Baseline Machine Learning Approaches To Predict Amyotrophic Lateral Sclerosis Disease Progression

Notebook for the iDPP Lab on Intelligent Disease Progression Prediction at CLEF 2022

Isotta Trescato[1]*, Alessandro Guazzo[1]*, Enrico Longato[1], Enidia Hazizaj[2], Chiara Roversi[1], Erica Tavazzi[1], Martina Vettoretti[1] and Barbara Di Camillo[1,3]

[1] *Department of Information Engineering, University of Padova, Padova, Italy*

[2] *Department of Pharmaceutical and Pharmacological Sciences, University of Padova, Padova, Italy*

[3] *Department of Comparative Biomedicine and Food Science, University of Padova, Padova, Italy*

* *These authors contributed equally*

Abstract

Amyotrophic lateral sclerosis (ALS) is a rapidly progressive neurodegenerative disease that typically leads to death within 3-5 years, characterised by a heterogeneous progression across the patient population. This heterogeneity has hindered efforts to assess the efficacy of developmental treatments designed to delay disease progression and prolong survival. As such, prediction of disease progression has been a long-standing interest in the field as a means of enabling better drug development using cheaper, more accurate clinical trials, as well as deriving new and important insights into disease mechanisms and manifestations. So far, this critical point has not yet been sufficiently addressed due to limited access to patient-level data and sophisticated computational tools. This contribution aims at comparing the performance of different baseline machine learning approaches on a common dataset obtained via the integration of different datasets from different countries provided by the challenge organisers. Results show that the ability of different methods across different subtasks to discriminate among subjects at risk and to predict the time of adverse events improves as dynamic variables, monitoring the first six months of patient follow-up, are included as possible predictors.

Keywords

Cox Proportional Hazard Model, Survival Support Vector Machine, Random Survival Forest, Amyotrophic Lateral Sclerosis, Baseline

1. Introduction

Amyotrophic lateral sclerosis (ALS) is a disease that affects motor neurons, leading to progressive paralysis and death, mostly from respiratory failure, typically within 3-5 years although

CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

trescatois@dei.unipd.it (I. Trescato); guazzoales@dei.unipd.it (A. Guazzo); enrico.longato@dei.unipd.it (E. Longato); enidia.hazizaj@studenti.unipd.it (E. Hazizaj); roversic@dei.unipd.it (C. Roversi); erica.tavazzi@unipd.it (E. Tavazzi); martina.vettoretti@dei.unipd.it (M. Vettoretti); barbara.dicamillo@unipd.it (B. Di Camillo)

https://www.dei.unipd.it/~dicamill/ (B. Di Camillo)

0000-0003-0625-993X (I. Trescato); 0000-0001-5155-2567 (A. Guazzo); 0000-0001-5940-645X (E. Longato); 0000-0001-8327-4530 (E. Hazizaj); 0000-0001-5320-1082 (C. Roversi); 0000-0001-6188-6413 (E. Tavazzi); 0000-0002-5020-1818 (M. Vettoretti); 0000-0001-8415-4688 (B. Di Camillo)

CEUR Workshop Proceedings (CEUR-WS.org)

survival varies greatly between patients [1]. In recent years, several novel ALS treatments were tested in large clinical trials, but most of them failed to slow down disease progression significantly [2]. Clinical heterogeneity has a multifaceted impact on both managing the disease and planning treatments; and, regrettably, currently available clinical tools cannot differentiate between patients with different manifestations and prognoses. While genetic and molecular understanding of disease processes is still in its early stages, there is already clear evidence for different clinical manifestations and likely different disease mechanisms across ALS patients, suggesting that ALS might comprise several different clinical phenomena [3]. The heterogeneity in the course of ALS clinical progression and, ultimately, survival, coupled with the rarity of this disease, makes predicting disease outcomes and clinical events at the patient level very challenging. This presents substantial barriers to the planning and interpretation of clinical studies of new treatments, leading to large, expensive, and potentially unbalanced trials [4].

Previous efforts have been made to identify the factors most relevant for stratifying the ALS patient population into groups with different clinical manifestations and prognoses. Site and age of onset have been shown to affect overall survival time and rate of disease progression. Other factors connected to prognosis include body mass index (BMI), absolute weight, cognitive function, and genetic background [5]. Recently, serum-based predictors have been reported for the first time, raising hopes for novel biomarker development and new insights on disease mechanisms. These include uric acid, albumin, and creatinine [6, 7]. However, these efforts were hindered by the limited data size and the resulting inability to pursue complex analytical approaches.

In this paper we address the challenge proposed in the context of iDPP@CLEF 2022, in which participants were asked to rank subjects based on their risk of early occurrence of non-invasive ventilation (NIV), percutaneous endoscopic gastrostomy (PEG), or death (Task 1) and to predict the time window of the event (Task 2), using a large dataset of 2559 patients and 68 variables, 49 of which static and 19 dynamic. For each of these tasks, participants are given a dataset containing 6 months of visits and are asked to either only use data available until time 0, i.e., the time of the first ALSFRS-R questionnaire, or all data available until month 6. This aligns with the point of view of a clinician who wants understand what can be predicted the first time the patient is visited (time 0) and after a follow-up of 6 months [8, 9]. In particular, three different baseline survival analysis approaches were used, i.e., the Cox proportional hazard model, the survival support vector machine (SSVM) and the random survival forest (RSF). Following the iDPP challenge rules that limited the number of different submissions, only the two best-performing models for each subtask and each setting (0 or 6 months) were considered. The submitted models were selected according to their average performance across bootstrap iterations, by means of the average out-of-bag C-index. Performance was evaluated via the metrics identified by the challenge organisers, i.e., C-index, area under receiver operating characteristic curves (AUROCs), Brier scores (BS) for Task 1, and absolute values of the distances between predicted and true event time windows (Abs distance), specificity and recall for Task 2 [10, 11].

The paper is organized as follows: Section 2 introduces related works and the main methodological approaches implemented until now to address ALS progression prediction. Section 3 describes the approach in terms of data preprocessing and adopted machine learning approaches. Finally, section 4 discusses the findings. SSVM was always selected, while Cox and RSF models only for some subtasks. Section 5 draws conclusions and presents the outlook for future work.

