to comprehend long sequence data and to focus on specific parts of the input [23]. Even though originally designed for text translation, transformers were successfully adapted to many other machine-learning tasks, such as computer vision, natural language processing, dialogue systems, etc. The *SurvTRACE*, an acronym for Survival analysis using Transformers with Competing Events, is a recent and novel transformer-based model suited for survival analysis tasks [9]. It is an advanced multi-task transformer-based network designed for handling censored data and multiple competing risks. Competing risks emerge when it is acknowledged that a patient can suffer from several different diseases, and the result of diagnoses is not binary but can contain multiple events. In this case, it is even more challenging to estimate the effect on the predictions, and therefore, it is usually assumed that these events are independent, leading to a selection bias [9]. The SurvTRACE solves this issue using a distinct loss function – inverse propensity score [24], further described in the original paper. However, the datasets for the multiple sclerosis predictions have mostly a binary target and do not consider other events.

The model architecture is divided into three distinct components (see Figure 2.2). Starting with the Input & Embedding module, the raw data are divided into categorical and numerical sets and embedded via matrix multiplication as $t_i^c = V_c x_i^c$ (categorical) and $t_i^n = V_n x_i^n$ (numerical). These are concatenated and represent embedded information about the i-th patient. The second section is described as the Encoder module which leverages multi-head self-attention. The inputs are processed as follows:

$$\hat{t_i^j} = \sum_{k=1}^{D} \alpha_{j,k} (W_{value} t_i^k) \,, \tag{2.3}$$

where $\alpha_{j,k}$ is a product of softmax outputs of the attention function with two weight matrices W_{query} and W_{key}. The W_{value} is also a learnable weight matrix. This process can be stacked both vertically and horizontally. In addition, residual connections are established between the final horizontal stacks. The final embedding is obtained:

$$t_{res}^j = SELU(W_{res} \hat{t^j} + t^j) \,. \tag{2.4}$$

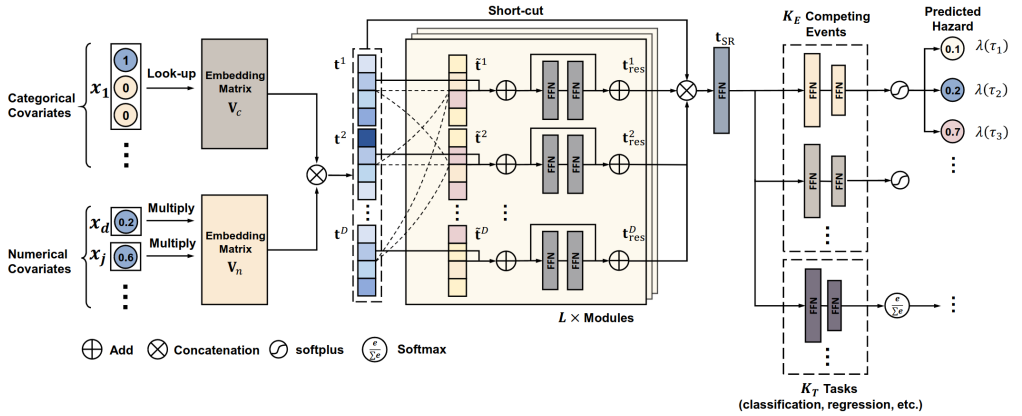

Figure 2.2: SurvTRACE architecture overview. The three main components are (i) Input & Embedding Module, (ii) Encoder Module, and (iii) Shared Representation & Sub-networks Module. Image is taken from [9].

The decision on the usage of the SELU (Scaled Exponential Liner Unit) is not elaborated on in the original paper. In the third module a shared representation t_{SR} is created from the encoder and is defined as:

$$t_{SR} = SELU(W_{SR}(\hat{t} \oplus t)) \,, \tag{2.5}$$

where $\hat{t}$ is the final output from the stacked transformers and is concatenated with the raw embeddings t. This output feature vector can be used for various downstream tasks which share the same input data but can vary in the prediction target. There are multiple learnable tasks described in [9], however, the Single-Event Survival Analysis, which aims to estimate the hazard function, will suffice for the means of this project. This hazard rate is formulated as $\lambda(\tau_j) = P(T = \tau_j | T > \tau_{j-1})$ where $T = \{\tau_1, ..., \tau_n\}$ is predefined set of time points. The sub-network uses the transformer output t_{SR} to predict the hazard rate prediction at specific time t:

$$\lambda(t) = log[1 + exp(f(\kappa(t)|t_{SR})] \,. \tag{2.6}$$

The $\kappa(t) = \{1, ..., n\}$ is a discrete index set of the time point set T. The model was originally trained on very large datasets (METABRIC [25], SUPPORT [26]) and compared to other traditional and deep learning models based on the time-dependent C-Index performance for the single event predictions. The results are presented in Table 2.3, where the CPH stands for the Cox Proportional Hazards model [27] and the RSF for Random Survival Forests [28]. The DeepSurv [29], DeepHit [30], PC-Hazard [31], and DSM [32] are other deep-learning-based models. It is clear, that the SurvTRACE transformer achieves comparatively the best results and surpasses all the other methods in different time intervals.

Algorithm	METABRIC			SUPPORT		
	25%	50%	75%	25%	50%	75%
CPH	0.628	0.627	0.632	0.549	0.564	0.586
DeepSurv	0.660	0.648	0.644	0.594	0.591	0.605
DeepHit	0.712	0.657	0.603	0.656	0.605	0.574
RSF	0.698	0.658	0.630	0.660	0.621	0.602
PC-Hazard	0.713	0.680	0.644	0.652	0.620	0.607
DSM	0.707	0.663	0.636	0.640	0.609	0.596
SurvTRACE w/o MTL	0.722	0.686	0.649	0.665	0.630	0.614
SurvTRACE	**0.728**	**0.690**	**0.655**	**0.670**	**0.633**	**0.617**

Table 2.3: SurvTRACE time-dependent C-Index benchmarks on METABRIC [25] and SUPPORT [26] datasets at different quantiles of event times (the % values). The best scores are in bold. The table is taken from [9].

2.3 Metrics

Eventually, the credibility of predictions has to be analysed and validated. Therefore, it is essential to define or select reliable metrics. There are two different tasks for which efficient metrics have to be provided. Following the selected dataset and the accompanying challenge, the C-Index [10] was used to validate the overall risk of worsening in patients with multiple sclerosis. Likewise, for the second task, the AUROC [11] and O/E ratio [12] were picked to evaluate the predictions of the cumulative probability of worsening in specific time intervals. A more detailed explanation of these terms is provided in the following section.

C-Index. The Harrell's concordance index or C-Index [10] is a metric that generalizes the AUROC (area under the receiver operating characteristic) by considering the possibility of censored data. Censoring occurs when the event of interest has not emerged by the end of the study. To explain the concordance, a pair of individuals is considered concordant if the individual with the higher risk score experiences the event (e.g., disease occurrence) sooner than the individual with the lower risk score. Conversely, a pair is discordant if the individual with the higher risk score experiences the event later than the individual with the lower risk score. The C-Index quantifies the proportion of concordant pairs to all informative pairs, measuring the model's ability to correctly rank individuals based on their predicted risk scores. In other words, it outlines how well a predicted risk score describes an observed sequence of events. It provides a comprehensive evaluation of the model's discrimination power, indicating its ability to reliably distinguish between different survival outcomes. As formulated in [5, 6], the value of C-Index can be determined as:

$$\text{C–Index} = \frac{\sum_{i=1}^{N} \Delta_i \sum_{j=i+1}^{N} I(T_i^{obs} < T_j^{obs}) I(M_i > M_j)}{\sum_{i=1}^{N} \Delta_i \sum_{j=i+1}^{N} I(T_i^{obs} < T_j^{obs})}, \tag{2.7}$$

where Δ_i is a binary variable which equals 1 if the subject i experienced the event at some point and 0 if event is censored, M is a predicted risk score of a subject, and T is a censoring of event times. The value of C-Index is $\in \langle 0, 1 \rangle$ where 1 represents perfect concordance. A C-Index = 0.5 is equivalent to the performance of random prediction, while the values below indicate a counter-correlation.

AUROC. The Receiver Operating Characteristic (ROC) curve is generated by tabulating sensitivities and specificities for various thresholds of a continuous test measure [33]. Sensitivity (true positive rate) is plotted against 1-specificity (false positive rate) to visually assess the diagnostic performance of the test. A curve above the diagonal line indicates better-than-chance performance. Sensitivity and specificity can be mathematically expressed as:

$$\text{sensitivity} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negative}}, \tag{2.8}$$

$$\text{specificity} = \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}}. \tag{2.9}$$

The AUROC is the area under the ROC and quantifies the test's ability to discriminate between conditions, with values ranging from 0 to 1, where 0 indicates a completely inaccurate model and a value of 1 indicates a perfectly accurate model that can distinguish between individuals who will experience worsening and those who will not. Conversely, a value of 0.5 suggests a classifier that assigns labels randomly. Therefore, a higher AUROC reflects a better ability of the model to discriminate between different outcomes [5, 6].

