From ALS patient stratification towards prognostic using disease progression patterns

Lino Fernandes (n.74192)

Amyotrophic Lateral Sclerosis (ALS) is a progressive disease which results in a rapid degeneration of motor neurons in the brain and the spinal cord, leading to loss of bulbar and limb function. Although being one of the most unrelenting neurodegenerative disorders it is still understudied, especially when compared to others such as Alzheimer's or Parkinson's. Patients usually succumb after three to five years from disease onset, most due to respiratory failure. Since there is no cure available, predicting respiratory insufficiency is fundamental for extending survivability and providing quality of life.

We apply to the Portuguese ALS dataset a state of the art constrained hierarchical clustering to obtain patient independent snapshots that are representative of the his condition at time *i*. Using state of the art Machine Learning models, these snapshots are used to predict a given patients' requirement for Non Invasive Ventilation (NIV) in time windows of 90, 180 and 365 days after time *i*. Later, these results are compared with other predictions obtained by using the complete patient history.

Our results show that using individual snapshots can accurately predict the NIV status of a patient showing an area under the Receiver Operating Characteristic curve of 75%, 77% and 78% for the three time windows. Moreover, when analyzing the complete patient history we found ALS to be much more dependent on the most recent evaluation. We finally divided the population by their disease progression rate, a novel approach that yielded promising results.

I. INTRODUCTION

The ALS research panorama can be divided into two major problems: Patient Diagnosis and Prognostic Prediction. The former is mostly focused on understanding the impact of diagnostic delay[1] and exploring the importance of clinical features that can lead to an earlier, more accurate diagnosis[2]. The latter involves studying the several predictors associated with ALS. There have been a wide range of predictors explored, from functional tests such as ALSFRS to demographic attributes such as the patient's Sex, BMI and Age at Onset[3]. Since most patients perish due to hypoventilation, predictors obtained from respiratory tests such as VC, FVC and MIP/MEP are also of major importance[4]. Neurophysiological studies have also been made to find how Phrenic Nerve[1] irregularities impact ALS[5].

Regarding predictors related to the respiratory function, the application Non-Invasive Ventilation (NIV) has been shown to extend survival rate and improve the patients' quality of life[6]. However, the decision for its application is based on guidelines which rely on clinical observations, respiratory tests and consensus agreement[7]. Recently, a novel strategy was proposed using a constrained hierarchical clustering to stratify ALS patients based on their NIV status[8]. Along with the application of Machine Learning models, this strategy can provide physicians with the probability of a given patient requiring NIV within a chosen time window. This can be further explored by dividing patients according to their disease progression rate[9]. Providing this hindsight to clinicians can have a great impact in their ability to manage the patients' condition, with hopes of improving and extending their quality of life.

In this work, we apply the aforementioned hierarchical clustering technique to create patient snapshots using the Portuguese ALS dataset. These are used to create learning examples in time windows (k) of 90, 180 and 365 days in which the class is the NIV status in a time *k* after the snapshot. To answer the following questions this strategy is applied to the most recent iteration of the of the Portuguese ALS dataset:

1. Given individual patient evaluations, can we predict if a given patient will require NIV within a time window after his last evaluation?
2. Given a set of consecutive patient evaluations, can we predict if a given patient will require NIV within a time window after his last evaluation?
3. Does partitioning the patients into disease progression groups affect the two previous problems?

II. TEST DATA

The dataset used in this context consists of the most recent clinical data from ALS patients followed in the ALS clinic of the Translational Clinical Physiology Unit, Hospital de Santa Maria, Lisbon. The first patients entered the study in 1995 and currently it holds data from 1135 patients followed until March of 2017. It is composed mainly of repeated clinical evaluations of functional, res-

[1] Nerve that passes motor information to the diaphragm.

piratory tests and neurophysiological assessments in the form of a multivariate time series, in a total of 120 different temporal features. In addition to these there are also 77 different static features that include demographics, medical and family history and onset evaluation. More importantly the date the patient started NIV. These add up to almost 200 different attributes which, with the help of clinicians, was narrowed down to 33 (excluding the patient identification). Recently genetic biomarkers from the most commonly associated genes with ALS were introduced in the database, but they were left out since most patients have yet to be tested.

III. CREATING LEARNING EXAMPLES

Snapshots

The Portuguese ALS dataset consists of a mixture of static features with multivariate timeseries composed of a combination of tests the patients were subjected throughout a period of three months. To create the patient snapshots, a hierarchical clustering stratification strategy was used, where two constraints were applied:

1. Two evaluations of the same test cannot belong to the same snapshot as they belong to different batches of tests;
2. The feature of interest (class), NIV in this case, must be coherent in all tests of a snapshot.

