Improving Features for Multiple Sclerosis Disability Progression Prediction through Temporal Alignment of Hospital Visits

Karel Fonteyn, Tom Dhaene and Dirk Deschrijver
IDLab
Ghent University - imec
Ghent, Belgium
{karel.fonteyn, tom.dhaene, dirk.deschrijver}@ugent.be

2024 International Conference on Machine Learning and Applications (ICMLA) | 979-8-3503-7488-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICMLA61862.2024.00171

***Abstract*—Predicting the disability progression in multiple sclerosis remains a significant challenge. Statistical features extracted from evoked potential signals are often discussed in literature as interesting biomarkers, yet their practical adoption remains limited due to their low effectiveness. Conversely, the use of deep neural network feature encoders, which allow for extracting more expressive biomarkers, is hindered by the limited availability of data. This study proposes a novel method of enhancing statistically extracted features by aligning hospital visits to embed the progression of the disease within the features. The results demonstrate a significant improvement in predicting disability progression using the statistically extracted features. Insight is provided through a study of correlation and importance of the enhanced features.**

***Index Terms*—Feature extraction, Representation learning, Predictive modeling, Multiple sclerosis**

I. Introduction

Deep neural network (DNN) feature encoders capture complex relationships within raw data signals by directly optimizing features to enhance task performance. However, complex DNNs require large volumes of data and substantial computational resources. For data sets with a limited number of samples, DNNs are prone to overfitting, compromising generalization. In such cases, computing statistical features that capture information like the morphology, statistical properties, and dynamic behavior from a signal is often preferable. Additionally, while DNN-learned features can discern intricate patterns, statistical features are more easily interpreted.

Healthcare data sets are rarely made publicly available due to the high value associated with expert annotations and privacy considerations. This results in a limited number of public data sets, often with only a small number of patients. Consequently, research and development of DNN feature encoders are restricted for such data, making the use of statistical feature extractors prevalent. This is also evident in research on monitoring the progression of Multiple Sclerosis (MS).

Various biomarkers have been documented in literature to monitor the individual progression of MS in patients. These include, but are not limited to, features derived from magnetic resonance imaging (MRI) scans [1] and features from evoked potentials (EPs) time series [2]. Despite extensive research, the practical adoption of these biomarkers remains limited due to their low effectiveness [3]. This study investigates whether statistically extracted features, more specifically EP biomarkers, can be optimized to enhance their utility in tracking disability progression. To this end, the potential of the Structuring Whitened Embeddings (SWE) approach [4] for statistical feature enhancement is explored.

This research received funding from the Flemish Government via the AI Research Program.

II. Background

A. EP biomarkers for MS progression

MS is a chronic autoimmune disease of the central nervous system, affecting millions of people worldwide. With no curative therapies available, treatment aims to prevent episodic inflammation and disability, supported by the understanding and monitoring of the disability progression in patients [5]. Patient disability can be monitored using EPs, measuring conduction of nerve pathways. By stimulating specific nerves and recording activity elsewhere, EPs reveal lesions via decreased conduction. While various types of EPs have been studied as MS biomarkers, this study focuses on motor EPs (MEP) [6]. MEPs relevant as biomarkers for MS are recorded in the abductor pollicis brevis (APB) muscles when stimulated in the hand areas of the motor cortex, and in the abductor hallucis (AH) muscles when stimulated in the leg areas of the motor cortex (Fig. 1) [7].

B. Structuring Whitened Embeddings

This study explores the use of Structuring Whitened Embeddings (SWE) [4] to enhance statistically extracted features. The SWE approach constructs inter-sample relations by aligning samples along an auxiliary relational variable. For MS disability progression, visits of a patient can be aligned using the time elapsed since the patient's first visit. This alignment enables capturing the evolution of the patient over time, encapsulating valuable information about the disability progression.