To address the survival data, an extended version of the AUROC is used, which acknowledges sensitivity and specificity as time-dependent measures. Defined in the *scikit-survival* [8] library, the cumulative dynamic AUROC at a time t is:

$$\widehat{\text{AUC}}(t) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} I(y_j > t) I(y_i \leq t) \omega_i I(\hat{f}(\mathbf{x}_j) \leq \hat{f}(\mathbf{x}_i))}{(\sum_{i=1}^{n} I(y_i > t))(\sum_{i=1}^{n} I(y_i \leq t) \omega_i)}, \tag{2.10}$$

where $\hat{f}(\mathbf{x}_i)$ is predicted risk score of i-th patient, and ω_i are weights derived from the train dataset.

O/E Ratio. The O/E (Observed-to-Expected) ratio is a measure commonly used in epidemiology, statistics, and medical research to assess the relationship between observed and expected values [12]. It provides a way of calibration for the model's predictions. It compares the actual number of observed worsening events to the number of events expected based on the model's predictions. Ideally, the O/E ratio should be close to 1, indicating good predictive performance and alignment between predicted and observed outcomes. A ratio significantly above 1 suggests that the observed events are occurring more frequently than expected, while a ratio below 1 indicates that the observed events are occurring less frequently than expected.

This ratio is especially useful in assessing the performance of diagnostic tests, evaluating the effectiveness of interventions, or studying the prevalence of diseases within specific populations. It provides a quantitative measure to understand whether observed outcomes deviate from what would be anticipated based on specific parameters or historical data. Monitoring the O/E ratio at each time interval allows for assessing the model's calibration performance over time [5, 6].

Methodology 3

This chapter describes and explains the steps taken during the assembly of the machine learning pipeline used to produce predictions of the progression of multiple sclerosis. Generating these predictions requires undergoing many data pre-processing steps in which numerous technical issues of the dataset are addressed. The data are converted into a form suitable for the model input. These steps are taken to maximize the prediction accuracy of machine learning methods and are followed by the selection of models capable of performing these predictions. Subsequently, the procedures employed for training, validating, and fine-tuning the models' hyper-parameters are specified as they are crucial to maximizing the predictive performance. The program workflow is visualised in a diagram in Figure 3.1, and comprises data pre-processing, model selection, training and validating the selected models, hyper-parameter fine-tuning, and combining best-performing models to ensembles.

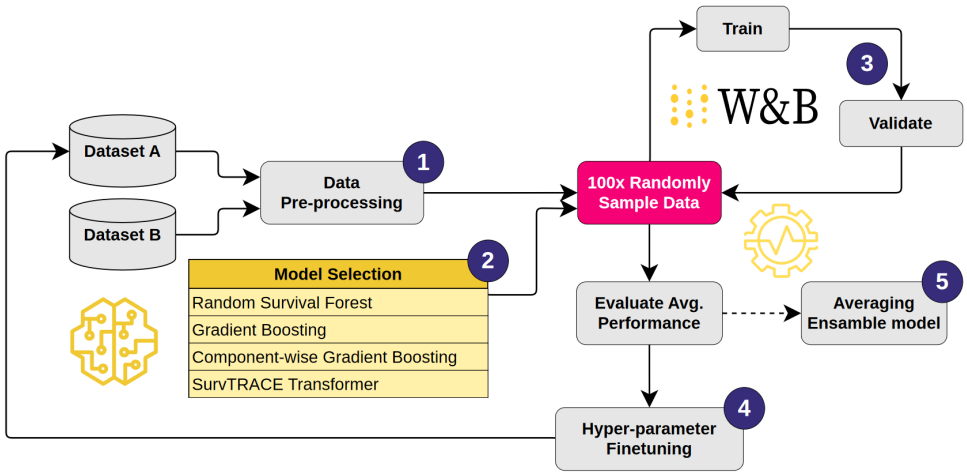

Figure 3.1: Machine-learning workflow. Stages of model training and validation. Methods consist of data pre-processing (1), selection of the specific model (2), and k-fold validation based on C-Index performance in 100 iterations (3). Model hyper-parameters are fine-tuned (4). Afterwards, additional methods are created by ensembling the best-performing models of each type (5).

The whole pipeline was programmed in the Python language, version 3.10. The source code is freely available in the GitHub repository[1]. All experiments were conducted on a Lenovo Notebook with a 6-core AMD Ryzen 5 5600H CPU and 16 GB of DDR4 RAM with a Windows 10 operation system.

3.1 Objective

The main objective of this paper is split between two tasks, namely, to predict the overall risk of worsening (Task 1) and to predict the cumulative probability of worsening over extended time windows (Task 2) of patients with multiple sclerosis. More precisely, the goal is to design and produce a machine learning program which would be able to assess the likelihood of a patient experiencing the event of worsening based on the provided clinical data. The single source of real-world data will be the iDPP 2023 dataset because it contains the most data and is openly available. Additionally, participation in the outgoing competition provides many research opportunities. Therefore, the tasks are inspired by the iDPP challenge [5, 6].

As the tasks belong to the field of survival analysis since only parts of the patient's outcomes are observable, the proposed methods will differ from traditional machine learning. For instance, in the case of clinical studies, the patients are usually monitored over a specific period. When an event occurs during this period, it can be precisely recorded. The data is uncensored. Alternatively, suppose no event is registered during the study period. In that case, the data are right-censored as it remains unknown whether or not an event occurred after the end of the study [8]. This atypical characteristic has to be considered during the method proposal. Furthermore, it is essential to formulate and define reliable metrics for both tasks to evaluate the performance of newly proposed methods.

3.1.1 Predicting Risk of Disease Worsening

The first task of this project focuses on the critical task of ranking subjects based on the overall risk of worsening, measured by a risk score ranging between 0 and 1, and it should reflect how early a patient experienced the "worsening" event. The definition of worsening is based on the EDSS [13], adhering to clinical standards, and is divided into two distinct sub-tasks. The effectiveness of predictions made by proposed methods will be measured using Harrell's Concordance Index (C-Index) [10] according to the competition rules. The definition is provided in Equation 2.7.

Task 1a. A patient experiences a worsening if the EDSS crosses the threshold of 3 ($EDSS \geq 3$) at least twice within an interval of a single year.

[1] https://github.com/Silvador386/IDPP2023

Task 1b. There are three specific instances when worsening occurs depending on the first recorded value of EDSS, according to the current clinical standards.

- EDSS $\in \langle 0, 1)$; worsening occurs when EDSS increments by 1.5 points.
- EDSS $\in \langle 1, 5.5)$; worsening occurs when EDSS increments by 1.0 points.
- EDSS $\in \langle 5.5, 10)$; worsening occurs when EDSS increments by 0.5 points.

3.1.2 Predicting Cumulative Probability of Worsening

The second task builds upon the first task and extends it by explicitly assigning the cumulative probability of worsening at different time windows between years (0, 2, 4, 6, 8, and 10). It is also divided between two sub-tasks (A and B) based on the two definitions of worsening. To evaluate the predictive performance of produced cumulative predictions, two following distinct metrics will be employed. Namely, the AUROC Curve and the O/E Ratio, described in Section 2.3.

3.2 Data Pre-processing

The success of machine learning in achieving optimal performance on a given task relies on various factors, with the representation and quality of the instance data being of critical importance. Effective knowledge discovery during training phases becomes arduous when irrelevant and redundant information or noisy and unreliable data are present. Data preparation and filtering steps, including data cleaning, normalization, transformation, feature extraction, and selection, are vital in achieving the best validation performance on the specific dataset [34]. However, there are many ways how to perform these steps. So how do we choose the best? In this case, the C-Index validation performance on the dataset A was employed as the ruling factor when selecting features or pre-processing methods. The premise is that the data in both datasets should be almost equivalent except for the EDSS feature definition, implying that the particular decision considering pre-processing should be generally valid for both cases.

First, all available information in the medical records is loaded from provided dataset files and subsequently grouped by unique patient ID. To be precise, many patients have multiple medical records based on the number of repetitions of different medical examinations. In this manner, the presence of all the available information concerning each patient is ensured. The features are divided into several groups derived from their characteristics, e.g., static and time-dependent features and categorical and numerical features. Static variables represent the time-invariant features.

If these are likewise categorical, then the one-hot encoding is applied, which elegantly solves the issue of missing values as the missing is its own category. Currently, these features are the patient's sex, residence, ethnicity, centre, multiple sclerosis diagnosis in pediatric age, and the record of the presence of several symptoms derived from their physical location. However, the *"time_since_onset"* and *"diagnostic_delay"* features tended to result in poorer performance and were, therefore, omitted.

