2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA) | 978-1-7281-8470-8/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICMLA51294.2020.00090

Domain adaptation from clinical trials data to the tertiary care clinic – Application to ALS

Ben Hadad
Department of Industrial Engineering and Management
Ben-Gurion University of the Negev
Beer Sheva, Israel
benhada@post.bgu.ac.il

Boaz Lerner
Department of Industrial Engineering and Management
Ben-Gurion University of the Negev
Beer Sheva, Israel
boaz@bgu.ac.il

***Abstract*—Amyotrophic lateral sclerosis (ALS) is a devastating and incurable disease affecting motor neurons, leading to progressive paralysis and death on average within three to five years from onset. The disease is characterized by highly variable patterns and rates of progression, which pose challenges to developing reliable and accurate ALS disease state prediction models to be used on a daily basis in clinics with little data. To meet these challenges, we suggest domain adaptation from a large, but unfortunately biased, clinical trials database to that of a tertiary care ALS clinic. To evaluate the reliability and accuracy of the suggested paradigm, we examine a naïve approach by which training is based only on the clinical trials data compared with a domain adaptation approach of an initial training using this same data followed by fine-tuning training using the clinic data. We also allow summarization of the clinical longitudinal data to evaluate non-temporal models, e.g., random forest (RF), XGBoost (XGB), and multilayer perceptron (MLP), partially exploiting the dynamic information hidden in patient clinical records, in comparison to the long short-term memory (LSTM) recurrent neural network, fully exploiting the temporal information in the data. First, we notice the XGB outperformance in terms of the ALS disease state prediction error to the RF and MLP, but surprisingly also to the LSTM regardless of prediction time (up to 24 months ahead). We contribute the inferiority of the highly parametrized neural network to the impact of the curse of dimensionality. Second, we show that this error does not significantly increase when the model is trained using only the clinical trials data, especially for LSTM in long prediction times. Finally, we demonstrate that fine-tuning of the clinical trials-based pre-trained model using the clinic data improves the LSTM and MLP performance compared to using solely the clinical trials or clinic data.**

Keywords—Amyotrophic lateral sclerosis (ALS), clinical trials data, disease-state prediction, LSTM, domain adaptation

I. Introduction

Amyotrophic lateral sclerosis (ALS) is a neurodegenerative disease of the motor neurons, with a highly uncertain pathogenesis, leading to progressive paralysis and death [1]. Despite medical and clinical progress since its discovery, this paralysis is still not visibly affected by the different therapies currently available [2, 3, 4]. ALS is characterized by a progressive decline of function of the upper and lower motor neurons, leading to progressive paralysis that affects the muscles of limbs, speech, swallowing, and respiration. The rapid progression of the disease is difficult to handle and known to be terminal an average of five years within onset. The inner workings and mechanisms of this disease remain unknown [5]. Since the 1990s, however, there has been a rapidly growing interest in the disease among the scientific and medical communities. It has been understood that extending the life expectancy and improving the quality of life of those afflicted depends on our understanding of its pathogenesis [1, 2].

The disease is characterized by high heterogeneity among patients regarding its progression, which makes it difficult to achieve significant results in clinical trials for developing medications and in treatments. Thus, reliable models for disease state prediction could improve the ability to assess treatment influence in the clinic and clinical trials and reduce the number of patients necessary to achieve statistically significant results [6, 7]. In addition, an accurate way to anticipate disease progression might benefit patients and their families by better preparing them for the anticipated disease outcomes. Besides the heterogeneity of the disease, another challenge in developing an accurate prediction model for clinical use is the small amount of data documented in the clinic because of the rarity of the disease. Clinical trial databases are usually large, as they comprise several trials.

Because of the importance of developing ALS disease state prediction models, the DREAM-Phil Bowen ALS Prediction Prize4Life Challenge was launched in 2012, inviting participants to develop algorithms to predict the rate of ALS disease progression for individuals as measured by the ALS functional rating scale (ALSFRS). Following the success of the first challenge, the DREAM ALS Stratification Prize4Life Challenge was launched in 2015 and focused on patient clustering as well as progression rate forecasting. Both challenges were based on the pooled resource open-access ALS clinical trials PRO-ACT database [8] and asked solvers to use three months of clinical trials information to predict disease future progression in months 3–12. The progression of the disease was assessed by the slope of change in ALSFRS values, which assumes linear disease progression.