1946-0759/24/$31.00 ©2024 IEEE
DOI 10.1109/ICMLA61862.2024.00171

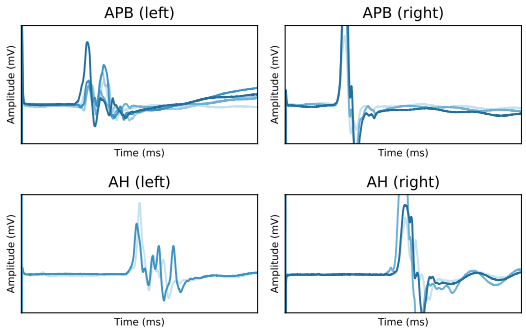


Fig. 1: Multiple four-signal MEP recordings are available per patient as MS biomarkers. Two channels are recorded in the abductor pollicis brevis (APB) muscles, and two in the abductor hallucis (AH) muscles.

The employed SWE approach was previously introduced to fit a DNN feature encoder for extracting features from raw ECG time-series in a self-supervised manner. Proportions among the relations set up along the auxiliary variable are mimicked in a whitened embedding, thereby constraining features to evolve smoothly between adjacent samples, effectively capturing continuity along the relational variable [4].

III. Methodology

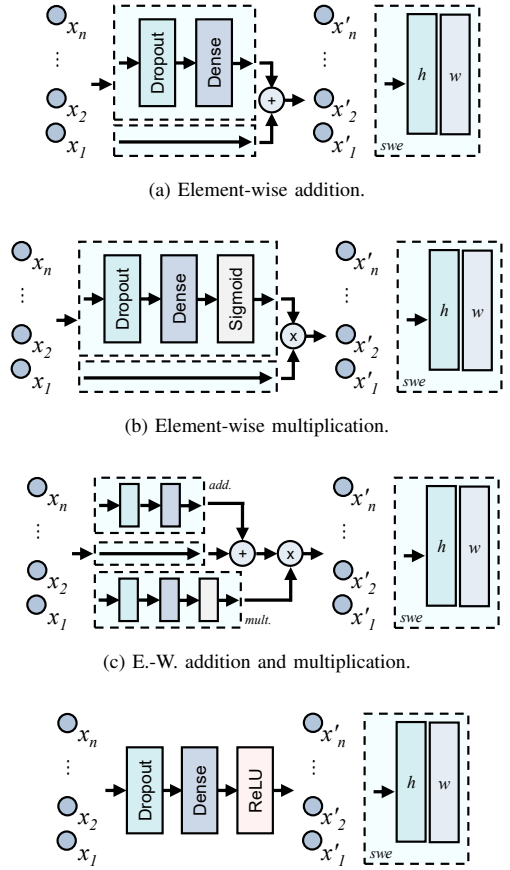
Four sets of enhanced features are obtained by transforming features statistically derived from the MEP [8]. Weights are fitted using SWE by structuring the inputs in a whitened 6-dimensional projection. Dropout is utilized for regularization. A graphical overview of the architectures for the discussed approaches is given in Fig. 2.

(a) Element-wise addition. The first approach explores feature improvement through linear transformation. Learned values, implemented using a Dense layer with neuron count matching the number of input features, are added element-wise to the original features.

(b) Element-wise multiplication. The second approach applies element-wise multiplication between the input features and the Dense layer's neurons. A Sigmoid activation reduces the weight of less important features.

(c) Element-wise addition followed by element-wise multiplication. The two preceding methods are combined for a "best-of-both-worlds" solution. Scaling renders the transformation non-linear.

(d) Dense layer features. Given the limited amount of samples, the raw MEP signals are too complex to fit a DNN feature extractor. However, by employing a single fully connected Dense layer with ReLU-activation, statistical features can be used as input for computing features optimized for progression. While previous approaches transform individual features, this technique aggregates information into a new set of features, essentially mapping patterns into a new space.



(a) Element-wise addition.

(b) Element-wise multiplication.

(c) E.-W. addition and multiplication.

(d) Dense layer features.

Fig. 2: Overview of the architectures used for the evaluated approaches that transform the original statistical features ($\mathbf{x}$) into improved features ($\mathbf{x}'$). The SWE module used for training consists of a projection (h) and a whitening (w) layer.