In the case of time-dependent features, it is crucial to extract the temporal context of measured values. A *"sliding time-window"* segmentation approach [2] was implemented to achieve this goal, where a 6-month-long non-overlapping time window, together with a series of gradually extending, cumulative time windows (6, 12, 18, 24, and 30 months) are applied. This process is visualised in Figure 3.2. Various statistical functions were applied to the segmented features to extract the information. To name a few, mean, standard deviation, median, mode, and, in some instances, a sum of all occurrences of a feature was calculated. In the case of the EDSS and Relapses dataset sub-sets, the mean, one standard deviation, median, and mode of the EDSS score and the relapse occurrences were computed, respectively. Meanwhile, in the Evoked Potentials sub-set, a sum of all occurrences where the altered potential feature was positive and a sum of all occurrences where the altered potential feature was both positive and negative was measured. These steps should capture information about the percentage of diagnosed altered potentials. In the MS-type dataset section, generally, no more than two different recordings were present for each patient. These determine the diagnosed MS type and the time of the diagnosis. The one-hot encoding was tried and tested on an MS-type record, and the time of diagnosis was added. However, this did not lead to any improvement in performance, and thus, the data were omitted. The application of the MRI dataset section mostly led to similar results, and with a single exception, the data were unused.

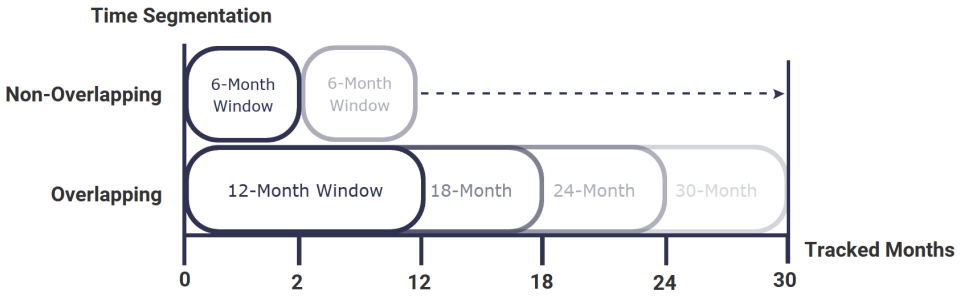

Figure 3.2: Visualisation of time-windows segmentation of the dynamic time-dependent data. For each dynamic feature (e.g., EDSS score, Evoked potentials, etc.), five 6-month non-overlapping and four accumulative (overlapping) time windows were created.

The missing feature values were handled in the last step of pre-processing. The numerical features were normalized to reduce the scale of the data, leading to an improvement in the numerical qualities of the dataset. The retaining the maximum amount of data available. Therefore, all the missing values were filled via the use of *Fast.ai* [35] functions, namely *TabularPandas*. This function creates a new categorical feature for every feature with at least one missing value. In this new feature, there are two categories representing missing or not missing values. In the original feature, the missing values are filled with a median of the whole feature. Afterwards, they are normalized by subtracting the mean and dividing by one standard deviation, which are both derived from the newly filled feature. Additionally, several methods were tested to fill in missing time-dependent values for each patient. Yet, they led to worse overall prediction performance.

3.3 Models

There are numerous approaches when estimating the survival function (Equation 2.1) or cumulative hazard function (Equation 2.2), which are addressed by traditional and deep learning survival analysis models. To design a method that could automatically produce the target predictions, it is important to select the most capable survival analysis machine learning models. General regression models can also be used to make predictions, however, based on the iDPP results [5, 6] and others, they tend to perform worse than their survival analysis counterparts, as they do not address the specific characteristics of the survival data. Therefore, they are omitted from the selection. The validation C-Index on the dataset A was used in the early stage of development as a ruling factor to narrow the model selection.

Various models from the popular Python library *scikit-survival* [8], built on top of *scikit-learn*, were tested to address the more traditional approaches. These include Survival Trees, Survival Support Vector Machines, and Ensemble Models. From these, the best-performing and the most stable were three ensemble models. Namely, Random Survival Forest, Gradient Boosting Survival Analysis, and Component-wise Gradient Boosting Survival Analysis (CGBSA). The linear Coxnet Survival Analysis was tested as well but failed due to linear dependencies in the pre-processed data. From deep learning approaches, a novel transformer [23] model suited for survival analysis, the *SurvTRACE* [9] transformer, was selected as it achieved the best results according to the benchmarks in Table 2.3. Other deep learning models were not tested. The deep learning models tend to require large training datasets to deliver good results and it is challenging to find these in survival analysis. Nonetheless, a large correctly trained transformer model should be able to contain all the available information, i.e., data and the time context, and should deliver good prediction capabilities.

Additionally, the best-performing, trained models of every model type were combined to create a new ensemble model and their predictions were averaged. This should lead to a compensation of extreme predictions made by these models, improving their stability. Likewise, it was attempted to use the maximal predicted values from combined models for prediction instead. However, this approach performed worse than the averaged predictions and was omitted.

3.4 Training & Validation Strategy

The training and validation procedure is described as follows: Each model was trained and validated in 100 iterations. In each iteration, the pre-processed dataset is randomly split into training and validation sets in an 80/20 ratio (i.e., the same ratio as in the provided training and test datasets). The C-Index is measured for both training and validation sets. Afterwards, the best model is chosen based on the achieved average and one standard deviation of the C-Index calculated across all the performed iterations. Furthermore, the *MinVal* 95/5 split ratio for several runs was experimented with. This meant that models were provided with an even larger proportion of available data in the training stage. Multiple iterations were performed to mitigate the effect of randomly split data, as it is likely that some splits are easier for model fitting and can lead to unreasonably high performance. Thus, it is better to evaluate model performance by averaging multiple iterations, reducing the effect of randomness by averaging it out. The transformer was trained with Adam optimizer for 40 epochs with 10-epoch early-stopping. Finally, the run with the highest validation C-Index value (one from 100 iterations) was selected to make the final predictions, i.e., predictions for the submission. The same seed was used for every run to prevent the introduction of an additional bias.

3.5 Hyper-parameter Fine-tuning

The selected models start with multiple available hyper-parameters, whose initial settings are presumably inadequate for the current task. Meaning, optimal model settings can be found through extensive iterative hyper-parameter search and lead to further maximization of validation performance.

The Weights & Biases framework [36] was used for this task, more precisely, the *"Sweeping tool"* with a mix of a random search for continuous hyper-parameters and grid search for categorical hyper-parameters in specific intervals. These were obtained by first applying a random search over a broader range of parameter space, which led to localizing the roughly optimal intervals. These intervals were then exhaustively searched through. The target was the maximum validation C-Index performance on the dataset A. The full list of all pre-selected search hyper-parameters

is provided in Table A.1[2] in the Appendix A. In total, approximately 3,500 separate runs were performed. After the extensive search through the parameters of selected models, performed in Table A.1, the optimal hyper-parameters converged to the following values:

- **Random Forest**:
  n_estimators=300, max_depth=6, min_samples_split=10,
  min_samples_leaf=3, min_weight_fraction_left=0.0, max_features='sqrt'
- **Gradient Boosting**:
  n_estimators=500, subsample=0.5, dropout_rate=0.2,
  ccp_alpha = 0.0, learning_rate=0.5, max_depth=3, min_samples_split=4,
  min_samples_leaf=1, min_weight_fraction_left=0.0, max_features='sqrt'
- **CGBSA**:
  loss=coxph, n_estimators=300, learning_rate=0.5, subsample=0.75,
  dropout_rate=0.2
- **SurvTRACE**:
  batch_size: 64, weight_decay: 0.0000284, learning_rate: 0.006157,
  hidden_size: 16, intermediate_size: 64, num_hidden_layers: 2,
  num_attention_heads: 4, hidden_dropout_prob: 0.2444,
  attention_probs_dropout_prob: 0.1143

	Random Forest	Random Forest MRI	Gradient Boosting	Component-wise Gradient Boosting	SurvTRACE
Before	0.731	0.738	0.741	0.719	0.561
After	0.741	0.747	0.750	0.725	0.698
Δ	+0.95%	+0.93%	+0.92%	+0.67%	+15.36%

Table 3.1: Improvement of the methods due to the hyper-parameter search measured by the average validation C-Index performance on the dataset A.

After setting the optimal hyper-parameter values, only marginal improvements for standard models were achieved, which likely indicates that the key part to achieving better performance lies in proper data pre-processing. On the other hand, the fine-tuning proved critical for the transformer's performance, as the average validation C-Index improved by 15.36%. However, the transformer predictions remained still quite volatile. The complete results of the hyper-parameter fine-tuning are available in Table 3.1.