However, this setting is not necessarily applicable in the clinic. Patients who enroll in clinical trials have more longitudinal data than casual clinic patients, and their demographic and clinical characteristics are different regarding, e.g., age and progression. While some studies [9] suggest to use only baseline information (data from onset and the first clinical visit) to develop more clinically applicable models, in practice, the clinic patient population is much more heterogenic in terms of number and interval of clinic visits. On the one hand, requiring patients to have three months of successive data results in an unpractical model for clinical use. On the other hand, using only baseline information for prediction does not exploit all necessary information for most patients. Therefore, some changes in the methodology are expected, especially if it is part of a more general approach that should also be implemented on other neurodegenerative diseases.

978-1-7281-8470-8/20/$31.00 ©2020 IEEE
DOI 10.1109/ICMLA51294.2020.00090

We suggest a methodology for model training and evaluation that does not require three months of data and utilizes all the patient information we possess to make future ALS disease state predictions. The suggested methodology simulates a more realistic and practical clinical scenario in which a patient who visits the clinic might have two or more past clinical visits in varying time periods, and the doctor is interested to predict their future disease state. Using the suggested methodology, we compared predictions of four state-of-the-art machine-learning algorithms on two databases: the PRO-ACT database (clinical trials data) and the TAS Medical Center (TASMC) database (clinic data). Three of the algorithms are non-temporal: (1) Random forest (RF) [10], the most common method for ALS predictions; (2) XGBoost (XGB), an open-source software library which provides a gradient boosting framework [11]; and (3) the feed forward multilayer perceptron (MLP) neural network, whose input structure is identical to the ensemble tree algorithms (1 and 2), but also is suitable for further incremental learning and domain adaptation. The fourth is the long short-term memory (LSTM) [12], which is more appropriate to sequential and temporal data as opposed to models such as RF and XGB. The latter two models aggregate data and may therefore discard important information, or only partly exploit it [13].

In addition, the commonly used PRO-ACT database includes information from ALS patients who participated in industry clinical trials, but this population does not reflect the clinical patient population due to some requirements that needed to be met in order to participate in that trial. For example, the clinical trial patients, compared to those in our clinic database, were younger, higher functioning (expressed by low disease state deterioration), and more homogeneous in terms of clinic visit frequency. Although relevance to the clinic of models trained on the PRO-ACT have been tested [14], we asked whether using this biased database as the basis of the trained model is justified, and if so, whether it could improve clinic patient predictions. We evaluated the clinic patients' disease state prediction accuracy with two approaches using the PRO-ACT patient data and compared this to a model that was trained only on the clinic patients. The first approach is a naïve one, relying on training a model on the PRO-ACT patients and evaluating its predictions on the clinic patients. The second, the domain adaptation approach, uses the PRO-ACT trained model as the initial model, and the training is continued only on the clinic (TASMC) patients for fine-tuning to the clinic. This approach is suitable for ALS clinics that based on the openly available large PRO-ACT database can utilize their small databases (since ALS is a rare disease) to achieve better predictions on their own patients. Note also that this approach is not restricted to ALS and can be applied to other (neurodegenerative) diseases.

II. Background and Related Work

The most popular rating instrument for monitoring the progression of disability in ALS patients is the ALSFRS. ALSFRS scores range from 0=no functionality to 4=full functionality. Each of ten ALSFRS items describe different physical functionality, e.g., breathing, speaking, walking, etc. that sums to the total ALSFRS score (0–40) [15]. In 1999, the revised ALSFRS (ALSFRS-R) was designed in order to balance the weighing between the limb and bulbar as compared to the respiratory function by incorporating additional assessments of dyspnea and orthopnea [16]. Although today the ALSFRS-R is a more popular target variable, in this study, we used the ALSFRS due to the amount of missing data that the TASMC database contained in the additional two respiratory ALSFRS items.