2. Related Work

Different approaches have been proposed in the literature for describing and predicting the prognosis of ALS patients, with most of them modelling progression by considering the impairment of functional capabilities [12, 13], the need for support interventions [14, 15, 16], the rate of future progression [17, 18], or a survival endpoint that often includes the administration of tracheotomy as a proxy for death [19, 20, 21, 22, 23].

Depending on the specifics of the research question and the corresponding selected outcome, in the literature, prediction tasks are modelled via different machine learning frameworks, namely classification, regression, or survival analysis.

The most popular and successful regression techniques for forecasting ALS prognosis include random forests (RFs) [17, 21, 24, 25, 26], gradient boosting [25, 26], and generalised linear models [21, 27]. More recently, graphical modeling techniques such as the dynamic Bayesian networks have been used to model disease progression in its entirety [12, 13]. Prognostic classification tasks, such as the categorization of patients as slow/fast progressors or the prediction of the occurrence of a clinical endpoint at a given time, are mainly based on logistic regression [15, 19, 20], RF [14, 18, 20, 21], support vector machines (SVM) [14, 16], and naïve Bayes classifier [14, 28]; while the Cox proportional hazard model [22, 23, 25] is the most frequent approach in the survival analysis setting.

Noticeably, while most of these works use a set of variables collected at a baseline visit (e.g., the visit at the time of ALS diagnosis, or the first visit of a clinical trial) as predictors, and attempt to predict a clinical outcome associated with disease progression over a follow-up period, some approaches enrich the baseline information by including the data collected over a time window instead of at a single time point [14, 25, 28]. These dynamic data are preprocessed to extract features – such as the mean value or the slope – describing initial progression trends, and then used to predict how the disease will further progress. As another option for using the clinical information collected during the patients' screening visits, some literature works model the entire disease progression course using all available information, and investigate the relationships among the variables and their effect on the outcomes over time [12, 13].

3. Methodology

In this paragraph, the experimental workflow will be described using the nomenclature used in the iDPP challenge guidelines. The two tasks will be referred to as:

- **Task 1**: ranking risk of impairment for ALS;
- **Task 2**: predicting time of impairment for ALS;

and the three subtasks as:

- **subtask a**: NIV or (competing event) Death, whichever occurs first;
- **subtask b**: PEG or (competing event) Death, whichever occurs first;
- **subtask c**: death.

For each subtask, two types of submission were required:

- **dataset M0**: submissions of predictions obtained using only data available until time 0, i.e., the time of the first registered visit or, equivalently, the first available ALSFRS-R questionnaire for each patient;
- **dataset M6**: submissions using data available until month 6.

A common preprocessing and model-training pipeline was devised and applied to all subtask-specific datasets provided by the organisers. The following sections document: the preliminary data analysis aimed at evaluating overall data quality (section 3.1.1); the data manipulation steps needed to obtain the final set of static and dynamic input variables (sections 3.1.2 and 3.1.3); the differences between the baseline (M0) and progression (M6) datasets as defined by the challenge guidelines (section 3.1.4); the normalisation (section 3.1.6) and imputation (section 3.1.7) procedures; and the model-dependent workflows for training, feature selection, and hyperparameter optimisation via bootstrap resampling (sections 3.1.5, 3.2.1, and 3.2.2).

3.1. Preprocessing

3.1.1. Quality Control

A preliminary data quality analysis was performed to verify that the patient *id*s were consistent between the provided csv files, that there were no subjects with more than 20% of missing values, and that each subject had at least one ALSFRS-R score measurement. These steps resulted in the exclusion of one subject (*id* = 0x712eabc24c9a530d1739ada22320183d) whose *onsetDate* was > 0. A positive *onsetDate* suggests that the subject had been administered the ALSFRS-R questionnaire before the disease's onset, which appears to be inconsistent with standard clinical practice. Thus, all the recordings related to this subject were removed.

3.1.2. Static Variables' Preprocessing

The variable *sex* was reported as a string with two levels, "male" and "female". It was mapped to a boolean variable such that 0 = "male" and 1 = "female". Similarly, the variables *smoking*, *hypertension*, *diabetes*, *dyslipidemia*, *thyroid_disorder*, *autoimmune_disease*, *stroke*, *cardiac_disease*, and *primary_neoplasm* were converted into booleans with values 0/1: missing values were converted to 0 and "TRUE" values to 1.

In the original version of the datasets, genetic mutations were reported separately for the two clinical centres by which the data were made available. The eight variables *lisbon_FUS_openbox*, *lisbon_SOD1_openbox*, *lisbon_TARDBP_openbox*, *lisbon_C9orf72*, *turin_FUS*, *turin_SOD1*, *turin_TARDBP*, and *turin_C9orf72_kind* were then combined into four boolean variable: *FUS*, *SOD1*, *TARDBP*, and *C9orf72*.

The datasets also included detailed information on traumas and surgical interventions, specifying the part of the body involved (*head, neck, cervical, thoracic, lumbosacral, cervical spine, thoracic spine, upper limb, lower limb, or abdominal*) and the time interval in which the event occurred (denoted by the suffixes *_last_5_ years* and *_more than_5_years*). We created two boolean variables *trauma_after_onset* and *surgery_after_onset* that were equal to 1 if and only if a subject had had any type of trauma or surgery, respectively, at any time after the onset. The two columns related to traumas and surgical interventions before onset (*major_trauma_before_onset,* *surgical_interventions_before_onset*) were kept as in the original version, but they have been renamed to *trauma_before_onset* and *surgery_before_onset*, respectively.

The variable *occupation* was given as a string with a controlled vocabulary of 41 different levels. This variable was mapped into a new variable, *instruction_level*, for which the value 1 corresponds to an 8th grade diploma, the value 2 to a high school diploma, the value 3 to a university degree, and the value 4 to a highly specialised or a managerial profile. The implicit assumption, here, was that a patient's current job and the minimum instruction level needed to qualify for that job were tightly correlated.

Five variables related to onset site were available: *onset_bulbar*, *onset_axial*, *onset_limbs*, *onset_generalized*, and *onset_limb_type*. While *onset_bulbar* and *onset_axial* are self-explanatory, *onset_limbs* and *onset_generalized* are further detailed in the last variable, *onset_limb_type*. Precisely, if and only if one between *onset_limbs* and *onset_generalized* was equal to 1, the same subject also had a value in *onset_limb_type*, reported as a string with 29 levels. The levels specified the impacted side (*left* or *right*), and distinguished between *distal/proximal* and *lower/upper*. The first two distinctions were ignored, whereas the third one, being, a priori, the most discriminative, was mapped into two new boolean variables (*onset_limb_upper* and *onset_limb_lower*) as follows.