[2]Default values were used for all non-mentioned hyper-parameters. For details, please refer to the scikit-survival documentation. Link to the original web page: https://scikit-survival.readthedocs.io/en/stable/api/index.html

3.6 Validation Results

This section covers the validation results of the best-performing methods, measured by the C-Index score on the validation set for both datasets (A, B). The previously discussed models (Random Forrest, Gradient Boosting, Component-wise Gradient Boosting, SurfTRACE transformer, and averaging Ensemble model) were used together with the *Random Forest MRI*. An extended version of pre-processed data was used in this method, enhanced with specific MRI data. For a few submissions, a *MinVal* strategy was tested, where the data were randomly split into training and validation sets in a 95/5 ratio during the K-Fold training. The rationale is that with a larger size of the training set, more information is provided to the models in the training stage, possibly decreasing the likelihood of over-fitting. The overall validation performance after 100 independent iterations is visible in the following boxplots in Figures 3.3 and 3.4.

In the case of the dataset A, the averaging Ensemble model achieved the top performance. This outcome is easily explained as the model already consists of the pre-trained, most well-performing models and is only tested for the average prediction accuracy over 100 separate iterations. The traditional models provide very similar results with only minor differences. However, the Gradient Boosting model delivers slightly better. Compared to the other methods, the transformer

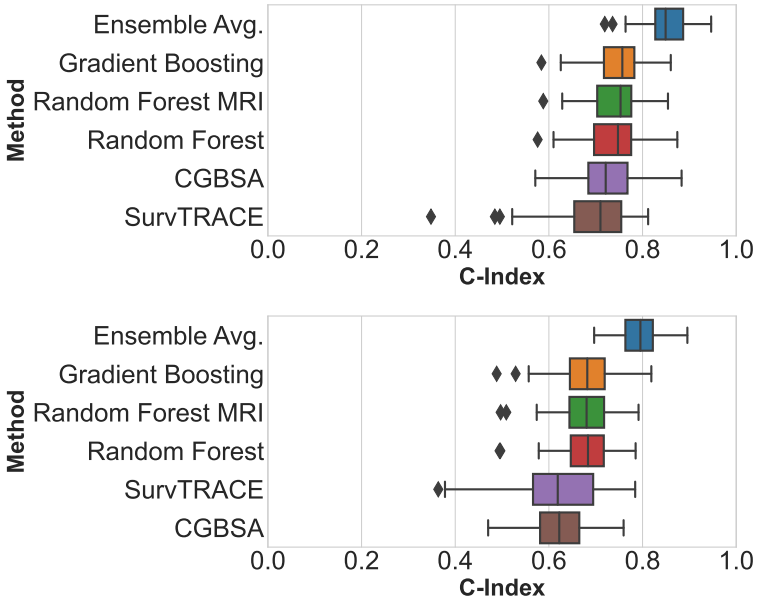

Figure 3.3: Results of validation C-Index performance of selected methods after 100 iterations with baseline splits. (Up) – dataset A, (Down) – dataset B. The results are sorted based on the best mean C-Index.

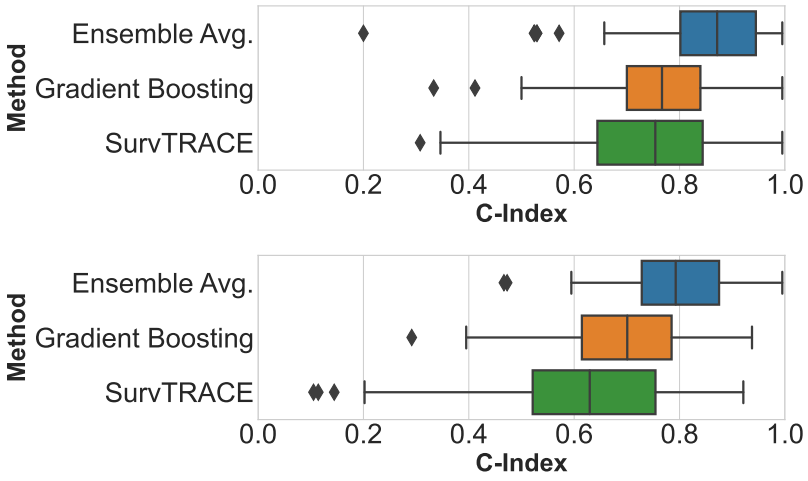

Figure 3.4: Results of validation C-Index performance of selected methods after 100 iterations with *MinMax* splits. (Up) – dataset A, (Down) – dataset B. The results are sorted based on the best mean C-Index.

suffers from substantial deviation, likely due to insufficient convergence and early stopping during model training. The predictive performance on the dataset B is generally worse than that of the dataset A, partially due to the optimisation being solely done on the dataset A. Although, the most substantial impact is presumably caused by the definition of the EDSS score, or possibly by the different data distribution, which would subsequently require different data pre-processing steps. Additionally, the CGBSA model performance comparatively drops even more. While the *MinVal* strategy led to a slight improvement in validation performance for both datasets, the strategy has an expected side-effect of amplified variance between different iterations. This is explicable by significantly reducing the size of the validation set, making it more unbalanced in the process.

The runs with the best overall validation C-Index were used to make predictions for final submissions for both tasks and their respective sub-tasks. These predictions were made on two distinct test datasets, A and B, which were originally provided without ground-truth outcomes. In total, 18 files for Task 1 and 10 files for Task 2 were submitted to the iDPP 2023 competition. It was permitted to submit up to 10 runs per sub-task.

Results 4

This section provides the official results of the submitted test runs for Task 1 and Task 2 and their respective sub-tasks. The scores achieved for the numerous results are discussed, and a case study of the best model's predictions is provided. Moreover, the final part compares the best test runs of the top 5 teams participating in the iDPP competition. Astonishingly, the proposed methods provided the best predictions in several tasks, with the best method achieving the test C-Index of 0.834. These submissions were selected in part based on the best-achieved validation C-Index values and in part to test all the previously proposed models and strategy variants.

The names of the displayed methods correspond to the model names. Concerning the competition, the names correspond to the *freefield*[1] section of the formal submission name described in the competition naming convention. To specify, survRf is the Random Forest, survGB is the Gradient Boosting, and AvgEnsemble is Ensemble Avg. The rest of the submission names remain more or less the same as defined.

4.1 Predicting Risk of Disease Worsening

Here follows the discussion of the results of the prediction accuracies described by the C-Index values and their 95% confidence intervals corresponding to the submissions made in Task 1, the sub-tasks (datasets) A and B. Additionally, a comparison of validation to test performance is provided to highlight the training effectiveness. The official results are displayed in Figure 4.1 together with the validation performance. Methods are sorted descendingly based on the mean C-Index. The dashed line corresponds to the score of predictions made by random choice.

Starting with the results on the dataset A, the overall highest C-Index score of 0.834 *(0.741–0.927)* was achieved by the Random Forest MRI model (enhanced with data from MRI scans), which was just slightly better than the second-best model, the Averaging Ensemble model which scored a C-Index of 0.828 *(0.739–0.917)*. The Random Forest MRI significantly surpassed its validation performance by roughly 8%,