Using the PRO-ACT database, [17] used a non-linear Weibull model to describe ALS disease progression. The parameters of the Weibull model were estimated using a non-linear mixed-effect modeling approach: Patients are first assigned to one of two clusters based on their deterioration rate—slow progression or fast progression, and then the correspondent Weibull function is applied. The patient's deterioration rate is calculated based on the difference between the baseline value of the ALSFRS-R score and the last value. In [18], different algorithms. e.g., a pre-slope model, a generalized linear model (GLM), and an RF algorithm used for prediction disease progression based on the PRO-ACT database and suggested that past disease progression is a strong predictor of future disease progression. They also found that larger variability in initial ALSFRS scores is linked to faster future disease progression. Another conclusion reported was that an RF model using only baseline data could accurately predict disease progression for a clinical trial research dataset, as well as for a population being treated at a tertiary care clinic. The RF outperformed the pre-slope and GLM models mainly at farther time points, while at early time points, the GLM and the RF were quite similar. The most important features found were the time from baseline (prediction time), the ALSFRS-R score at baseline, and the ALSFRS slope. Another work used three months of patient data to predict the changes in ALSFRS scores over time [19]. They applied model-based (linear models) and model-free [RF and Bayesian adaptive regression trees (BART)] methods. The BART was slightly better than the RF, but the authors reported that both were only moderately successful. In another study, [14] developed models based on the PRO-ACT for classifying patients into two classes: slow and fast progressors. Others [13] applied the LSTM to the PRO-ACT and used the patients' last visit data to predict further visits.

III. Methodology

A. Data

Two databases we transferred between are: the clinical trials database—the PRO-ACT, and a tertiary care ALS clinic database—the TASMC. In this section, we will compare between patient characteristics in the two databases to anticipate the influence of these characteristics on the success of the transfer.

The PRO-ACT database. The PRO-ACT database was created by Prize4Life and the Neurological Clinical Research Institute (NCRI) at Massachusetts General Hospital in order to enhance ALS research by building a data set that would merge data from a large number of completed ALS clinical trials [8]. We used data of ALSFRS, demographics, family history, laboratory data, vital signs, and forced vital capacity (FVC).

The TASMC database. The data were collected in the ALS clinic of this medical center during the years 2000–2019. The clinic is a large tertiary referral center for ALS that today follows approximately 100 new cases annually. The database contains records of patients with clinically probable or definite ALS who were followed in the clinic. For all patients, age at disease onset, age at diagnosis, gender, ethnicity, and disease form at onset, as well as ALSFRS in each visit were recorded.

Not only does the PRO-ACT contain records of many more patients (3,171 patients in the PRO-ACT and 1,328 in TASMC), it also contains more clinical visits per patient than the TASMC (6.97 vs. 4.39 on average, respectively). A patient in the PRO-ACT database had more visits with shorter intervals between each consecutive visits. Recall that the PRO-ACT merges data of clinical trials which go on for a limited time period, where patients are asked to visit the clinic at fixed time intervals. Additionally, patients who visited the clinic frequently for a long period of time were considered stable, and using the last visits of those patients might cause an over-fitting problem to our models. Thus, for fair comparison between the databases, from the TASMC database, we used only visits that occurred in the first two years from the first visit. Missing data that could be imputed using the clinical staff were completed.

Patients' age and time since onset (the estimated time between disease onset and the first clinical visit) were quite similar in both databases, as were as gender and onset site distribution.

B. Models and data preparation

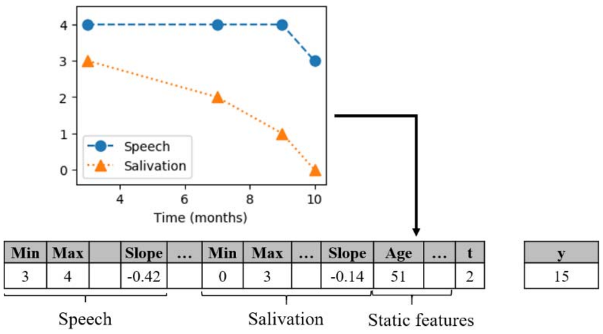
We used four state-of-the-art models suitable for regression problems: RF, XGB, MLP, and the LSTM artificial neural network. Although RF is the most popular method for ALS prediction problems, we also used the XGB, one of the most popular gradient boosting algorithms that is found to be a powerful tool in many domains. Both the RF and the XGB are non-temporal models that require "flattening" of the temporal data by some aggregation. For each temporal feature, we extracted main temporal characteristics, e.g., mean, standard deviation, slope, minimum, maximum, etc. (Fig. 1).

In contrast, one of the advantages of the LSTM model is that it can process entire sequences (longitudinal clinical data), i.e., the full information in the patient's data is extracted and used for model training. In addition, the neural network architecture of the LSTM can be used for the domain adaptation problem by fitting a layer's weights for a certain task (i.e., the source) and then using these weights as the initial ones for further training on a different but similar task (i.e., the target). This approach can be very beneficial when the target domain is, e.g., a small data set of ALS patients from the clinic (say TASMC), while the source domain is a large data set, e.g., of clinical trials data (say PRO-ACT). There is no straightforward way to implement domain adaptation with the XGB and the RF, but an MLP model, whose input is also flattened data, has the ability to adapt to new domains similarly to the LSTM.