- If *onset_limb_type* contained "upper", then *onset_limb_upper* $= 1$ and *onset_limb_lower* $= 0$;
- If *onset_limb_type* contained "lower", then *onset_limb_lower* $= 1$ and *onset_limb_upper* $= 0$;
- If *onset_limb_type* contained "upper-lower" or did not contain information on the involved part, then *onset_limb_upper* $= 1$ and *onset_limb_lower* $= 1$.

Ultimately, to describe the onset site, *onset_limbs* and *onset_generalized* were removed in favour of the two new variables *onset_limb_upper* and *onset_limb_lower*, which were maintained together with *onset_axial*. *Onset_bulbar* was removed because it was a deterministic function of the other three variables ($onset\_bulbar = 1$ if and only if $onset\_limb\_upper = onset\_limb\_lower = onset\_axial = 0$).

Three variables specifying motor neuron (MN) prevalence were available: *prevalentLMN*, *prevalentUMN*, and *mixedMN*. At first, we created the variable *not_specified_MN*, which was equal to 1 when all the other tree MN variables were NAs. Then, we transformed these variables in three dummy variables, removing *prevalentLMN*, and mantaining *prevalentUMN*, *not_specified_MN*, and *mixedMN*.

BMI was computed for each patient as $BMI = weight/height^2$ combining the measures of the variables *weight* and *height*, which were then removed.

The variable *ethnicity* was removed because less than 1% of the observations were different from *Caucasian*. Moreover, all the variables with more than 70% of missing values were removed: all the twenty variables related to blood tests, *retired at diagnosis*, *packYear*, *dailyCigarettes*, *smoking_startYear*, *smoking_endYear*, and *alive*.

3.1.3. Dynamic Variables' Preprocessing

In addition to the static variables described above, dynamic information was available, allowing for the creation of dynamic variables to capture the changes in patient characteristics over time. Dynamic variables were all in the form of monthly rates of change, between *onsetDate* and time 0, or, only for dataset M6, between time 0 and the last available visit.

In the original datasets, two measures of weight (*weight* and *weight_before_onset*) were provided. Since weight change is to be considered an indicator of ALS progression [29, 30], the slope of weight change in kg/month was computed as the angular coefficient of a linear fit between the points (*onsetDate; weight_before_onset*) and (*time0; weight*).

Separate csv files, containing one row for each visit in which a forced vital capacity (FVC) value or an ALSFRS-R score was recorded, were also available.

The variable *fvcValue* had $35 - 40\%$ of missing values in the three datasets (42% subtask a; 37% subtask b; 38% subtask c). Approximately 40% of patients had one recorded *fvcValue*, 20% had more than one, and 40% had none. To homogenise this variable across subjects, a single value of FVC per subject was retained according to the following criteria.

- If the subject had only one value, then it was kept as is.
- If a subject had more than one value, then the minimum was selected.
- If a subject had no values, then a value was imputed (more details in section 3.1.7).

The questions used to compute the ALSFRS-R score can be split into five sub-domains with questions 1 to 3 counting for bulbar impairment, 4 and 5 for upper limbs impairment, 6 and 7 for trunk impairment, 8 and 9 for lower limbs impairment, and 10 to 12 for breathing impairment. For each sub-domain, the questions comprising it were manipulated as follows to obtain three different assessments of the ALSFRS-R score: the score at time 0, the ALSFRS-R slope in the time period between *onsetDate* and time 0, and the ALSFRS-R slope during the observation time (i.e., between time 0 and month 6).

1. A virtual visit with the maximum score, equal to 4 for each question and summing to a total of 8 (upper limbs, trunk, lower limbs) or 12 (bulbar, breathing), depending on the number of questions in each sub-domain, was added as if there were a visit confirming that the subject did not have ALS before the onset time (hence, the maximum score).
2. The ALSFRS-R score at time 0 was renamed to *alsfrs_baseline_<sub-domain>*.
3. The slope between the virtual visit at onset time and the visit at time 0 was computed as the angular coefficient of the corresponding linear fit. The resulting variable (*slope_baseline_<sub-domain>*) was an indicator of the speed of worsening of the patient's condition in each sub-domain between onset and first visit.
4. The slope between the visit at time 0 and the last available visit of each subject within the first six months from time 0 was computed as the angular coefficient of the linear fit over all available visits. The resulting variable (*slope_progression_<sub-domain>*) was an indicator of the speed of worsening of the patient's condition in each sub-domain during the first 6 months, to be included only in dataset M6.
5. For approximately 20% of the patients, a direct computation of *slope_progression_<sub-domain>* was not possible, as they did not have multiple visits during the six-months period.

For those patients, *slope_progression_<sub-domain>* was set equal to *slope_baseline_<sub-domain>* under the assumption of a constant deterioration rate before and after the first visit.

3.1.4. Baseline (M0) and Progression (M6) Datasets

Following the challenge's guidelines, two different versions of the datasets were identified. For dataset M0, comprising only the information available at time 0, all the *slope_progression_<sub-domain>* variables were removed, as well as all subjects whose *outcome_time* was equal to 0. For dataset M6, including all information until month 6, all the variables, preprocessed as described in sections 3.1.1, 3.1.2, and 3.1.3 were retained, but all subjects whose *outcome_time* was lower than 6 were removed.

3.1.5. Bootstrap

For the dual purpose of performing feature selection and hyperparameter optimisation (section 3.2.1), and extracting 100 independent classifiers to obtain an ensemble classifier (section 3.2.2), 100 bootstrap sets of size equal to that of the entire subtask dataset were created. Each bootstrap set, containing approximately 63.2% of the subjects, possibly with repetition, was used as an internal training set, whereas the remaining 36.8% out-of-bag subjects were used as the corresponding validation set.

3.1.6. Normalisation

Normalisation is helpful to avoid introducing bias related to the different dynamic range of each variable, and to promote consistency between the scale of the coefficients that might be estimated during model training. Here, min-max scaling was used. In practice, let x be the variable to be normalised, this method constrains x into the range [0-1] according to the following equation.

$$x_{scaled} = \frac{x - min(x)}{max(x) - min(x)} \tag{1}$$

The normalisation parameters were derived separately on the 100 bootstrap-generated training sets and applied to the respective 100 validation sets.