IV. Experimental Setting

After the feature transformation networks are fitted through SWE, the quality of the feature sets is assessed by predicting MS disability progression after two years, quantified as a binary problem. This study uses the standard definition of disability progression as defined by Kalincik et al. [9].

The data set "Motor Evoked Potentials for Multiple Sclerosis: A Multiyear Follow-up Dataset" [10] is utilized. From this data, 380 patients undergoing treatment are considered, totaling 1969 individual records. 85 individuals have at least one record indicating MS progression. Each record consists of 4 MEP signals. Statistical features from each signal are extracted using HCTSA [8], [11], of which the top-ranked features relevant to the MS disability progression as reported by Yperman et al. are considered [7]. Additionally, latency and the peak-to-peak amplitude are considered for every signal, as well as the time elapsed since the patient's first visit, resulting in a total of 17 features. The inclusion of this last feature

allows for fair comparison between a model's baseline and the SWE improved feature sets, as all now share the same information. Clinical meta features are excluded to analyze the isolated impact on the statistically extracted features.

Feature optimization by SWE aligns recurring patients' visits, using the time since a patient's first visit as the auxiliary relational variable to establish inter-sample relations. At least three patient visits are required to build up relational proportions, limiting the number of patients available for feature optimization to 265. For each patient, any combination of three chronological visits is considered as an input.

Each set of improved features is evaluated both as a replacement for, and in addition to, the original feature set. The impact is assessed on the performance of predicting disability progression, using Logistic Regression as a linear model and Random Forest as a nonlinear one. The performances obtained using the enhanced feature sets are benchmarked against the selected model's performance on the original feature set, referred to as the baseline.

V. Results & Discussion

Results for the average Area Under the Curve (AUC) of the Receiver-Operating Characteristic (ROC) with 95% confidence intervals are given in Fig. 3. Scores are obtained through 20 repetitions of 5-fold cross validations, totaling 100 measurements, with hyperparameter optimization per fold. Normal distribution of the results is determined by the Shapiro-Wilk test. Statistical significance ($p < 0.05$) relative to the baseline is calculated using the paired T-Test if the data is normally distributed, the Wilcoxon signed-rank test is used otherwise.

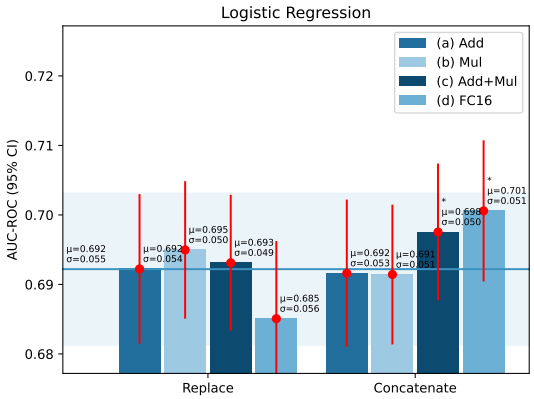
A. Logistic Regression

Fig. 3a presents the results for the Logistic Regression model around the baseline score of ($\mu = 0.692$, $\sigma = 0.055$). Replacement of the original features does not show any significant difference. Concatenating feature sets shows significant improvement for approaches (c) and (d). Concatenating the non-linearly transformed features to the original feature set allows the model to benefit from both linear relationships in the original set and from information added after non-linear mapping. When replacing the original features, performance is reduced, as the linear model struggles to fully leverage the introduced non-linear relationships.

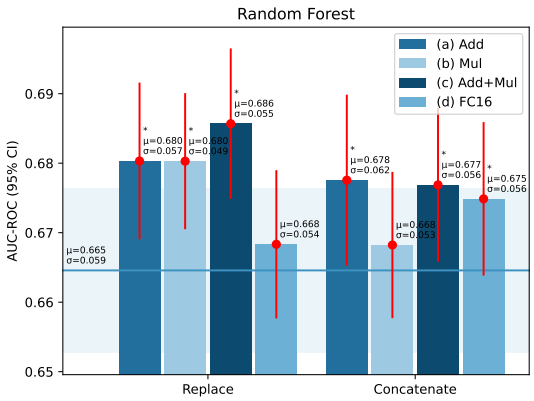
B. Random Forest

Results for the non-linear Random Forest model are plotted in Fig. 3b. While baseline performance ($\mu = 0.665$, $\sigma = 0.059$) is not on-par with the Logistic Regression, the Random Forest can be used to study feature importance.