Then, according to whether or not the patient had the need for NIV, a single binary feature is computed. In Figure 1 we can see the output of this stratification from the original data. Since the multiple exams that comprise a test batch are usually done in different days, there is not an exact date of occurrence of a snapshot. As such the NIV status is calculated using the median date of the test batch and the NIV date. If the median date is after the NIV date then the feature is attributed a 1, and a 0 else wise.

To answer the questions posed, learning examples were created using time windows of k equals to 90, 180 and 365 days with a new binary class E. Given the median date i of the snapshot (Figure 2), if a patient began NIV between i and $i + k$ then E = 1 (figure 2a). If NIV starts after $i + k$ or the patient never started NIV [2], the snapshots are labeled as E = 0 (figure 2b and 2c). Some snapshots are discarded of one of two conditions meet, if there is no information about NIV after $i + k$ and/or the patient already started NIV at time i (figure 2d and 2e).

Set of Snapshots

Using the individual snapshots built before, we created a representation of the patients' history by extracting all patients with at least T snapshots available and concatenated them. In this setting, the class will be the binary NIV status given by E, k days after the *T-th* snapshot. Doing so allows us to use the complete set of records of a given patient up until time i. Unlike the Single Snapshot case, here we have several datasets where each patient is represented by one instance with the all the records from the first T snapshots used as features (*FirstT* dataset in figure 4). Parallelly we created other data sets in which the earliest snapshots were discarded and only the last one was used (*Last* dataset in figure 4). In summary, in the former case we are using the first T snapshots to predict the NIV status k days after the *T-th* snapshot while in the latter we use only the *T-th* snapshot to make the same prediction.

We have limited the number of snapshots T to six, as that grants at least 100 training instances for the training scheme. In total, ten datasets were created for every value of k: each *FirstT* and *Last* data set has up to five consecutive snapshots. Note that in the case of the dataset with only one snapshot *FirstT* and Last are the same. Although this dataset provides us no useful information about the problem at hand in this section, it will be useful for further discussion.

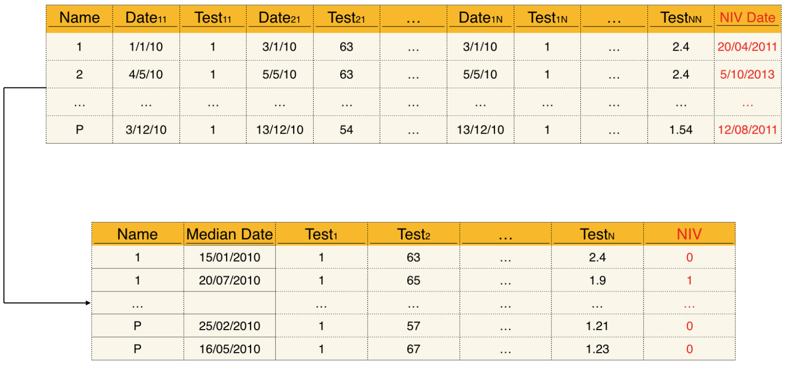
Progression Groups

By measuring the change in ALSFRS over time, ΔALSFRS, we can have an estimation of how the disease is progressing and infer about the survivability of the patient[10]. A study has found that computing this feature using the date of onset improves the predictive power of future progression rate[9]. The change in ALSFRS can then be computed using equation 1, where $ALSFRS_{1stVisit}$ is the ALSFRS score of a given patient at the beginning of the trial and ΔT is the time in months since the first symptoms appeared until the first visit.

$$\Delta ALSFRS = \frac{40 - ALSFRS_{1stVisit}}{\Delta T} \tag{1}$$

Alternatively we can use a different functional test ALSFRS-R as shown in equation 2. This is a revised

[2] As long as the patient has at least one snapshot after $i + k$.



Name	Date$_{11}$	Test$_{11}$	Date$_{21}$	Test$_{21}$	…	Date$_{1N}$	Test$_{1N}$	…	Test$_{NN}$	NIV Date
1	1/1/10	1	3/1/10	63	…	3/1/10	1	…	2.4	20/04/2011
2	4/5/10	1	5/5/10	63	…	5/5/10	1	…	2.4	5/10/2013
…	…	…	…	…	…	…	…	…	…	…
P	3/12/10	1	13/12/10	54	…	13/12/10	1	…	1.54	12/08/2011

Name	Median Date	Test$_1$	Test$_2$	…	Test$_N$	NIV
1	15/01/2010	1	63	…	2.4	0
1	20/07/2010	1	65	…	1.9	1
…		…	…	…	…	…
P	25/02/2010	1	57	…	1.21	0
P	16/05/2010	1	67	…	1.23	0

FIG. 1: Transformation of the original data into patient snapshots (adapted from AV Carreiro *et al.* [8]).