3.1.7. Imputation

Imputation of the remaining missing values in the preprocessed input variables was performed using the mice R package [31] with 20 multiple imputations, 20 iterations, and random forest as imputation method (*m=20, maxit=20, method="rf"*). The imputation parameters were estimated separately on the 100 bootstrap-generated training sets and applied to the respective 100 validation sets. To check the robustness of the imputation process, we compared the distributions of each variable before and after imputation. The variables imputed were: *fvc* ($\sim 49\%$ missing), *instruction_level* ($\sim 30\%$ missing), *bmi* ($< 10\%$ missing), and *slope_weight* ($\sim 25\%$ missing).

3.2. Model Training

Three survival analysis methods were considered, namely: Cox, SSVM, and RSF. They were chosen to represent a broad spectrum of baseline models including parametric (SSVM), semi-parametric (Cox), linear (Cox, SSVM), and nonlinear (RSF) models. The same training workflow, based on feature selection and hyperparameter optimisation was considered for the Cox and SSVM models (section 3.2.1), and an ad-hoc one, leveraging their ability to differentiate variables influence on prediction according to their importance, for the RSF (section 3.2.2). For each subtask, two versions of all models were independently trained, one on dataset M0, and the other on dataset M6.

The Cox model and the RSF can only output risk scores, which can be used to address Task 1 by ranking ALS patients according to their risk of impairment, but do not provide a straightforward solution to predicting Task 2's time of impairment. To extend these two methodological approaches to Task 2, the predicted time of impairment for a given patient was selected as the median predicted time to impairment, i.e., the time at which the estimated survival function crossed the 0.5 threshold. Instead, the SSVM can be used either as a ranker or a time regressor depending on how the risk ratio hyperparameter is set during model training. Here, the SSVM was initially trained as a time regressor to address Task 2 directly. Then, its predicted times were converted into risk scores in the range [0-1], as requested by the challenge rules, via Platt scaling [32].

3.2.1. Cox Model and SSVM Training Workflow

The same training workflow was used for Cox and SSVM models. For each bootstrap set, an optimal model was obtained by performing hyperparameter optimisation and feature selection. The optimal set of features was obtained using the forward recursive feature selection (FRFS) approach [33]. At each FRFS step, the best feature to be added was chosen as the one that led to the maximum Harrell's concordance index (C-index) on the validation set of the considered bootstrap set. The early stopping criterion for FRFS was a relative variation of the C-index $< 0.01\%$.

Hyperparameters were optimized using a random search approach [34] over 200 possible values. For the Cox model, the only hyperparameter that needed optimisation was the strength of the L2 regularisation (α, randomly sampled from a log-uniform distribution with support $[10^{-8}$-$10^{3}]$). For the SSVM model, a linear kernel was used and the rank ratio was set equal to 0 so as to obtain a time regression model applicable to both tasks (1 and 2) as explained in section 3.2. The only hyperparameter of the SSVM was the penalisation weight of the squared hinge loss objective function (β, randomly sampled from a log-uniform distribution with support $[10^{-7}$-$10^{4}]$). For both approaches, the best hyperparameters (i.e., the subset of most important features, the strength of regularisation α for the Cox model, and the penalisation weight β for the SSVM) were chosen as those that maximised the C-index on the validation set of each bootstrap set.

The features selected for the optimal models of each bootstrap set were used to obtain the optimal set of features needed to train a single final model. This last set of features was selected based on the average number of FRFS steps at which a variable was added to the set of predictors in the 100 bootstrap sets. If a variable was never selected in a given set, its associated number of steps was set to the maximum number of variables plus one. The final model was then trained on the whole training set using only the variables whose average number of steps was smaller than average. The final model's hyperparameters (α for the Cox model and β for the SSVM) were optimised again via a 10-fold cross validation and a random search approach. The best hyperparameters were chosen as those that led to the maximum cross-validation C-index.

3.2.2. Random Survival Forest Training Workflow

The workflow used to train the RSF models did not include the feature selection step. Instead, the 100 different bootstrap sets were used to train 100 models, to be used as an ensemble classifier.

For each bootstrap set, 200 models were trained, varying the combination of two model hyperparameters, i.e., the number of trees to be grown and the maximum depth of each tree. The values tested for the number of trees were randomly sampled in the range [50-750] while the values for the maximum depth were randomly sampled in the range [10-200], adding the option "max_depth" (i.e., the longest path between the root node and the leaf node, where a deeper tree means a more complex model). For each bootstrap set, the best hyperparameter combination among the 200 tested was chosen as the one that maximised the C-index on the validation set.

The obtained 100 independent RSF models were applied to the test set, obtaining 100 predictions for each subject. The final prediction was obtained via the ensemble averaging the 100 predictions.

4. Results

For each subtask (a, b, c) and each dataset version (M0 and M6), the two best-performing models according to the average C-index on the 100 bootstrap sets were used to obtain the runs requested for performance evaluation. Each selected model was used to obtain a run for each task as explained in section 3.2, for a total of 24 submitted runs (2 tasks, 3 subtasks, 2 dataset versions). The SSVM was selected among the best models for all sub-tasks and all dataset versions. The Cox model was selected for subtask c when using dataset M0 and for subtasks a and c when using dataset M6. Finally, the RSF was selected for subtasks a and b when using dataset M0, and for subtask b when using dataset M6.

4.1. Task 1 Results

Table 1 shows the performance metrics of our submitted models for all Task 1 subtasks when using dataset M0. The models submitted for each of the three subtasks are reported in the columns of the table. The table rows are divided into three sections: the C-index is reported in the first section, area under receiver operating characteristic curves (AUROCs) at various time points from 12 to 60 months are reported in the second one, and Brier scores (BS), computed at the same time points considered for the AUROCs, are reported in the third one.