C. Experimental design

In our suggested methodology, inspired by [13], each patient does not contribute only one training or test observation. Any clinical visit of the patient is a candidate to be the target visit whose value we want to predict or an explanatory visit to train the model with.

Note that for non-temporal models, in order to flatten the temporal data (for example, to calculate the slope), values of two visits are required. Hence, this is the only restriction in order to consider a split of patient visits to explanatory visits and a target visit. Each training or test patient provides several valid observations. As a result, the training and test sets contain as many observations as possible with different prediction times, which contribute to better generalizing and evaluating the model for any prediction time.



Min	Max		Slope	…	Min	Max	…	Slope	Age	…	t
3	4		-0.42		0	3		-0.14	51		2

y
15

Fig. 1. Example of a feature vector for a 51-year old patient. In this example, the patient had four explanatory visits, and the target visit was two months after the last explanatory one (with a true value of 15). The deterioration pattern of the patient in speech and salivation is shown in the graph. For each temporal feature, the minimum, maximum, slope (deterioration rate between the first and last visit), etc. are calculated and concatenated with the static features and prediction time.

In the first experiment we performed, the goal was to compare between the prediction performances of the four algorithms. In the second experiment, we tested whether a clinical trials-based model could be used to predict future outcomes of clinic patients. In the third experiment, we used domain adaptation to test whether initial weights of a PRO-ACT-trained LSTM and MLP models could be fine-tuned using the clinic's data to improve the ALS clinic's patient disease state predictions. Note that, in the second and third experiments, we used only shared features of the PRO-ACT and TASMC databases.

Experiment 1—Model comparison. For each of the four algorithms, we searched randomly 60 sets of hyper-parameters using 60 permutations of training–validation–test of the PRO-ACT data, and averaged performance over the 60 test sets for the best configuration.

Experiment 2—Clinical trials vs. clinic data-based prediction models. Using the same permutations and best configurations, we evaluated performance on the TASMC test sets of a model trained using the PRO-ACT data compared with another trained using the TASMC training set.

Experiment 3—Domain adaptation from clinical trials to the tertiary-care clinic. As illustrated in Fig. 2, the experiment involves two training phases: (1) fitting initial MLP and LSTM models on the PRO-ACT database, and (2) fine-tuning the trained models using the clinic (only training) patients' data. Like in Experiment 2, performance is evaluated only with the clinic test patients, averaged over 60 permutations of the training–test patient split.

D. Evaluation

The mean absolute error (MAE) and the RMSE are both suitable metrics to evaluate the ALS predictions because both metrics express the average model prediction error in the units of the variable of interest, which in our case, is the ALSFRS score of the target visit. Although the RMSE gives relatively high weight to large errors compared to MAE, in this study, and due to their importance, we provided both RMSE and MAE in each experiment. While a test set contains observations with different prediction times, most are short-term predictions (as the disease average duration is between 3 and 5 years), the RMSE and MAE calculated over the whole test set might not be informative enough for evaluating the prediction accuracy for long-term predictions. Hence, we also provide prediction evaluations for four time-intervals: from 0 to 6 months, 6 to 12 months, 12 to 18 months, and 18 to 24 months. For each prediction time interval, we estimate the RMSE and MSE using bootstrapping analysis by computing for each prediction time interval the RMSE and the MAE for 1,000 test subsets, each containing 800 test observations randomly sampled without replacement.

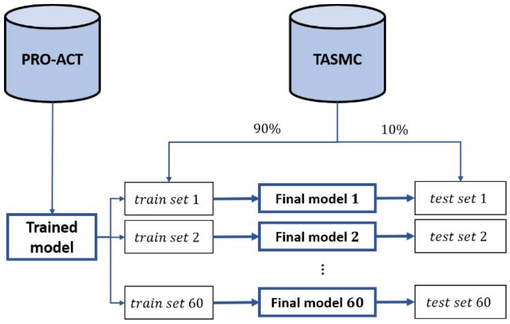


Fig. 2. The first stage in Experiment 3 is fitting LSTM and MLP models using the PRO-ACT database. In the second stage, these models are fine-tuned using the TASMC training set. Then prediction performance using the TASMC test set is reported as the average over 60 (90/10% TASMC training/test sets) random patient splits.