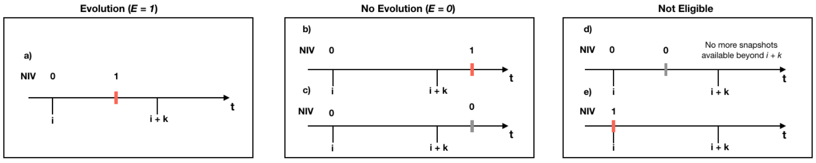


FIG. 2: Binary class *Evolution* representing the patient NIV status $k + i$ days after the snapshot date i (adapted from AV Carreiro *et al.*[8]).

version of ALSFRS which includes two extra functional questions.

$$\Delta ALSFRS - R = \frac{48 - ALSFRSR_{1stVisit}}{\Delta T} \tag{2}$$

To stratify the patients accordingly, we computed the ΔALSFRS and analyzed its distribution. Then, if the ΔALSFRS distribution was close to a Normal distribution we would claim that the slow and fast progressors were contained outside the 1-sigma region, in the first and second tail respectively. However, the actual distribution did not follow a Gaussian. The solution suggested by the experts was to take the patients beyond the third quartile (Q_3) as fast progressors and the slow progressors those until the first quartile (Q_1). The patients encompassed by the first and second quartile were labeled as neutral, *i.e.* patients with regular progression.

$$ALSFRS = \begin{cases} Q_1 = 0.28 \\ Q_2 = 0.57 \\ Q_3 = 1.04 \end{cases} \tag{3}$$

Out of the 1135 patients recorded in the Portuguese ALS dataset, 152 either have not been subjected to ALSFRS evaluations or have the ΔT information unavailable. Were we to use the the criteria based on the revised ALSFRS, equation 2, only 894 patients would be available for analysis.

Classifier	**Parameter**	**Range**
NB	Kernel	{True,False}
SVM	Polynomial Degree	{1,2,3}
RF	Number of Trees	{5,10,15,20}

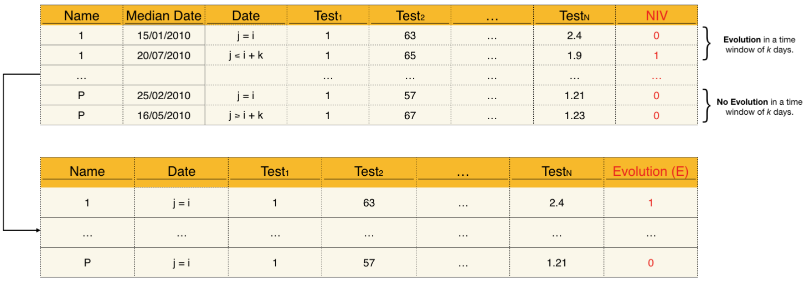
TABLE I: Parameters tested for each classifier.

IV. METHODOLOGY

The datasets were used in a 5x10-fold CV scheme using 100% of the created datasets to train three models imported to Python from WEKA[11] using the Python WEKA Wrapper[12]: Naive Bayes (NB), Random Forests (RF) and Support Vector Machines (SVM).

For each problem, several data preprocessing techniques were applied and those which maximized the Area Under the ROC curve (AUC) were chosen and presented in section V: Missing Value Imputation (MVI), Correlation based Feature Selection (FS), One Hot Encoding (OHE) and SMOTE.

For each problem, the models' parameters were also optimized for maximizing the AUC. The parameters tuned and the range of values tested are shown in table I

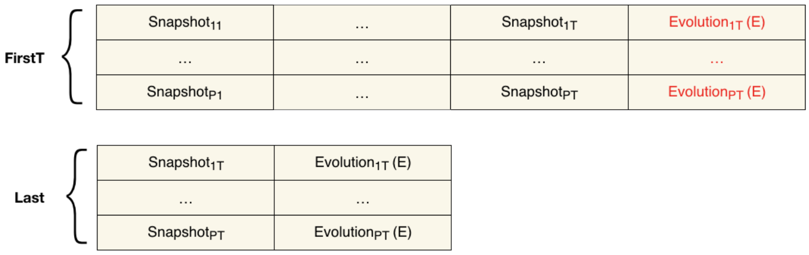


Name	Median Date	Date	$Test_1$	$Test_2$	...	$Test_N$	NIV
1	15/01/2010	j = i	1	63	...	2.4	0
1	20/07/2010	j ⩽ i + k	1	65	...	1.9	1
...			...	...	...	...	...
P	25/02/2010	j = i	1	57	...	1.21	0
P	16/05/2010	j ⩾ i + k	1	67	...	1.23	0

Evolution in a time window of *k* days.
No Evolution in a time window of *k* days.