[1] http://brainteaser.dei.unipd.it/challenges/idpp2023/

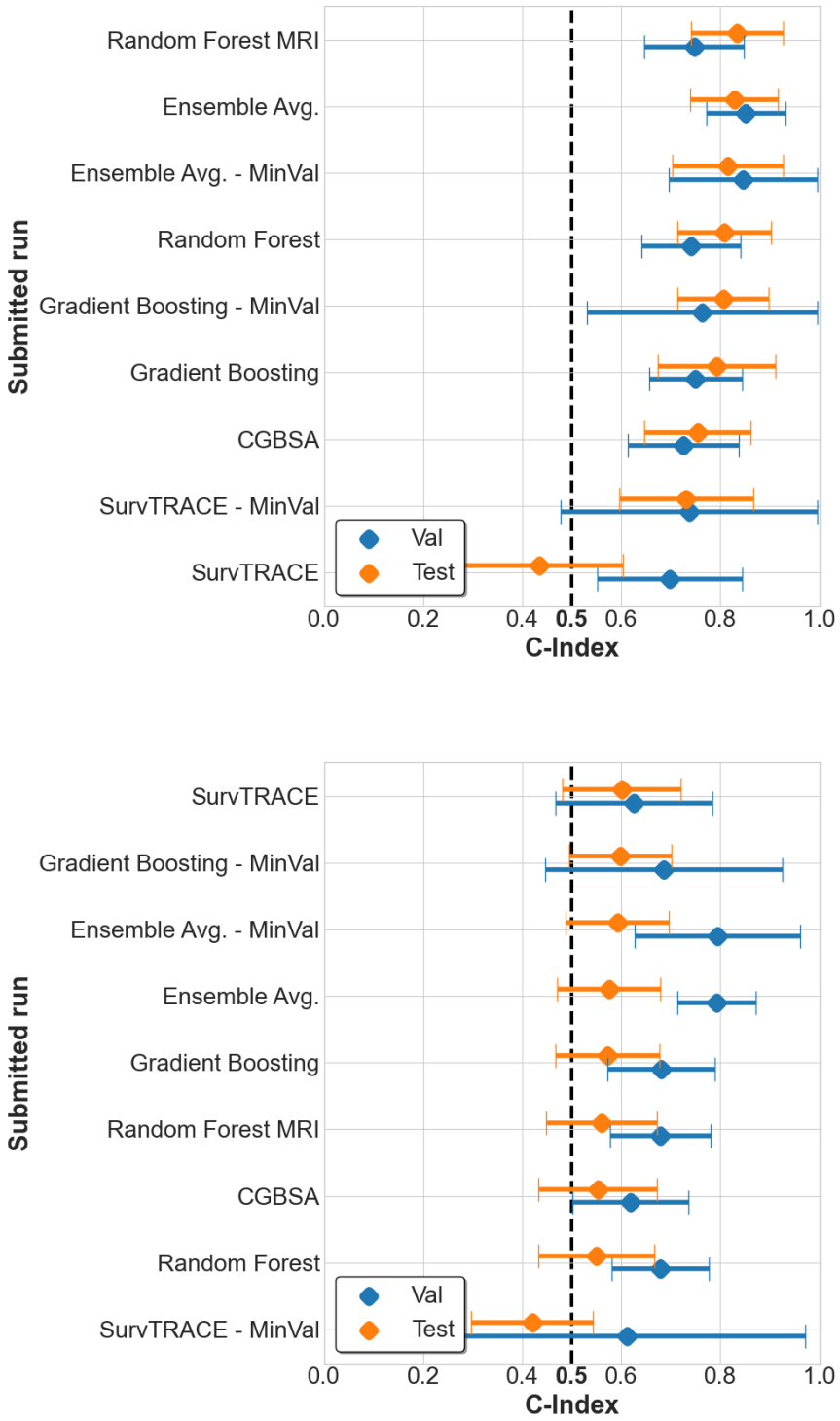

Figure 4.1: Achieved C-Index scores of submitted test runs and their respective 95% confidence intervals for both sub-tasks. (Up) – sub-task A (Task 1), (Down) – sub-task B (Task 1). Blue scores stand for validation results, and orange for test resultsForest. The dashed line represents random predictions.

leading to an assumption of a favourably selected model run. Contrarily, the Averaging Ensemble model slightly under-performed. However, the results are much closer to the expectations. Other submitted models perform similarly well. Although, a majority of them scored better than expected. It is also apparent that the *MinVal* strategy under-delivers, yet it keeps its high variance. At the lowest point is the SurvTRACE transformer model, completely missing the expectation, scoring even lower than the random chance. This likely leads to a conclusion that there must have been some unanticipated issue in the learning process.

Moving to the test results on the dataset B, the first place was acquired by the SurvTRACE transformer with the test C-Index of 0.601 *(0.482–0.721)*, delivering the closest performance to the expectation. Contrarily, the majority of other methods significantly under-delivers. For instance, the Averaging Ensemble performance decreased on the test dataset by almost 20% than was expected. A theoretical explanation of this unanticipated difference could be that the test dataset consisted of data differing in distribution from the training dataset. Additionally, it is possible that during the 100-randomly-sampled training process, the best-performing model runs were fitted to divergent data samples, leading to significantly worse performance on the test sample. The worst performance was again achieved by the SurfTRACE transformer, which is likely due to a selection of badly converged model runs for the submission. The sub-optimal convergence is explainable by a model over-fitting over a specific data subset.

4.2 Predicting Cumulative Probability of Worsening

In the second task of predicting the cumulative probability of worsening, the official results of prediction are provided to discuss the findings. The test scores consist of cumulative AUROC values, O/E ratios, and their respective 95% confidence intervals of all five submitted methods. These are based on tree models: Random Forest, Gradient Boosting, and CGBSA, and are separated into five distinct 2-year time intervals from 0 up to 10 years. The results are sorted in descending order based on the mean AUROC score produced from means of all target time intervals. The data is presented in two distinct forms. The first form is a set of two radial graphs where one depicts the AUROC score – the larger the covered surface, the better the method performed. The second one shows the O/E ratio, which represents the ratio of observed to expected events at each time interval. The ideal value is equal to 1 and is highlighted by the black circle. The results are supplemented by tables that contain detailed performance with precise mean values and the respective confidence intervals. The following name enumeration is applied:

I – Random Forest MRI
II – Random Forest
III – Gradient Boosting - MinVal
IV – Gradient Boosting
V – Component-wise Gradient Boosting

Starting with scores for sub-task A (dataset A), the Random Forest MRI, similarly to the first task, provides the best with the mean AUROC score of 0.881 which is visualized in Figure 4.2 and in Table 4.1.

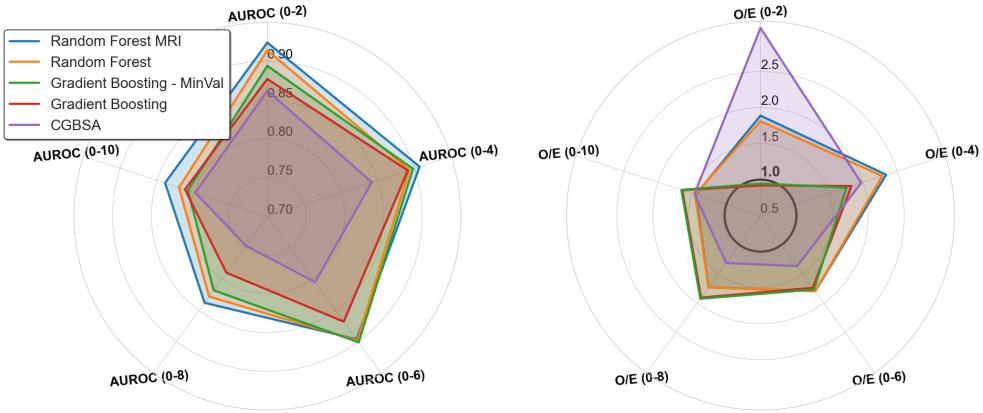

Figure 4.2: The test results of Task 2, dataset A. (Up) – average AUROC, (Down) – average O/E ratio. For the AUROC, the larger surface means better predictions. In this case, the best method is blue (Random Forest MRI). For the O/E ratio, the closer the value to 1, the better. In this case, the green method seems the best (Gradient Boosting - MinVal). Methods are ordered based on achieved mean AUROC.

ID	AUROC 0–2	AUROC 0–4	AUROC 0–6	AUROC 0–8	AUROC 0–10
I	0.924 (0.800-1.000)	0.907 (0.816–0.998)	0.896 (0.801–0.991)	0.838 (0.713–0.964)	0.839 (0.699–0.979)
II	0.914 (0.784–1.000)	0.893 (0.798–0.989)	0.898 (0.808–0.989)	0.828 (0.702–0.954)	0.820 (0.672–0.968)
III	0.894 (0.787–1.000)	0.898 (0.810–0.985)	0.901 (0.800–1.000)	0.818 (0.677–0.959)	0.808 (0.648–0.967)
IV	0.877 (0.745–1.000)	0.891 (0.796–0.986)	0.868 (0.753–0.984)	0.790 (0.641–0.938)	0.812 (0.654–0.969)
V	0.862 (0.731–0.993)	0.842 (0.713–0.971)	0.805 (0.670–0.941)	0.747 (0.597–0.898)	0.798 (0.649–0.947)
ID	**O/E Ratio 0–2**	**O/E Ratio 0–4**	**O/E Ratio 0–6**	**O/E Ratio 0–8**	**O/E Ratio 0–10**
I	1.889 (0.937–2.842)	2.339 (1.391–3.287)	1.797 (1.068–2.525)	1.731 (1.065–2.396)	1.447 (0.875–2.019)
II	1.811 (0.879–2.744)	2.283 (1.347–3.220)	1.796 (1.067–2.524)	1.732 (1.066–2.398)	1.458 (0.884–2.032)
III	0.946 (0.272–1.620)	1.759 (0.937–2.581)	1.768 (1.045–2.490)	1.926 (1.224–2.628)	1.658 (1.046–2.270)
IV	0.919 (0.255–1.583)	1.831 (0.993–2.670)	1.739 (1.022–2.455)	1.906 (1.207–2.604)	1.644 (1.035–2.254)
V	3.106 (1.885–4.327)	1.975 (1.104–2.847)	1.366 (0.731–2.002)	1.312 (0.732–1.891)	1.467 (0.891–2.042)

Table 4.1: The AUROC and O/E ratio cumulative test scores in the given time intervals for all submitted methods to the Task 2, sub-task A. Methods are denoted based on the ID described in the accompanying text.