Model discrimination was acceptable, with C-index and AUROC values around 0.7 for all submitted models across all subtasks. The models performed better in sub-task b (prediction

Table 1

Task 1 metrics, datset M0: C-index and AUROC are reported with their estimated value as well as their 95% confidence intervals, only the computed value is reported for the brier score instead

	sub-task a		sub-task b		sub-task c	
Metric	SSVM	RSF	SSVM	RSF	SSVM	Cox
C-index	0.652 (0.620-0.683)	0.643 (0.612-0.675)	0.692 (0.664-0.720)	0.690 (0.662-0.717)	0.686 (0.658-0.713)	0.685 (0.658-0.713)
AUROC (12m)	0.742 (0.688-0.796)	0.642 (0.610-0.675)	0.768 (0.715-0.822)	0.748 (0.696-0.799)	0.784 (0.731-0.838)	0.760 (0.703-0.817)
AUROC (18m)	0.722 (0.668-0.776)	0.696 (0.637-0.755)	0.766 (0.720-0.812)	0.768 (0.722-0.815)	0.748 (0.702-0.795)	0.743 (0.696-0.790)
AUROC (24m)	0.720 (0.664-0.775)	0.708 (0.653-0.762)	0.757 (0.711-0.803)	0.741 (0.693-0.790)	0.751 (0.707-0.795)	0.753 (0.709-0.796)
AUROC (30m)	0.735 (0.678-0.793)	0.699 (0.643-0.755)	0.771 (0.723-0.818)	0.781 (0.733-0.828)	0.751 (0.706-0.795)	0.749 (0.705-0.794)
AUROC (36m)	0.739 (0.681-0.797)	0.730 (0.669-0.790)	0.774 (0.723-0.824)	0.782 (0.732-0.832)	0.751 (0.704-0.798)	0.747 (0.700-0.794)
AUROC (48m)	0.746 (0.678-0.815)	0.771 (0.704-0.839)	0.790 (0.731-0.849)	0.807 (0.754-0.860)	0.781 (0.728-0.834)	0.783 (0.731-0.835)
AUROC (60m)	0.701 (0.612-0.789)	0.779 (0.692-0.866)	0.783 (0.718-0.848)	0.797 (0.737-0.857)	0.785 (0.726-0.844)	0.789 (0.730-0.847)
BS (12m)	0.580	0.645	0.595	0.705	0.594	0.814
BS (18m)	0.449	0.499	0.438	0.522	0.479	0.667
BS (24m)	0.336	0.374	0.330	0.395	0.375	0.525
BS (30m)	0.260	0.290	0.238	0.282	0.282	0.398
BS (36m)	0.211	0.234	0.198	0.233	0.229	0.322
BS (48m)	0.124	0.136	0.126	0.145	0.138	0.190
BS (60m)	0.080	0.082	0.105	0.119	0.107	0.140

of PEG or death) than in the other subtasks according to all discrimination metrics. Model calibration, evaluated via the BS, was good only for long prediction horizons (PHs) (BS > 0.5 with PH = 12 months vs. BS < 0.15 with PH = 60 months).

Table 2 shows the performance metrics of our submitted models for all Task 1 subtasks when using dataset M6. The structure of the table is the same described above for Table 1. Model discrimination was, once again, acceptable with C-index values ranging around 0.7 and AUROC values greater than 0.75 for subtask a. Discrimination was better for subtasks b and c, with C-index values over 0.7 and AUROC values around 0.78. Model calibration was good only for long prediction horizons (PHs) (BS > 0.5 with PH = 12 months vs. BS < 0.15 with PH = 60 months).

Discrimination and calibration improved across all subtasks as dynamic variables were included in the pool of possible predictors (dataset M0 vs. dataset M6). This interesting result suggest that, for ALS, a fast progressing disease, collecting dynamic features for the first 6 months to describe the initial disease progression, instead of just focusing on baseline status, provides a good set of predictors that leads to better predictive power.

Higher AUROC values were obtained when the prediction horizon (PH) was short (12-18 months) or long (48-60 months). Such an oscillating behaviour was partially expected due to

Table 2
Task 1 metrics, dataset M6: C-index and AUROC are reported with their estimated value as well as their 95% confidence intervals, only the computed value is reported for the brier score instead

	sub-task a		sub-task b		sub-task c	
Metric	SSVM	Cox	SSVM	RSF	SSVM	Cox
C-index	0.696 (0.666-0.727)	0.694 (0.664-0.725)	0.708 (0.680-0.736)	0.716 (0.690-0.742)	0.713 (0.686-0.739)	0.706 (0.679-0.733)
AUROC (12m)	0.795 (0.744-0.845)	0.790 (0.739-0.841)	0.779 (0.723-0.834)	0.781 (0.731-0.832)	0.806 (0.752-0.859)	0.790 (0.734-0.845)
AUROC (18m)	0.789 (0.741-0.836)	0.783 (0.735-0.832)	0.785 (0.740-0.830)	0.800 (0.758-0.843)	0.788 (0.744-0.831)	0.775 (0.731-0.820)
AUROC (24m)	0.774 (0.724-0.824)	0.765 (0.714-0.817)	0.786 (0.743-0.829)	0.786 (0.742-0.830)	0.795 (0.754-0.835)	0.780 (0.738-0.835)
AUROC (30m)	0.768 (0.715-0.821)	0.763 (0.708-0.818)	0.787 (0.742-0.833)	0.809 (0.764-0.854)	0.775 (0.733-0.818)	0.764 (0.721-0.808)
AUROC (36m)	0.765 (0.708-0.821)	0.756 (0.698-0.814)	0.788 (0.739-0.836)	0.813 (0.765-0.861)	0.777 (0.733-0.821)	0.774 (0.730-0.819)
AUROC (48m)	0.768 (0.699-0.838)	0.771 (0.702-0.840)	0.806 (0.751-0.861)	0.823 (0.770-0.875)	0.802 (0.753-0.852)	0.801 (0.750-0.852)
AUROC (60m)	0.757 (0.670-0.845)	0.756 (0.667-0.846)	0.797 (0.736-0.857)	0.813 (0.754-0.872)	0.808 (0.754-0.862)	0.807 (0.752-0.861)
BS (12m)	0.587	0.675	0.599	0.701	0.608	0.815
BS (18m)	0.448	0.521	0.439	0.517	0.487	0.668
BS (24m)	0.333	0.389	0.328	0.390	0.375	0.525
BS (30m)	0.258	0.303	0.235	0.278	0.281	0.399
BS (36m)	0.206	0.242	0.194	0.229	0.226	0.322
BS (48m)	0.118	0.138	0.122	0.142	0.133	0.190
BS (60m)	0.073	0.082	0.104	0.117	0.103	0.140

the typical trade-off between PH length and number of recorded events as functions of time. Short PHs are expected to result in increased predictive power, because they are more tightly correlated (at the very least, temporally) with the input data, but, at the same time, might end too close to the start of follow-up for many events to have happened; conversely, longer PHs have a looser link to the input data (and are, thus, harder to predict), but more events are available, generally resulting in a more robust model training.