IV. Results

A. Experiment 1—Model comparison

Feature importance. The average relative feature importance by the RF, which is calculated as the decrease in the RF node impurity weighted by the probability of reaching that node, over 60 models (a model for each data permutation) demonstrated a similar trend between the PRO-ACT and TASMC databases. In both cases, the two most important features to predict the next ALSFRS score are the last known ALSFRS score ("ALSFRS_last") and the prediction time ("t_pred"). In addition, in both databases, the ALSFRS items are more important in general than the other temporal features the database contains (e.g., laboratory tests in the PRO-ACT database). Hence, we will use only the ALSFRS items in the next two experiments.

Prediction evaluation. Table I shows the estimated RMSE and MAE in four prediction time intervals (and in their union) for the PRO-ACT database. It can be seen that the XGB model outperforms all the other models, demonstrating the lowest RMSE and MAE values regardless of the prediction time (except for the RMSE in 18–24 months). As expected, regardless of the algorithm, the farther away the target visit is, the more challenging the prediction task is, and as a result, the prediction error is greater.

Note that the RF performances are very similar to those of the XGB algorithm, whereas those of the LSTM are inferior to both. In addition, the MLP achieved mediocre performances, better than the LSMT but inferior to the two ensemble methods, indicating that, for this type of longitudinal data where the sequences are relatively short, flattening the data is the most beneficial method to use.

B. Experiment 2—Clinical trials vs. clinic data-based prediction models

Fig. 3 shows increasing prediction errors as a function of the prediction time (regardless of the algorithm and the training set: PRO-ACT or TASMC) and the almost always superiority of the XGB over the RF, MLP, and LSTM. It can be seen that the predictions for early-time intervals (up to 6 months) are more accurate using the PRO-ACT trained model.

The salient improvement belongs to the LSTM model, suggesting that the larger PRO-ACT dataset better utilizes this model's abilities. The LSTM is a model in which the prediction time does not increase the differences between using the TASMC or PRO-ACT patients. The LSTM model can better capture the temporal dimension of the data, compared with the flattened data for the non-temporal models, and might reduce the risk of over-fitting to short-term predictions. In general, the ALS disease predictions using the LSTM for the clinic's (test) patients is quite similar when the model is trained on the PRO-ACT or on the clinic (training) patient population. Thus, the PRO-ACT-based LSTM model can be generalized and used in the ALS clinic, even if the clinic lacks sufficient records to create its own model using its own patients, and might even be more accurate for short-term predictions tasks.

In addition, comparing the RMSE to that of the previous experiment, we can see that feature removal (omitting all the features from the TASMC database that do not exist in the PRO-ACT) did not change the estimated prediction error, strengthening the conclusion from the feature importance analysis, that the features of the clinical trials data are also appropriate for the clinic data.

C. Experiment 3—Domain adaptation from clincal trials to the tertiary-care clinic

Although the naïve PRO-ACT-trained model was found (Fig. 3) reasonable to use on the clinic patients, we might be able to improve its predictions further by also using relatively small amount of information already available in the clinic.

In this experiment, we applied domain adaptation to combine information from the two datasets with the hope of improving the LSTM and MLP predictions for the clinic's patients. The training process included two stages: (1) Train an LSTM/MLP model using the PRO-ACT patients, and (2) fine-tune training using the TASMC (training) patients. This type of training and re-training allowed us to fit a model using a large, even if biased, clinical trials database while reducing the risk of over-fitting the data.

TABLE I. RMSE AND MAE BOOTSTRAPPING RESULTS (AVERAGE AND STANDARD DEVIATION) OVER 1,000 REPETITIONS OF SAMPLING 800 TEST OBSERVATIONS WITHOUT REPLACEMENT FOR EACH ALGORITHM AND PREDICTION TIME INTERVAL IN THE PRO-ACT DATABASE.