Moreover, it even managed to score more than 90% in the first two years (0.924) and the second two years (0.907). Conversely, the O/E ratio is relatively high across all the predictions. This likely indicates a practical issue when there is a greater portion of observed patient events than was expected. Other methods performed similarly well. In the case of the Gradient Boosting methods, the MinVal strategy is better than the default one, in contrast to the first task. Additionally, they both perform the best in the case of the O/E ratio.

Lastly, the test results for sub-task B are provided in Figure 4.3 and Table 4.2. The Gradient Boosting MinVal strategy method scored first with a mean AUROC of 0.607 while delivering one of the best O/E ratios. The overall results are yet again significantly worse than would be expected from scores achieved on dataset A. A similar trend is observable here as in Task 1, sub-task B, where the scores worsen across all predictions. This behavior could be explained with the same reasoning as previously discussed. These are:

1. Optimisation based on the validation C-Index for dataset A during training.
2. Different definitions of the EDSS feature.
3. Different properties of the test data or fitting of the best-performing model on very divergent data subset.

ID	AUROC 0–2	AUROC 0–4	AUROC 0–6	AUROC 0–8	AUROC 0–10
III	0.606 (0.437–0.776)	0.612 (0.468–0.756)	0.602 (0.451–0.754)	0.587 (0.433–0.742)	0.626 (0.465–0.787)
V	0.514 (0.311–0.717)	0.580 (0.423–0.737)	0.604 (0.452–0.756)	0.627 (0.477–0.777)	0.628 (0.463–0.793)
IV	0.569 (0.392–0.747)	0.597 (0.454–0.741)	0.589 (0.440–0.737)	0.580 (0.427–0.733)	0.594 (0.430–0.758)
I	0.596 (0.421–0.770)	0.561 (0.407–0.715)	0.559 (0.407–0.711)	0.525 (0.369–0.681)	0.491 (0.324–0.658)
II	0.590 (0.410–0.769)	0.552 (0.401–0.704)	0.549 (0.398–0.700)	0.522 (0.367–0.678)	0.506 (0.340–0.672)
ID	**O/E Ratio 0–2**	**O/E Ratio 0–4**	**O/E Ratio 0–6**	**O/E Ratio 0–8**	**O/E Ratio 0–10**
III	0.920 (0.353–1.486)	1.228 (0.716–1.740)	1.375 (0.896–1.854)	1.430 (0.979–1.880)	1.489 (1.052–1.926)
V	1.818 (1.021–2.615)	0.774 (0.367–1.180)	1.515 (1.012–2.017)	1.295 (0.866–1.724)	1.166 (0.780–1.553)
IV	1.045 (0.441–1.649)	1.259 (0.741–1.778)	1.363 (0.886–1.840)	1.404 (0.957–1.850)	1.454 (1.022–1.885)
I	2.257 (1.370–3.145)	1.525 (0.955–2.096)	1.364 (0.886–1.841)	1.302 (0.872–1.732)	1.316 (0.905–1.726)
II	2.292 (1.398–3.187)	1.523 (0.953–2.093)	1.351 (0.876–1.826)	1.304 (0.873–1.734)	1.313 (0.903–1.724)

Table 4.2: The AUROC and O/E ratio cumulative test scores in the given time intervals for all submitted methods to the Task 2, sub-task B. Methods are denoted based on the ID described in the accompanying text.

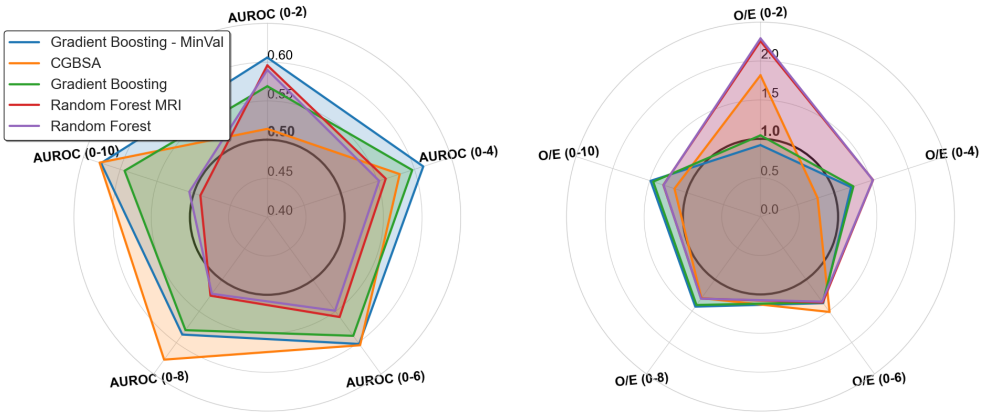

Figure 4.3: The test results of Task 2, dataset B. (Up) – average AUROC, (Down) – average O/E ratio. For the AUROC, the larger surface means better predictions. The 0.5 circle shows random prediction. In this case, the best method is blue (Gradient Boosting - MinVal). For the O/E ratio, the closer the value to 1, the better. In this case, the blue method seems the best as well. Methods are ordered descendingly based on achieved mean AUROC.

4.3 Case Study

To better understand the prediction mechanism, we can visualize the estimated survival functions of the best-performing model Random Forest MRI for two sets of patients in Figure 4.4. Both sets contain four randomly selected patients from the test set of the dataset A. Some are censored due to the end of the study or missing follow-up, and in some cases, the worsening occurred. The survival models try to estimate the real survival function as stated in Equation 2.1 for the whole time frame. However, the predicted values are not as important as the relative order of survival functions in the relationship to the time of worsening or censoring. It should be noted, that both C-Index (Task 1) and AUROC (Task 2) are evaluated based on the pair ranking of estimated survival probabilities. Consequentially, evaluating specific cases is complicated as the scores have meaning only when compared to other predictions. In other words, the AUROC measures how many pair estimates are correctly ordered in a specific time t, while the C-Index evaluates the whole time frame. At the time of worsening of an individual, their estimated probability of survival should be the lowest compared to others.

Starting with the upper graph, the predictions are almost flawless relative to each other. The first worsening event C) occurred in the function with the lowest probability of survival at the time. The same is true for D). The survival functions of censored data should be above the survival functions that ended in worsening at the time of censoring, which is also true for both shown cases. The average AUROC

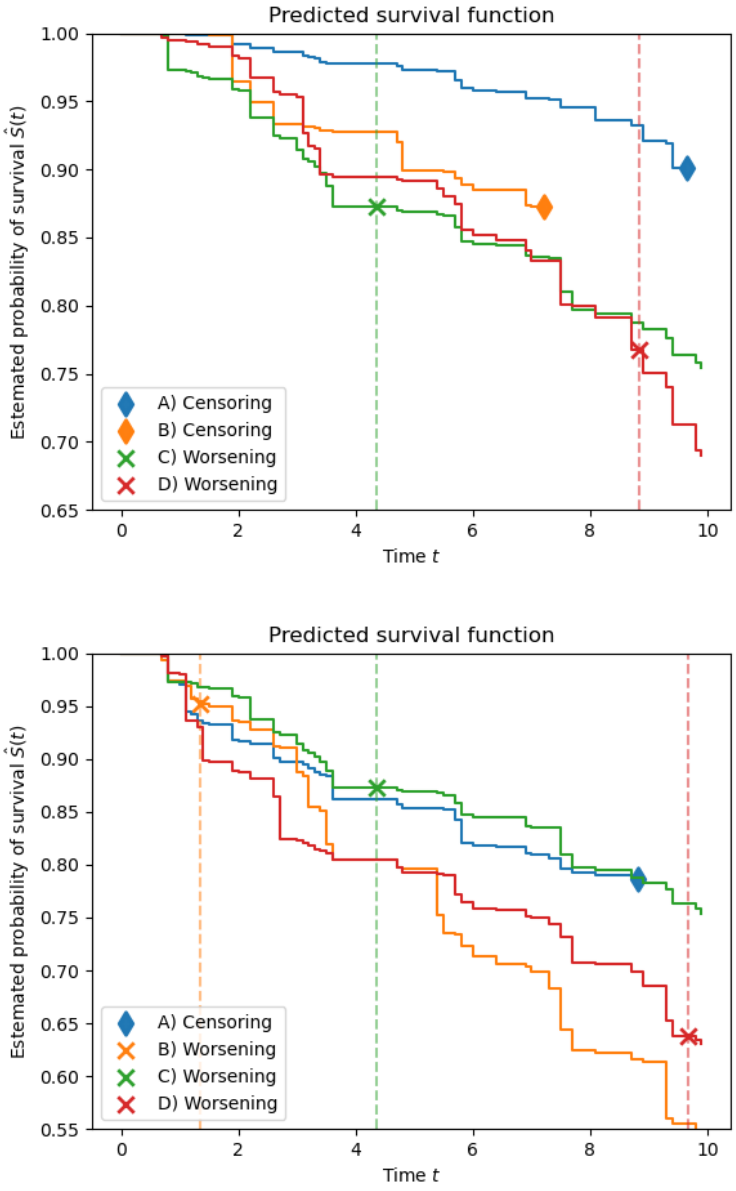

Figure 4.4: Example of survival functions estimated by Random Forest MRI for 4 randomly selected patients per graph, from the original 110 of the test dataset A, Task 2. (Up) – Correct predictions, (Down) – Wrong predictions – in the relative context to each other. The markers show the ground truth time of censoring or occurrence of worsening.

score calculated by Equation 2.10 drawn solely from these four samples reaches near 1.0. In the lower graph are examples of very poor predictions relative to each other. The first event of worsening B) has the second largest estimated probability of survival at the time of the event, even though it should be the lowest. In the second event C), the estimated survival probability should be the lowest, but it is above the A), and D). The survival function of the censored data A) should have ideally been above the others at times of worsening. The average AUROC score drawn solely from these four samples reaches only about 0.35. That is an extremely poor result compared to the average AUROC of 0.881 of the whole test set (110 individuals).