For replacement of the original feature set, significant improvement in means is obtained by approaches (a), (b), and their combination (c). The improvement after linear transformation (a) indicates that the base features can be made more informative for the model to split on. The element-wise multiplication (b) demonstrates that the Random Forest benefits from added regularization following from feature



(a) Results for the Logistic Regression model.



(b) Results for the Random Forest model.

Fig. 3: Experimental AUC-ROC scores (with 95% confidence intervals) of the improved feature sets, both for replacement of and concatenation with the original feature set. Scores are aligned around the model's baseline performance on the original feature set (horizontal line indicating mean score, shaded box indicating 95% confidence interval) to highlight model-specific performance gains. Scaling is consistent across the plots. Statistical significance is denoted by an asterisk $(*)$.

weighing. Given significant improvement for both methods, it is unsurprising that the best performance for the Random Forest model is achieved using the combination of element-wise addition and multiplication (c). Although significance can also be noted when combining feature sets, the performance of replacing the original set is not matched. The performance drop is likely due to the information being shared over both sets of features, as Random Forests can be sensitive to redundant features.

One might be surprised to see the performance disparity between approaches (c) and (d), given their shared operational basis. The fully connected approach uniformly transforms all inputs through shared weights, aggregating information into a new feature set. Conversely, the element-wise addition and multiplication approach tailors transformations to each feature

individually. These individual adjustments are particularly beneficial for random forest models, which rely heavily on the quality and relevance of individual input features.

C. Feature Importance

Analyzing the feature importance as set by the Random Forest reveals that the baseline model, fitted to the original features, assigns the most value to the latency, as well as to the peak-to-peak amplitudes (up to 0.080). This aligns with literature, where both are recognized as important MS biomarkers. All other statistical features fall in the range of 0.047 till 0.057. Upon replacement by SWE-transformed features, the importance distribution becomes more uniform. All statistical features, including latency and peak-to-peak amplitudes, now exhibit feature importance values between 0.055 and 0.061.

The importance assigned to the features of the baseline is heavily skewed towards a select few, leading to an over-reliance that compromises generalization. The use of SWE for feature transformation enables the model to recognize a wider range of features as useful, introducing improved robustness.

D. Feature Correlation

Excluding self-correlations, the original feature set has correlation values ranging from 0.04 to 0.77, with correlations to the time since the first visit ranking the lowest. After transformation by approaches (a) and (c), feature correlations become more averaged, narrowing the range to $[0.25, 0.58]$ and $[0.23, 0.52]$ respectively. Notably, the highest correlations are all with the 'time since first visit'-feature, indicating successful embedding of temporal evolution, as well as decorrelation of the other features. Similarly, for the Dense layer features of (d), all correlations between learned features are minimal (in the range of $[0.27, 0.37]$), while correlations with the feature 'time since first visit' rank highly ($[0.48, 0.58]$).

The original statistical features used in the baseline exhibit high correlation, leading to multicollinearity that destabilizes the Random Forest model. Following SWE-fitted transformations, feature correlation decreases. While this balance mitigates issues related to multicollinearity for the Random Forest, this likely also harms the predictive power of the Logistic Regression model. Nonetheless, with the temporal evolution being embedded, the features become more suited for the prediction setting of disability progression.

VI. Conclusion

This research explored various statistical feature enhancement methods using SWE, each with distinct strengths and potential applications, and with suitability depending on the objective and the machine learning model used. This study assessed improvement in predicting MS disability progression. Feature sets were evaluated using Logistic Regression as a linear model, and Random Forest as a non-linear one.

It was observed that mapping statistically extracted features to a new space through a Dense layer using SWE can capture complex data patterns without the overfitting issues typically associated with DNN-feature extraction. Including these transformed features in the Logistic Regression model significantly improved its baseline performance. Similar improvement was observed in the performance of the Random Forest model after individual feature transformations.