The BS decreased as the PH widened. This result was expected as the submitted models were trained without setting an artificial censoring time on the training set, thus leading to a better calibration in the long term, and to a general tendency to overestimate short-term event probability.

4.2. Task 2 Results

Table 3 shows the performance metrics of our submitted models for all Task 2 subtasks when using dataset M0. The models submitted for each one of the three subtasks are reported in the columns of the table. The table is divided in three sections along the rows according to the considered metric: the mean of the absolute values of the distances between predicted and true event time windows (Abs distance) is reported in the first section, the specificity and recall of the

Table 3
Task 2 metrics, dataset M0

	sub-task a		sub-task b		sub-task c	
Metric	SSVM	RSF	SSVM	RSF	SSVM	Cox
Abs distance	8.847	8.599	7.714	7.407	8.115	8.187
Specificity (6-12m)	0.968	0.967	0.958	0.976	0.977	0.987
Specificity (12-18m)	0.787	0.635	0.847	0.733	0.853	0.843
Specificity (18-24m)	0.641	0.728	0.734	0.800	0.766	0.727
Specificity (24-30m)	0.729	0.847	0.789	0.859	0.787	0.816
Specificity (30-36m)	0.944	0.935	0.855	0.877	0.849	0.898
Specificity (>36m)	0.988	0.968	0.958	0.935	0.903	0.876
Recall (6-12m)	0.082	0.0918	0.157	0.112	0.080	0.022
Recall (12-18m)	0.272	0.436	0.333	0.511	0.164	0.232
Recall (18-24m)	0.479	0.333	0.333	0.350	0.328	0.385
Recall (24-30m)	0.281	0.218	0.277	0.240	0.323	0.292
Recall (30-36m)	0.136	0.136	0.125	0.208	0.216	0.216
Recall (>36m)	0.055	0.164	0.289	0.350	0.382	0.426

predicted time window classification are reported in the second and third sections respectively for each allowed window.

The performance of the submitted models was good according to the mean absolute value distance as the predicted event times were, on average, only 7-8 months inaccurate with respect to the true event time. Specificity was high (> 0.7) for all time windows and all models when considering subtasks b and c; while, for subtask a, it was lower than 0.7 for two time windows, confirming the apparently higher difficulty of predicting NIV or death vs. the other challenge endpoints. Finally, recall was low for all models across all subtasks.

Table 4 shows the performance metrics of the submitted models for all Task 2 subtasks when using the dataset M6. The structure of the table is the same described above for Table 3. As observed when considering results obtained using dataset M0, the performance of our submitted models was good according to the mean absolute value distance. The event time prediction was, on average, within 6-7 months with respect to the true event time. Specificity was around 0.8 for all time windows and all models when considering subtasks b and c, meanwhile, for subtask a it was still lower than 0.7 for few time windows. Recall was once again low for all models across all subtasks as our models continued to struggle to correctly classify the time window despite the introduction of dynamic variables.

In both settings, when considering the extreme time windows (6-12, 30-36, and >36 months), specificity was higher and recall was lower than the corresponding values obtained in central windows (12-18, 18-24, and 24-30 months), which had a higher number of observations. The submitted models had high specificity, but, as evidenced by low recall across the board, correctly identifying the exact time window was quite the challenging task. Nevertheless, as the values obtained when considering the average of the absolute value of distances between predicted and true time windows were low (6-8 months), the predicted time window was not far from the true one. In fact, on average, it was most frequently the previous or following time window, which might be considered an acceptable margin of error from a clinical perspective.

Table 4
Task 2 metrics, dataset M6

Metric	sub-task a SSVM	 Cox	sub-task b SSVM	 RSF	sub-task c SSVM	 Cox
Abs distance	7.662	7.796	7.358	6.857	7.582	7.683
Specificity (6-12m)	0.955	0.971	0.967	0.964	0.970	0.990
Specificity (12-18m)	0.793	0.659	0.834	0.739	0.893	0.849
Specificity (18-24m)	0.691	0.681	0.784	0.809	0.786	0.733
Specificity (24-30m)	0.802	0.872	0.774	0.883	0.807	0.821
Specificity (30-36m)	0.910	0.962	0.907	0.912	0.846	0.888
Specificity (>36m)	0.988	0.976	0.931	0.905	0.875	0.878
Recall (6-12m)	0.234	0.173	0.186	0.174	0.238	0.119
Recall (12-18m)	0.290	0.454	0.357	0.464	0.246	0.246
Recall (18-24m)	0.437	0.458	0.300	0.283	0.371	0.414
Recall (24-30m)	0.281	0.187	0.407	0.166	0.276	0.292
Recall (30-36m)	0.227	0.090	0.208	0.208	0.135	0.108
Recall (>36m)	0.164	0.164	0.394	0.535	0.401	0.407

5. Conclusions and Future Work

In this work, two different approaches to address the two pilot tasks proposed in the iDPP challenge, i.e., ranking patients based on event risk (Task 1) and predicting the time to event (Task 2) for subjects with ALS, were applied. For each task, three different types of event had to be predicted: NIV or Death (if this occurred before NIV), PEG or Death (if this occurred before NIV) and Death, for subtasks a, b, and c, respectively.

Particularly, three baseline machine learning techniques for survival analysis (namely, the Cox model, the SSVM and the SRF) were applied to carry out each task. The application of these approaches, widely used in the literature for ALS progression prediction as discussed in Section 2, provided a first comparison of baseline models, including both parametric and semi-parametric, linear and non-linear techniques, on the same dataset and on different types of events. Moreover, for each methodological approach and each task, the prediction ability of the available features was investigated via two different version of the dataset: dataset M0, in which only variables available until the baseline visit are used, and dataset M6, in which also dynamic features are considered, up to 6 months after the baseline.