	XGB		**RF**		**MLP**		**LSTM**	
Prediction time interval (moths)	***RMSE***	***MAE***	***RMSE***	***MAE***	***RMSE***	***MAE***	***MAE***	***RMSE***
(0, 6]	**2.65 (.10)**	**1.98 (.06)**	2.69 (.10)	2.01 (.06)	2.91 (0.12)	2.15 (0.07)	2.98 (.11)	2.19 (.07)
(6, 12]	**3.98 (.12)**	**3.09 (.09)**	4.10 (.12)	3.20 (.09)	4.15 (0.14)	3.19 (0.09)	4.35 (.14)	3.34 (.10)
(12, 18]	**4.92 (.13)**	**3.93 (.11)**	5.07 (.14)	4.05 (.11)	5.19 (0.13)	4.17 (0.11)	5.42 (.15)	4.32 (.12)
(18, 24]	5.57 (.14)	**4.42 (.12)**	**5.56 (.12)**	4.53 (.11)	5.97 (0.14)	4.78 (0.13)	6.27 (.15)	5.06 (.13)
(0, 24]	**3.42 (.13)**	**2.54 (.09)**	3.51 (.12)	2.61 (.08)	3.65 (0.13)	2.71 (0.09)	3.78 (.13)	2.78 (.09)

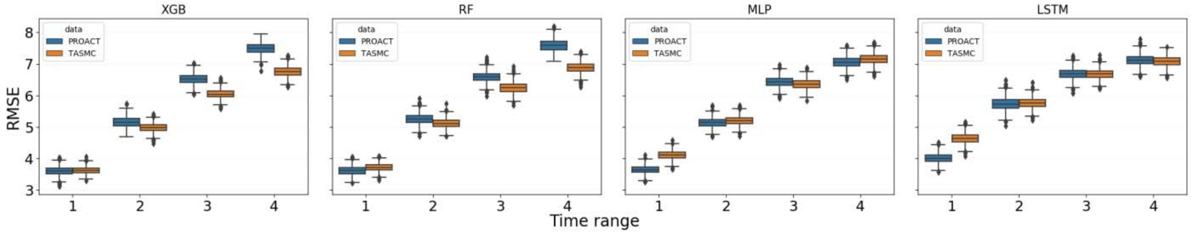


Fig. 3. Boxplots of the RMSE over 1,000 test subsets (each having 800 random observations) divided into four prediction time intervals .

Tables II and III show that training a model that was first trained on the PRO-ACT database and next fine-tuned on patients belonging to the TASMC test patient population improves the predictions for any prediction time, expressed by lowering of both the RMSE and MAE of TASMC test patient predictions (except for short-term prediction performances of the MLP model). Since the PRO-ACT, compared to the clinic data set, is characterized by frequent visits in a short period of time, its contribution to training is for short period predictions. However, adaptation to the domain of clinic data for which the intervals between visits are longer gives the adapted model advantage for longer periods at the expense of short period predictions, what explains the larger advantage of domain adaptation to long period predictions over short ones.

Not only that domain adaptation helped LSTM improve performance compared to when it used only, but the entire, clinic data set (Table II), but also compared to XGB (Fig. 3). XGB is not significantly better with RMSEs of: 3.50, 4.73, and 5.89 compared with RMSEs of 3.75, 5.13, and 6.27 with domain adaptation for prediction time intervals (0, 6], (6, 12], and (12, 18]. As for prediction to farther time intervals (18–24 months), the adapted LSTM (Table II) is even more accurate than the XGB (Fig. 3), achieving a lower RMSE of 6.68 vs. 6.90.

Fig. 4 demonstrates the prediction improvement when using the PRO-ACT trained MLP model and fine-tuning using the TASMC clinic data rather than training a model only using TASMC. When using up to 60% of the data of the training patients (320 patients), fine-tuning the PRO-ACT pre-trained MLP model was better than training a model based only the clinic training set. In other words, data of at least 320 patients are necessary to develop a model accurate enough for the clinic usage as if it used a pre-trained model using a much larger database of clinical trials.

TABLE II. COMPARISON OF TASMC (TEST) PATIENTS' PREDICTIONS FOR LSTM TRAINED BASED ONLY ON: (1) TASMC (TRAINING) PATIENTS, (2) PRO-ACT PATIENTS, AND (3) BOTH DATASETS (DOMAIN ADAPTATION).