4.4 Competition Results & Comparison

Thanks to the participation in the iDPP 2023 challenge, it is possible to evaluate the achieved results in comparison with other participating teams. The proposed methods delivered the overall best performance in several tasks while remaining competitive in the others. From 45 registered teams, 9 made submissions into at least one of Task 1 and Task 2. One hundred twenty-four runs were submitted into these tasks in total (76 and 48) [5, 6]. We decided to show and compare runs with the best predictive performance from the top 5 participating teams for each task and their respective sub-tasks.

Starting with Task 1, the results are presented as the mean test C-Index score for both sub-tasks. Each submission is named by a team shortcut and with an abbreviation of the method name. These are shown in Table 4.3 for sub-tasks A and B. They are sorted by the by the highest mean C-Index. In sub-task A, the methods delivered the best performance of mean C-Index 0.834, surpassing the competition by 3.2% with the Random Forest MRI method. It was the highest achieved performance altogether, leading to the conclusion that our methods were designed and optimized appropriately for this task. In the second sub-task, we scored third place with a mean C-Index of 0.601 with the SurvTRACE transformer method, still providing comparable performance even for the not-optimized dataset. This leaves a considerable space for further improvement. Other well-placed methods provided by competing teams are also based on Random Survival models or completely different models, namely Fast Kernel SVM models [37] or CoxNet models [38]. These provide further options to test and experiment on.

Moving to the competition results for Task 2, we again selected the best-performing method, best on mean AUROC score, per every top 5 team. The methods are sorted descendingly by averaged AUROC score over all time intervals. They are displayed in Table 4.4. In sub-task A, our Random Forest MRI method delivered the highest AUROC score with an average of 0.881, while providing the best predictive performance in the first 2 and 4 years, even surpassing the 90% threshold

(0.924% and 0.907% respectively). These are exceptional results and point to a good optimization. In sub-task B, we placed in second place with an average AUROC score of 0.607 across all intervals with the Gradient Boosting MinVal method. This is only about 5.24% behind the first competing team, which implies that the method delivers competing predictions even though it was not primarily optimized for this metric. All achieved results and findings were presented at the Conference and Labs of the Evaluation Forum (CLEF 2023) in Thessaloniki, Greece.

	Task 1 A		Task 1 B	
Rank	**Submissions**	**C-Index**	**Submissions**	**C-Index**
1.	**uwb_T1a_survRFmri**	**0.834**	fcool_T1b_FastKernelSurvSVM	0.690
2.	CBMUniTO_T1a_coxnet	0.802	CBMUniTO_T1b_coxnet	0.634
3.	fcool_T1a_RandomSurvivalForest	0.801	**uwb_T1b_SurvTRACE**	**0.601**
4.	HULATUC3M_T1a_survcoxnet	0.774	uhu-etsi-1_T1b_s02	0.598
5.	sisinflab-aibio_T1a_RF2	0.771	sisinflab-aibio_T1b_GB2	0.587

Table 4.3: Official competition results – risk of worsening – for top 5 teams for Task 1, sub-task A and sub-task B, sorted by mean C-Index. The results of the proposed methods are in bold.

	Task 2 A				
Rank	**Submission**	**2 years**	**4 years**	**6 years**	**8 years**	**10 years**
1.	**uwb_T2a_survRFmri**	**0.924**	**0.907**	**0.896**	**0.838**	**0.839**
2.	HULATUC3M_T2a_survcoxnet	0.864	0.898	0.938	0.859	0.831
3.	CBMUniTO_T2a_coxnet	0.890	0.900	0.856	0.787	0.796
4.	sisinflab-aibio_T2a_RF1	0.754	0.873	0.871	0.746	0.745
5.	uhu-etsi-1_T2a_05	0.774	0.740	0.774	0.703	0.722
	Task 2 B					
Rank	**Submission**	**2 years**	**4 years**	**6 years**	**8 years**	**10 years**
1.	CBMUniTO_T2b_cwgbsa	0.632	0.626	0.655	0.673	0.709
2.	**uwb_T2b_survGB_minVal**	**0.606**	**0.612**	**0.602**	**0.587**	**0.626**
3.	sbb_T2b_Cox	0.642	0.567	0.601	0.594	0.622
4.	sisinflab-aibio_T2b_GB2	0.614	0.639	0.629	0.616	0.527
5.	uhu-etsi-1_T2b_s02	0.644	0.590	0.610	0.567	0.609

Table 4.4: Official competition results – cumulative probability of worsening – for top 5 teams. Mean AUROC score for Task 2, sub-task A and sub-task B. The results of the proposed methods are in bold.

Conclusion 5

This work was dedicated to studying, designing, and advancing the options to estimate the overall risk and the cumulative probability of worsening of patients with multiple sclerosis. Various openly available multiple sclerosis datasets, machine learning models, and metrics were searched and described. To provide further comparison, the proposed methods were submitted to the iDPP 2023 challenge, which focused on providing clinicians with new machine-learning-based methods to predict the progression of multiple sclerosis.

An extensive overview of the newly designed methods was provided, followed by the achieved results based on the defined metrics. The proposed data analysis, pre-processing, feature selection methods and strategies, model selection, training and validation strategy, and hyperparameter search were fully discussed. Likewise, the steps to assemble the entire machine-learning pipeline were extensively described. The best-performing survival-analysis-based models were selected for development, comprising random-forest-based and gradient-boosting-based decision tree methods alongside the recent SurvTRACE transformer. To further enhance baseline performance, an extensive hyper-parameter search for picked methods was conducted, and the benefits of fine-tuning were described, resulting in a 16% improvement of transformer predictions and approximately 1% improvement in the others (Table 3.1).

The performance was measured and optimized based on the mean validation C-Index achieved after 100 randomly sampled separate iterations on the dataset A. These metrics became a ruling factor for most of the decisions. The runs with the highest score were then used to make the final submissions to the competition. The achieved validation results (Figures 3.3 and 3.4) were thoroughly discussed. These were then compared with the test results in Figures 4.1 for the first task and in Figures 4.2, and 4.3 for the second task. Furthermore, a case study of the predictions estimated by the best model – Random Forest MRI, was conducted, with an example provided in Figure 4.4.

Participation in the competition proved to be valuable for the comparison of achieved performance. The results of the submitted methods were compared with

the top 5 competing teams in Table 4.3, for Task 1, and in Table 4.4 for Task 2. The best predictive performance was achieved in multiple categories, and the findings were summarized. In terms of the C-Index and average AUROC, the first place was scored in both A sub-tasks with a C-Index of 0.834 and a mean AUROC score of 0.881, respectively. In the case of sub-task B, results comparable with others were achieved. In Task 1, third place was attained with a C-Index of 0.601, and second, in Task 2 based on the average AUROC score of 0.607. Considerable robustness towards overfitting on a specific dataset was demonstrated, as the proposed methods achieved third and second place, even though they were purely designed and optimized for sub-task A. This possibly suggests the importance of the EDSS score definition. Furthermore, the proposed design and findings were presented at the Conference and Labs of the Evaluation Forum (CLEF 2023) in Thessaloniki, Greece.