Among the three applied approaches, for each subtask the two best performing ones were selected based on the average C-index obtained on the 100 boostrap validation sets. The SSVM was selected for all the subtasks and dataset versions, confirming its versatility to adapt to different prediction problems on ALS. Specifically, in Task 1 it reached good discrimination ability with a C-index of about 0.7 and AUROC between 0.70 and 0.81 in all the subtasks and time windows, while model calibration was good only for long PHs (BS greater than 0.5 for a PH of 12 months and BS lower than 0.08 with a PH equal to 60 months). In Task 2, the SVM yielded an Abs distance between predicted and true time windows of 7-8 months.

Another notable finding of our analysis was that, for each subtask and model used, performance improved when dynamic features were added as predictors. This is particularly evident looking at Task 1's AUROC and Task 2's Abs distance obtained: the AUROC was always greater than 0.75 (range [0.75-0.82]) when dataset M6 was used, while for dataset M0 it was lower than 0.7 in some cases (range [0.64-0.8]). Similarly, the average Abs distance was in the range [7.4-8.8] with dataset M0, and [6.8-7.7] with dataset M6. This highlights the importance of using the first 6 months of data to more accurately determine future impairments in a fast progressing disease such as ALS, instead of just using data collected at the patients' first visit.

In both Task 1 and Task 2, the models had worse performance in subtask a, compared to subtasks b and c, suggesting a higher difficult in predicting the need for NIV than that of PEG, or only death. This is apparent from Task 1's AUROCs and Task 2's specificities. When considering, e.g., dataset M6, AUROC values were around 0.77 for subtask a (range [0.75-0.79]) and around 0.79 for the other two subtasks (range [0.76-0.82]). Similarly, specificity in subtask a was lower than 0.7 in several time windows in subtask a (range [0.65-0.98]), while around 0.8 in subtasks b and c (range [0.73-0.99]). This might be reasonable being NIV a non-invasive intervention whose timing is based on clinical decision and not only on patient condition. It would be interesting to compare these results with those of other participants to check if the pattern is maintained.

As stated above, this work constitutes a preliminary comparison of baseline approaches for the prediction of ALS progression. Future works will focus on a wider range of methods, including techniques such as gradient boosting and neural networks. Moreover, it could be interesting to expand the pool of features to include -omics data, to explore their possible effect on model performance. Other possibly ameliorative approaches might be based on patient stratification and the subsequent development of group-specific models.

We believe that facilitating the stratification of ALS patient populations and predicting the clinical progression of ALS can serve to improve our understanding of disease mechanisms, enable individual patient level prognosis, and improve the success rate of clinical trials.

References

[1] M. C. Kiernan, S. Vucic, B. C. Cheah, M. R. Turner, A. Eisen, O. Hardiman, J. R. Burrell, M. C. Zoing, Amyotrophic lateral sclerosis, The lancet 377 (2011) 942–955.

[2] L. Zinman, M. Cudkowicz, Emerging targets and treatments in amyotrophic lateral sclerosis, The Lancet Neurology 10 (2011) 481–490.

[3] A. Chiò, A. Calvo, C. Moglia, L. Mazzini, G. Mora, et al., Phenotypic heterogeneity of amyotrophic lateral sclerosis: a population based study, Journal of Neurology, Neurosurgery & Psychiatry (2011) jnnp–2010.

[4] E. Beghi, A. Chiò, P. Couratier, J. Esteban, O. Hardiman, G. Logroscino, A. Millul, D. Mitchell, P.-M. Preux, E. Pupillo, et al., The epidemiology and treatment of ALS: focus on the heterogeneity of the disease and critical appraisal of therapeutic trials, Amyotrophic Lateral Sclerosis 12 (2011) 1–10.

[5] A. Chio, G. Logroscino, O. Hardiman, R. Swingler, D. Mitchell, E. Beghi, B. G. Traynor, E. Consortium, et al., Prognostic factors in als: a critical review, Amyotrophic lateral sclerosis 10 (2009) 310–323.

[6] A. Abraham, V. E. Drory, Influence of serum uric acid levels on prognosis and survival in amyotrophic lateral sclerosis: a meta-analysis, Journal of neurology 261 (2014) 1133–1138.

[7] A. Chiò, A. Calvo, G. Bovio, A. Canosa, D. Bertuzzo, F. Galmozzi, P. Cugnasco, M. Clerico, S. De Mercanti, E. Bersano, et al., Amyotrophic lateral sclerosis outcome measures and the role of albumin and creatinine: a population-based study, JAMA neurology 71 (2014) 1134–1142.

[8] A. Barrón-Cedeño, G. Da San Martino, M. Degli Esposti, F. Sebastiani, C. Macdonald, G. Pasi, A. Hanbury, M. Potthast, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Thirteenth International Conference of the CLEF Association (CLEF 2022), Lecture Notes in Computer Science (LNCS) 13390, Springer, Heidelberg, Germany, 2022.

[9] G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), CLEF 2022 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, 2022.

[10] A. Guazzo, I. Trescato, E. Longato, E. Hazizaj, D. Dosso, G. Faggioli, G. M. Di Nunzio, G. Silvello, M. Vettoretti, E. Tavazzi, C. Roversi, P. Fariselli, S. C. Madeira, M. de Carvalho, M. Gromicho, A. Chiò, U. Manera, A. Dagliati, G. Birolo, H. Aidos, B. Di Camillo, N. Ferro, Intelligent Disease Progression Prediction: Overview of iDPP@CLEF 2022, in: A. Barrón-Cedeño, G. Da San Martino, M. Degli Esposti, F. Sebastiani, C. Macdonald, G. Pasi, A. Hanbury, M. Potthast, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Thirteenth International Conference of the CLEF Association (CLEF 2022), Lecture Notes in Computer Science (LNCS) 13390, Springer, Heidelberg, Germany, 2022.

[11] A. Guazzo, I. Trescato, E. Longato, E. Hazizaj, D. Dosso, G. Faggioli, G. M. Di Nunzio, G. Silvello, M. Vettoretti, E. Tavazzi, C. Roversi, P. Fariselli, S. C. Madeira, M. de Carvalho, M. Gromicho, A. Chiò, U. Manera, A. Dagliati, G. Birolo, H. Aidos, B. Di Camillo, N. Ferro, Overview of iDPP@CLEF 2022: The Intelligent Disease Progression Prediction Challenge, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), CLEF 2022 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, 2022.