	Only TASMC		**Only PRO-ACT**		**Domain adaptation**	
Prediction time (months)	***RMSE***	***MAE***	***RMSE***	***MAE***	***RMSE***	***MAE***
(0, 6]	4.65 (0.17)	3.44 (0.11)	4.02 (.15)	2.95 (0.09)	**3.75 (.11)**	**2.89 (.08)**
(6, 12]	5.74 (0.16)	4.46 (0.12)	5.70 (.18)	4.28 (0.12)	**5.13 (.16)**	**4.01 (.11)**
(12, 18]	6.67 (0.15)	5.46 (0.13)	6.72 (.16)	5.36 (0.14)	**6.27 (.14)**	**5.14 (.13)**
(18, 24]	7.08 (0.15)	5.63 (0.14)	7.12 (.16)	5.68 (0.15)	**6.68 (.16)**	**5.24 (.15)**
(0, 24]	5.45 (0.18)	4.15 (0.14)	5.20 (.17)	3.84 (0.12)	**4.76 (.15)**	**3.65 (.11)**

TABLE III. COMPARISON OF TASMC (TEST) PATIENTS' PREDICTIONS FOR MLP TRAINED BASED ONLY ON: (1) TASMC (TRAINING) PATIENTS, (2) PRO-ACT PATIENTS, AND (3) BOTH DATASETS (DOMAIN ADAPTATION).

	Only TASMC		**Only PRO-ACT**		**Domain adaptation**	
Prediction time (months)	***RMSE***	***MAE***	***RMSE***	***MAE***	***RMSE***	***MAE***
(0, 6]	3.84 (0.11)	3.01 (0.09)	**3.64 (0.13)**	**2.72 (0.09)**	3.68 (0.11)	2.89 (0.09)
(6, 12]	5.06 (0.13)	4.01 (0.11)	5.14 (0.15)	3.99 (0.12)	**4.94 (0.12)**	**3.92 (0.11)**
(12, 18]	6.30 (0.15)	5.08 (0.13)	6.42 (0.16)	5.17 (0.13)	**6.14 (0.15)**	**4.97 (0.13)**
(18, 24]	7.19 (0.16)	5.88 (0.14)	7.05 (0.17)	5.75 (0.14)	**7.00 (0.16)**	**5.63 (0.14)**
(0, 24]	7.19 (0.16)	5.88 (0.14)	4.85 (0.16)	**3.65 (0.11)**	**4.74 (0.14)**	**3.65 (0.11)**

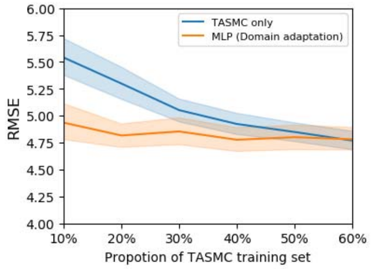


Fig. 4. The RMSE on the TASMC test set achieved using the MLP model trained only using TASMC clinic training data (blue) vs. the fine-tuned PRO-ACT pre-trained model (MLP) (orange) as a function of the TASMC training set size (% of the original training set).

V. Discussion and future work

The ability to accurately predict the ALS disease course is very important to ALS patients, their families, and doctors, as well as pharmaceutical companies. The high heterogeneity of the disease regarding its progression in the patient population is considered to be the main challenge in assessing accurate disease state prediction models. Furthermore, due to the rarity of the disease and its data, it is difficult for ALS clinics to establish a large enough dataset to be the basis of machine-learning models, which increases the necessity of alternative and creative ways to exploit the data the clinic possesses.

In this study, we compared four state-of-the-art algorithms and concluded that, in terms of prediction error, the most suitable approach is to use non-temporal algorithms such as XGB, which outperformed the RF, MLP, and the LSTM (as the latter highly parameterized model is more prone to overfitting). Nevertheless, for an ALS clinic, which might not possess a complete dataset of the patients' clinic visit history or observations for enough patients, using the clinical trials PRO-ACT database as the training set to initiate domain adaptation might be a good alternative for developing a prediction model using only the small clinic data alone. Indeed, domain adaptation for the LSTM and MLP using a pre-trained PRO-ACT model provided a model that is superior to both the PRO-ACT trained model and the TASMC trained model, and is even on par with the best XGB model. As the amount of clinic data increases, so will increase the gain from domain adaptation.

It is important to understand the advantages and disadvantages of each approach. The non-temporal model (say XGB) might be the best choice in terms of prediction accuracy and interpretation for ALS clinics with large enough databases. On the other hand, more effort in data pre-processing and feature engineering is needed compared to using the LSTM-based model. For ALS clinics with only small databases, it is better to use the PRO-ACT trained model or, even better, the domain adapted model. There is no straightforward way to apply an approach like domain adaptation, as we implemented with the LSTM and MLP, using tree-based models like the RF and XGB, hence, this may be a direction of future research.