Searched Hyper-parameters

ML Method	Hyper-parameters	Values	Search method
Random Forest	n_estimators	[100, 300, 500]	Grid search
	max_depth	[6, 8, 10]	
	min_samples_split	[8, 10, 15]	
	min_samples_leaf	[4, 6, 8]	
	min_weight_fraction_leaf	[0.0, 0.3]	
	max_features	[sqrt, log2]	
Gradient Boosting	n_estimators	[100, 200]	Grid search
	subsample	[0.2, 0.5, 1]	
	dropout_rate	[0, 0.2]	
	ccp_alpha	[0, 0.1, 1]	
	learning_rate	[0.1, 0.5]	
	max_depth	[3, 5, 7]	
	min_samples_split	[2, 4]	
	min_weight_fraction_leaf	[0.0, 0.3]	
	max_features	[sqrt, log2]	
CGBSA	loss	[coxph, squared, ipcwls]	Grid search
	n_estimators	[100, 200, 300, 500]	
	learning_rate	[0.1, 0.5, 1]	
	subsample	[0.2, 0.5, 1]	
	dropout_rate	[0, 0.2, 1]	
SurvTRACE	batch_size	(48–128)	Random search
	weight_decay	($5e^{-5}$–$1e^{-3}$)	
	learning_rate	($5e^{-4}$–$1e^{-2}$)	
	hidden_size	[16, 32, 64]	
	intermediate_size	[64, 128, 256, 512]	
	num_hidden_layers	[2, 4, 6]	
	num_attention_heads	[2, 4, 6]	
	hidden_dropout_prob	(0.2–0.4)	
	attention_probs_dropout_prob	(0.1–0.3)	

Table A.1: Hyper-parameter values of selected methods used for fine-tuning. Evaluation of performance was based on achieved C-Index values on the dataset A.

Bibliography

1. GOLDENBERG, Marvin M. Multiple sclerosis review. *Pharmacy and therapeutics*. 2012, vol. 37, no. 3, p. 175.
2. PINTO, Mauro F et al. Prediction of disease progression and outcomes in multiple sclerosis with machine learning. *Scientific reports*. 2020, vol. 10, no. 1, p. 21038.
3. BRUST, John CM. *Current diagnosis & treatment neurology*. McGraw Hill Professional, 2018.
4. WALTON, Clare et al. Rising prevalence of multiple sclerosis worldwide: Insights from the Atlas of MS, third edition. *Multiple Sclerosis Journal*. 2020, vol. 26, no. 14, pp. 1816–1821. PMID: 33174475.
5. FAGGIOLI, G. et al. Intelligent Disease Progression Prediction: Overview of iDPP@CLEF 2023. In: ARAMPATZIS, A. et al. (eds.). *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023)*. Lecture Notes in Computer Science (LNCS), Springer, Heidelberg, Germany, 2023.
6. FAGGIOLI, G. et al. Overview of iDPP@CLEF 2023: The Intelligent Disease Progression Prediction Challenge. In: ALIANNEJADI, M.; FAGGIOLI, G.; FERRO, N.; VLACHOS, M. (eds.). *CLEF 2023 Working Notes*. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073., 2023.
7. KURTZKE, John F. On the evaluation of disability in multiple sclerosis. *Neurology*. 1961, vol. 11, no. 8, pp. 686–686.
8. PÖLSTERL, Sebastian. scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn. *Journal of Machine Learning Research*. 2020, vol. 21, no. 212, pp. 1–6.
9. WANG, Zifeng; SUN, Jimeng. Survtrace: Transformers for survival analysis with competing events. In: *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. 2022, pp. 1–9.

10. HARRELL Frank E., Jr; CALIFF, Robert M.; PRYOR, David B.; LEE, Kerry L.; ROSATI, Robert A. Evaluating the Yield of Medical Tests. *JAMA*. 1982, vol. 247, no. 18, pp. 2543–2546. ISSN 0098-7484.

11. BRADLEY, Andrew P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition*. 1997, vol. 30, no. 7, pp. 1145–1159.

12. DEBRAY, Thomas PA et al. A guide to systematic review and meta-analysis of prediction model performance. *bmj*. 2017, vol. 356.

13. KURTZKE, John F. Rating neurologic impairment in multiple sclerosis. *Neurology*. 1983, vol. 33, no. 11, pp. 1444–1444. ISSN 0028-3878. Available from eprint: `https://n.neurology.org/content/33/11/1444.full.pdf`.

14. CHAWLA, Nitesh V.; JAPKOWICZ, Nathalie; KOTCZ, Aleksander. Editorial: Special Issue on Learning from Imbalanced Data Sets. *SIGKDD Explor. Newsl.* 2004, vol. 6, no. 1, pp. 1–6. ISSN 1931-0145.

15. YPERMAN, Jan; POPESCU, Veronica; VAN WIJMEERSCH, Bart; BECKER, Thijs; PEETERS, Liesbet M. Motor evoked potentials for multiple sclerosis, a multiyear follow-up dataset. *Scientific Data*. 2022, vol. 9, no. 1, p. 207.

16. PARCIAK, Tina et al. Introducing a core dataset for real-world data in multiple sclerosis registries and cohorts: Recommendations from a global task force. *Multiple Sclerosis Journal*. 2023, p. 13524585231216004.

17. MUSLIM, Ali M et al. Brain MRI dataset of multiple sclerosis with consensus manual lesion segmentation and patient meta information. *Data in Brief*. 2022, vol. 42, p. 108139.

18. COMMOWICK, Olivier et al. Multiple sclerosis lesions segmentation from multiple experts: The MICCAI 2016 challenge dataset. *NeuroImage*. 2021, vol. 244, p. 118589. ISSN 1053-8119. Available from DOI: `https://doi.org/10.1016/j.neuroimage.2021.118589`.

19. CHARIL, Arnaud et al. Statistical mapping analysis of lesion location and neurological disability in multiple sclerosis: application to 452 patient data sets. *NeuroImage*. 2003, vol. 19, no. 3, pp. 532–544. ISSN 1053-8119. Available from DOI: `https://doi.org/10.1016/S1053-8119(03)00117-4`.

20. WANG, Ping; LI, Yan; REDDY, Chandan K. Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)*. 2019, vol. 51, no. 6, pp. 1–36.

21. CLARK, Taane G; BRADBURN, Michael J; LOVE, Sharon B; ALTMAN, Douglas G. Survival analysis part I: basic concepts and first analyses. *British journal of cancer*. 2003, vol. 89, no. 2, pp. 232–238.

22. LOMBARDI, Angela et al. Time-to-event interpretable machine learning for multiple sclerosis worsening prediction: results from iDPP@ CLEF 2023. In: *CLEF*. 2023.

23. VASWANI, Ashish et al. Attention is all you need. *Advances in neural information processing systems*. 2017, vol. 30.

24. LITTLE, Roderick JA; RUBIN, Donald B. *Statistical analysis with missing data*. Vol. 793. John Wiley & Sons, 2019.

25. CURTIS, Christina et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*. 2012, vol. 486, no. 7403, pp. 346–352.

26. KNAUS, William A et al. The SUPPORT prognostic model: Objective estimates of survival for seriously ill hospitalized adults. *Annals of internal medicine*. 1995, vol. 122, no. 3, pp. 191–203.

27. COX, David R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1972, vol. 34, no. 2, pp. 187–202.

28. ISHWARAN, Hemant; KOGALUR, Udaya B; BLACKSTONE, Eugene H; LAUER, Michael S. Random survival forests. 2008.

29. KATZMAN, Jared L et al. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC medical research methodology*. 2018, vol. 18, pp. 1–12.

30. LEE, Changhee; ZAME, William; YOON, Jinsung; VAN DER SCHAAR, Mihaela. Deephit: A deep learning approach to survival analysis with competing risks. In: *Proceedings of the AAAI conference on artificial intelligence*. 2018, vol. 32. No. 1.

31. KVAMME, Håvard; BORGAN, Ørnulf. Continuous and discrete-time survival prediction with neural networks. *arXiv preprint arXiv:1910.06724*. 2019.

32. NAGPAL, Chirag; LI, Xinyu; DUBRAWSKI, Artur. Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks. *IEEE Journal of Biomedical and Health Informatics*. 2021, vol. 25, no. 8, pp. 3163–3175.

33. HOO, Zhe Hui; CANDLISH, Jane; TEARE, Dawn. What is an ROC curve? *Emergency Medicine Journal*. 2017, vol. 34, no. 6, pp. 357–359. ISSN 1472-0205. Available from DOI: 10.1136/emermed-2017-206735.

34. KOTSIANTIS, Sotiris B; KANELLOPOULOS, Dimitris; PINTELAS, Panagiotis E. Data preprocessing for supervised leaning. *International journal of computer science*. 2006, vol. 1, no. 2, pp. 111–117.

35. HOWARD, Jeremy; GUGGER, Sylvain. Fastai: A Layered API for Deep Learning. *Information*. 2020, vol. 11, no. 2, p. 108.

36. BIEWALD, Lukas. *Experiment Tracking with Weights and Biases*. 2020. Available also from: `https://www.wandb.com/`. Software available from wandb.com.

37. BRANCO, Ruben et al. Investigating the impact of environmental data on ALS prognosis with survival analysis. In: *CLEF*. 2023.

38. ROSSI, Ivan; BIROLO, Giovanni; FARISELLI, Piero. Multiple Sclerosis Survival Prediction Results from DSM-COMPBIO UNITO. In: *CLEF*. 2023.