[12] E. Tavazzi, S. Daberdaku, A. Zandonà, R. Vasta, B. Nefussy, C. Lunetta, G. Mora, J. Mandrioli, E. Grisan, C. Tarlarini, et al., Predicting functional impairment trajectories in amyotrophic lateral sclerosis: a probabilistic, multifactorial model of disease progression, Journal of Neurology (2022) 1–21.

[13] T. Leão, S. C. Madeira, M. Gromicho, M. de Carvalho, A. M. Carvalho, Learning dynamic bayesian networks from time-dependent and time-independent data: Unraveling disease progression in amyotrophic lateral sclerosis, Journal of Biomedical Informatics 117 (2021) 103730.

[14] A. S. Martins, M. Gromicho, S. Pinto, M. de Carvalho, S. C. Madeira, Learning prognostic models using diseaseprogression patterns: Predicting the need fornon-invasive ventilation in amyotrophic lateralsclerosis, IEEE/ACM Transactions on Computational Biology and Bioinformatics (2021).

[15] A. Ferreira, S. C. Madeira, M. Gromicho, M. d. Carvalho, S. Vinga, A. M. Carvalho, Predictive medicine using interpretable recurrent neural networks, in: International Conference on Pattern Recognition, Springer, 2021, pp. 187–202.

[16] A. V. Carreiro, P. M. Amaral, S. Pinto, P. Tomás, M. de Carvalho, S. C. Madeira, Prognostic

models based on patient snapshots and time windows: Predicting disease progression to assisted ventilation in amyotrophic lateral sclerosis, Journal of biomedical informatics 58 (2015) 133–144.

[17] M.-L. Ong, P. F. Tan, J. D. Holbrook, Predicting functional decline and survival in amyotrophic lateral sclerosis, PloS one 12 (2017) e0174925.

[18] K. D. Ko, T. El-Ghazawi, D. Kim, H. Morizono, Predicting the severity of motor neuron disease progression using electronic health record data with a cloud computing big data approach, in: 2014 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, IEEE, 2014, pp. 1–6.

[19] J. Ackrivo, J. Hansen-Flaschen, E. P. Wileyto, R. J. Schwab, L. Elman, S. M. Kawut, Development of a prognostic model of respiratory insufficiency or death in amyotrophic lateral sclerosis, European Respiratory Journal 53 (2019).

[20] V. Grollemund, G. L. Chat, M.-S. Secchi-Buhour, F. Delbot, J.-F. Pradat-Peyre, P. Bede, P.-F. Pradat, Development and validation of a 1-year survival prognosis estimation model for amyotrophic lateral sclerosis using manifold learning algorithm umap, Scientific reports 10 (2020) 1–12.

[21] S. R. Pfohl, R. B. Kim, G. S. Coan, C. S. Mitchell, Unraveling the complexity of amyotrophic lateral sclerosis survival prediction, Frontiers in neuroinformatics 12 (2018) 36.

[22] R. P. van Eijk, J. N. Bakers, M. A. van Es, M. J. Eijkemans, L. H. van den Berg, Implications of spirometric reference values for amyotrophic lateral sclerosis, Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration 20 (2019) 473–480.

[23] W. J. Scotton, K. M. Scott, D. H. Moore, L. Almedom, L. C. Wijesekera, A. Janssen, C. Nigro, M. Sakel, P. N. Leigh, C. Shaw, et al., Prognostic categories for amyotrophic lateral sclerosis, Amyotrophic Lateral Sclerosis 13 (2012) 502–508.

[24] R. Küffner, N. Zach, R. Norel, J. Hawe, D. Schoenfeld, L. Wang, G. Li, L. Fang, L. Mackey, O. Hardiman, et al., Crowdsourced analysis of clinical trial data to predict amyotrophic lateral sclerosis progression, Nature biotechnology 33 (2015) 51.

[25] R. Kueffner, N. Zach, M. Bronfeld, R. Norel, N. Atassi, V. Balagurusamy, B. Di Camillo, A. Chio, M. Cudkowicz, D. Dillenberger, et al., Stratification of amyotrophic lateral sclerosis patients: a crowdsourcing approach, Scientific reports 9 (2019) 690.

[26] S. Jahandideh, A. A. Taylor, D. Beaulieu, M. Keymer, L. Meng, A. Bian, N. Atassi, J. Andrews, D. L. Ennist, Longitudinal modeling to predict vital capacity in amyotrophic lateral sclerosis, Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration 19 (2018) 294–302.

[27] A. A. Taylor, C. Fournier, M. Polak, L. Wang, N. Zach, M. Keymer, J. D. Glass, D. L. Ennist, P. R. O.-A. A. C. T. Consortium, Predicting disease progression in amyotrophic lateral sclerosis, Annals of clinical and translational neurology 3 (2016) 866–875.

[28] E. Tavazzi, S. Daberdaku, R. Vasta, A. Calvo, A. Chiò, B. Di Camillo, Exploiting mutual information for the imputation of static and dynamic mixed-type clinical data with an adaptive k-nearest neighbours approach, BMC Medical Informatics and Decision Making 20 (2020) 1–23.

[29] C. Moglia, A. Calvo, M. Grassano, A. Canosa, U. Manera, F. D'Ovidio, A. Bombaci, E. Bersano, L. Mazzini, G. Mora, et al., Early weight loss in amyotrophic lateral sclerosis: outcome relevance and clinical correlates in a population-based cohort, Journal of Neurology, Neurosurgery & Psychiatry 90 (2019) 666–673.

[30] T. Shimizu, Y. Nakayama, C. Matsuda, M. Haraguchi, K. Bokuda, K. Ishikawa-Takata, A. Kawata, E. Isozaki, Prognostic significance of body weight variation after diagnosis in als: a single-centre prospective cohort study, Journal of Neurology 266 (2019) 1412–1420.

[31] S. Van Buuren, K. Groothuis-Oudshoorn, mice: Multivariate imputation by chained equations in R, Journal of statistical software 45 (2011) 1–67.

[32] A. Niculescu-Mizil, R. Caruana, Predicting good probabilities with supervised learning, in: Proceedings of the 22nd International Conference on Machine Learning, ICML '05, Association for Computing Machinery, New York, NY, USA, 2005, p. 625–632.

[33] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.

[34] J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization, Journal of Machine Learning Research 13 (2012) 281–305.