Another possible way to improve prediction accuracy for clinic patients is by handling the short-term prediction overfitting caused by the biased training set using, e.g., up-sampling long-term prediction observations, different architectures of the LSTM to benefit more from its abilities in domain adaptation, or the XGB input (the flattened temporal data) as an additional input on top of the LSTM layer. Our code is available *online*, and so is the processed *The PRO-ACT dataset*. The TASMC data set, however, cannot be shared due to privacy issues.

References

[1] J. Burrell, S. Vucic and M. Kiernan, "Isolated bulbar phenotype of amyotrophic lateral sclerosis", *Amyotrophic Lateral Sclerosis*, vol. 12, no. 4, pp. 283-289, 2011.

[2] J. Mitchell, G. Borasio, "Amyotrophic lateral sclerosis", *The Lancet*. vol. 369, pp. 2031–2041, 2007.

[3] A. Renton, A. Chiò and B. Traynor, "State of play in amyotrophic lateral sclerosis genetics", *Nature Neuroscience*, vol. 17, no. 1, pp. 17-23, 2013.

[4] L.P. Rowland, "Ameliorating Amyotrophic Lateral Sclerosis", *New England Journal of Medicine*, vol. 362, no. 10, pp. 953-954, 2010.

[5] J. Rothstein, "Current hypotheses for the underlying biology of amyotrophic lateral sclerosis", *Annals of Neurology*, vol. 65, no. 1, pp. S3-S9, 2009.

[6] R. Küffner, N. Zach, R. Norel, J. Hawe, D. Schoenfeld et al., "Crowdsourced analysis of clinical trial data to predict amyotrophic lateral sclerosis progression", *Nature Biotechnology*. vol. 33, pp. 51–57, 2014.

[7] N. Zach, D.L. Ennist, A.A. Taylor, H. Alon, A. Sherman et al "Being PRO-ACTive: What can a Clinical Trial Database Reveal About ALS?", *Neurotherapeutics*, vol. 12, no. 2, pp. 417-423, 2015.

[8] N. Atassi, J. Berry, A. Shui, N. Zach, N. Sherman et al., "The PRO-ACT database: Design, initial analyses, and predictive features", *Neurology*, vol. 83, no. 19, pp. 1719-1725, 2014.

[9] A.A. Taylor, C. Fournier, M. Polak, L. Wang, N. Zach et al "Predicting disease progression in amyotrophic lateral sclerosis". *Annals of Clinical and Translational Neurology*, vol. 3, pp. 866–875, 2016.

[10] L. Breiman, "Random forests", *Machine Learning*, vol. 4, no. 1, pp. 5-32 (2011).

[11] T. Chen, C. Guestrin, XGBoost. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 16, 2016.

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory", *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997.

[13] A. Nahon, B. Lerner, "Temporal modeling of ALS using longitudinal data and long-short term memory-based algorithm", *ESANN*, 2018.

[14] M. Ong, P. Tan and J. Holbrook, "Predicting functional decline and survival in amyotrophic lateral sclerosis", *PLoS ONE*, vol. 12, no. 4, p. e0174925, 2017.

[15] B. Brooks, R. Miller, M. Swash and T. Munsat, "El Escorial revisited: Revised criteria for the diagnosis of amyotrophic lateral sclerosis", *Amyotrophic Lateral Sclerosis and Other Motor Neuron Disorders*, vol. 1, no. 5, pp. 293-299, 2000.

[16] J.M. Cedarbaum, N. Stambler, E. Malta, C. Fuller, D. Hilt et al., "The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function", *Journal of the Neurological Sciences*, vol. 169, no. 1-2, pp. 13-21, 1999.

[17] R. Gomeni and M. Fava, "Amyotrophic lateral sclerosis disease progression model", *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, vol. 15, no. 1-2, pp. 119-129, 2013.

[18] T. Hothorn and H. Jung, "RandomForest4Life: A random forest for predicting ALS disease progression", *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, vol. 15, no. 5-6, pp. 444-452, 2014.

[19] M. Tang, C. Gao, S.A. Goutman, A. Kalinin, B. Mukherjee et.al., "Model-based and model-free techniques for amyotrophic lateral sclerosis diagnostic prediction and patient clustering." *Neuroinformatics*", vol. 17, no. 3, pp. 407-21, 2019.