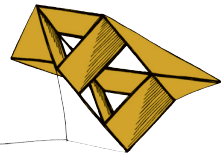

Predicting Risk of Multiple Sclerosis Worsening

Marek Hanzl[1]

1 Introduction

Multiple Sclerosis (MS) is a chronic autoimmune disease of the central nervous system. The progression of MS is highly varied and unpredictable and can lead to a wide range of physical, cognitive, and emotional symptoms. Furthermore, its progression and severity can vary widely between individuals. Therefore, it is difficult to predict the disease progression. The iDPP@CLEF 2023 aims to support this issue by opening an international challenge, to use AI to help predict disease progression and to provide clinicians with valuable information to support their decisions for individual patients' treatment. In this work, we describe our winning submission into the iDPP@CLEF competition.

2 iDPP@CLEF 2023 Competition

The iDPP competition offers 3 tasks. We participated in the first: *Predicting risk of disease worsening*, and the second: *Predicting the probability of worsening at different time windows*. Additionally, each task consists of two sub-tasks, (a) and (b) with different datasets. These consist of static and dynamic data, which are categorical and continuous. The dynamic data span over 2.5 years and include medical history (e.g., relapses, EDSS scores, evoked potentials, MRIs, MS type, etc.) of 1,192 patients. In order to allow robust evaluation of overall performance, several metrics were defined, i.e., C-Index (Harrell et al. (1982)) for Task 1, and AUROC curve for Task 2. Based on the effect on the validation score, we selected the most effective features and pre-processed the datasets.

3 Methodology

The prediction of disease worsening is usually referred to as survival analysis. In our experiments, we have evaluated various approaches, namely: *Gradient Boosting*, *Componentwise Gradient Boosting*, *Random Survival Forest* (Pölsterl (2020)) ensemble models, and a recent *SurvTRACE* transformer model (Wang et al. (2022)).

Training procedure: Each model was trained and validated 100×. In each iteration, we randomly split the pre-processed data into training and validation sets in an 80/20 ratio, and we measured the C-Index. We chose the best model based on the achieved average and std of the C-Index calculated from all the iterations. Furthermore, we performed a hyper-parameters search for all methods using the *Weights & Biases "Sweeping tool"* based on the performance on a dataset of sub-task (a).

We selected the models with the best overall validation score and used them to predict the outcomes of the test data for each sub-task separately. At last, among those are ensemble models that combine the best of every model type, and average their predictions.

[1] student in a Bachelor's program Cybernetics and Control Engineering UWB, e-mail: hanzl@students.zcu.cz

Risk of Worsening

Rank	C-Index
1.	**0.834**
2.	0.802
3.	0.801
4.	0.774
5.	0.771

Probability of Worsening [AUROC]

Rank	2 years	4 years	6 years	8 years	10 years
1.	**0.924**	**0.907**	**0.896**	**0.838**	**0.839**
2.	0.864	0.898	0.938	0.859	0.831
3.	0.890	0.900	0.856	0.787	0.796
4.	0.754	0.873	0.871	0.746	0.745
5.	0.774	0.740	0.774	0.703	0.722

Rank	C-Index
1.	0.690
2.	0.634
3.	**0.601**
4.	0.598
5.	0.587

Rank	2 years	4 years	6 years	8 years	10 years
1.	0.632	0.626	0.655	0.673	0.709
2.	**0.606**	**0.612**	**0.602**	**0.587**	**0.626**
3.	0.642	0.567	0.601	0.594	0.622
4.	0.614	0.639	0.629	0.616	0.527
5.	0.644	0.590	0.610	0.567	0.609

Table 1: Official competition results – best runs for top 5 teams (Tasks (a) & Tasks (b)). Our results are in bold. AUROC scores are calculated for each time interval separately.

4 Results and Conclusion

The official results of the top 5 teams in each category are visualized in Table 1. Our approach provides a reliable risk of worsening prediction with a C-Index value of 0.834 and scored first in 2 out of four tasks. Even though we optimized our method just for tasks (a) , we achieved competitive results in tasks (b) , scoring 2nd, and 3rd, place out of 10 teams. This outcome is likely due to the different way the EDSS features were defined. Achieved results revealed true potential in providing clinicians with additional information and allowing treatment interventions.

Acknowledgement

Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth, and Sports of the Czech Republic. The work has been supported by the grant of the University of West Bohemia, project No. SGS-2022-017.

References

Harrell, Frank E., Jr and Califf, Robert M. and Pryor, David B. and Lee, Kerry L. and Rosati, Robert A.(1982) Evaluating the Yield of Medical Tests. *The Journal of the American Medical Association*, Volume 247, pp. 2543–2546.

Pölsterl S. (2020) scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn. *Journal of Machine Learning Research*, Volume 21, pp. 1–6.

Wang, Zifeng and Sun, Jimeng (2022), Survtrace: Transformers for survival analysis with competing events. *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. pp. 1–9, 2022.