An analysis of feature importance and correlation provided valuable insights. The improved performance of the Random Forest model can be attributed to better generalization with SWE-improved features, for which importances are more uniformly assigned. The transformed features also exhibit a balanced correlation structure, reducing multicollinearity issues for this model. Conversely, for the linear Logistic Regression model, the lower correlation of the individually transformed features leads to less predictive power. While the correlation between different statistical features was reduced, the correlation with time, the relational variable employed for SWE, was enhanced, signifying an optimization of the features for predictive settings.

This research provided valuable insights into enhancing biomarkers for predicting disability progression in MS. Further studies can extend the application to other data, potentially exploring its performance on more extensive feature sets.

References

[1] A. Gajofatto, M. Calabrese, M. D. Benedetti, and S. Monaco, "Clinical, mri, and csf markers of disability progression in multiple sclerosis," *Disease markers*, vol. 35, no. 6, pp. 687–699, 2013.

[2] B. Kallmann, S. Fackelmann, K. Toyka, P. Rieckmann, and K. Reiners, "Early abnormalities of evoked potentials and future disability in patients with multiple sclerosis," *Multiple Sclerosis Journal*, vol. 12, no. 1, pp. 58–65, 2006.

[3] E. De Brouwer, T. Becker, L. Werthen-Brabants, P. Dewulf, D. Iliadis, C. Dekeyser, G. Laureys, B. Van Wijmeersch, V. Popescu, T. Dhaene *et al.*, "Machine-learning-based prediction of disability progression in multiple sclerosis: an observational, international, multi-center study," *PLOS Digital Health*, vol. 3, no. 7, p. e0000533, 2024.

[4] K. Fonteyn, L. Bontinck, T. Dhaene, and D. Deschrijver, "Structuring whitened embeddings: Augmentation-free representation learning for sparsely labeled transformation-sensitive data," 2024, submitted to Knowledge-Based Systems, Elsevier.

[5] T. Kalincik, A. Manouchehrinia, L. Sobisek, V. Jokubaitis, T. Spelman, D. Horakova, E. Havrdova, M. Trojano, G. Izquierdo, A. Lugaresi *et al.*, "Towards personalized therapy for multiple sclerosis: prediction of individual treatment response," *Brain*, vol. 140, no. 9, pp. 2426–2443, 2017.

[6] P. Fuhr, A. Borggrefe-Chappuis, C. Schindler, and L. Kappos, "Visual and motor evoked potentials in the course of multiple sclerosis," *Brain*, vol. 124, no. 11, pp. 2162–2168, 2001.

[7] J. Yperman, T. Becker, D. Valkenborg, V. Popescu, N. Hellings, B. V. Wijmeersch, and L. M. Peeters, "Machine learning analysis of motor evoked potential time series to predict disability progression in multiple sclerosis," *BMC neurology*, vol. 20, pp. 1–15, 2020.

[8] B. D. Fulcher and N. S. Jones, "hctsa: A computational framework for automated time-series phenotyping using massive feature extraction," *Cell systems*, vol. 5, no. 5, pp. 527–531, 2017.

[9] T. Kalincik, G. Cutter, T. Spelman, V. Jokubaitis, E. Havrdova, D. Horakova, M. Trojano, G. Izquierdo, M. Girard, P. Duquette *et al.*, "Defining reliable disability outcomes in multiple sclerosis," *Brain*, vol. 138, no. 11, pp. 3287–3298, 2015.

[10] J. Yperman, V. Popescu, B. Van Wijmeersch, T. Becker, and L. M. Peeters, "Motor evoked potentials for multiple sclerosis, a multiyear follow-up dataset," *Scientific Data*, vol. 9, no. 1, p. 207, 2022.

[11] B. D. Fulcher, M. A. Little, and N. S. Jones, "Highly comparative time-series analysis: the empirical structure of time series and their methods," *Journal of the Royal Society Interface*, vol. 10, no. 83, p. 20130048, 2013.