Name	Date	$Test_1$	$Test_2$	...	$Test_N$	Evolution (E)
1	j = i	1	63	...	2.4	1
...	...	...	...	...	...	...
P	j = i	1	57	...	1.21	0

FIG. 3: Creating the learning examples used for prognostic according to the rules presented in figure 2 (adapted from AV Carreiro *et al.*[8]).



$Snapshot_{11}$	...	$Snapshot_{1T}$	$Evolution_{1T}$ (E)
...	...	...	...
$Snapshot_{P1}$	...	$Snapshot_{PT}$	$Evolution_{PT}$ (E)

$Snapshot_{1T}$	$Evolution_{1T}$ (E)
...	...
$Snapshot_{PT}$	$Evolution_{PT}$ (E)

FIG. 4: Representation of the two datasets used in this section: *FirstT* and *Last*.

V. RESULTS AND DISCUSSION

A. Single Snapshot Prediction

Table II shows a summary of the raw dataset along with information about the class distribution. We see that in longer windows there are less snapshots available which is explained by the rules defined for the stratification strategy used (see section III). Remember that if there is no information regarding the NIV status k days after the time i of the snapshot, the snapshot is discarded. When using longer time windows this situation happens more often and so less snapshots are eligible. Naturally, by increasing the window we are evaluating the NIV status of the patient in a longer interval, so it is more likely that the patient starts NIV during it. This explains why the distribution of the class is more balanced for the larger windows.

Window	Snapshots	E=1	E=0
90	2357	475 (20.15%)	1882 (70.75%)
180	2241	743 (33.15%)	1498 (66.85%)
365	2042	1086 (53.18%)	956 (46.82%)

TABLE II: Class distribution for k ={90,180,365} days.

In table III the results for the best performing model are shown. The model performs better when performing predictions in longer time windows, as there is a steady increase in all metrics with k. A possible explanation is the class imbalance in smaller windows, despite SMOTE having been applied to all datasets. Using NB seems to be significantly better at predicting NIV requirements than the alternative proposed models. It is also clinically useful as it can provide a probability value along with the prediction which may be insightful of the the patient's condition: Think of two patients, one with 51% chance of needing NIV and other and 99%, both will be positively classified but one is probably in a more aggravated condition than the other.

To evaluate these results we make a comparative analysis to a similar case using an older version of the Portuguese ALS dataset[13]. This dataset contained records 758 patients, a 377 difference to the data used in this dissertation. Independent patient snapshots were created using the same constrained hierarchical clustering approach which were then transformed into learning examples, as explained in section **??**. Regarding the preprocessing techniques, AV Carreiro tested MVI and two different FS techniques (*sucRem* and *mRMR*). Given that we have more available data, it would be expectable that our results were an improved update to his work.

However, that was not the always the case. When comparing the best performing models in both situations (NB with MVI), his performs better in the 90 days window. For a 6 month prediction (180 days) both models perform similarly, and in the largest window ours is scoring a higher AUC. Regarding the other metrics, our predictive models are weaker at correctly capturing the negative class E = *0*, but classify the positive class E = *1* better.

The results obtained during this stage are on par with other published results and validate that using patient independent snapshots and time windows as a prognostic approach to assess NIV requirements.

Window	AUC	Precision	Recall	Specificity
90	75.7±1.67	48.28±6.49	75.11±3.1	90.32±1.69
180	77.41±2.81	60.83±5.78	59.77±7.71	79.53±5.53
365	79.53±1.9	75.11±3.1	74.44±3.29	70.82±6.53

TABLE III: Single Snapshot Prediction results for three different windows (90, 180 and 365 days). These results were obtained in a 5x10-FoldCV using the NB model with Kernel with 50% of SMOTE instances being added.

B. Set of Snapshots[3]

Where as in the previous section we were trying to predict whether a patient will require NIV after k days after time i, here we aim to use a set of consecutive patient evaluations to predict the same endpoint k days after the last observation. Furthermore we can test how the several exams the patient is subjected throughout the clinical trial actually reflect the disease progression along time, by consecutively removing the earlier snapshots and analysing how it affects the performance of the classifier. By creating the sets of snapshots we are increasing the number of features by a significant factor: the number of snapshots (T) times the number of features (33). For higher T the number of predictors significantly increases to a point where we can have more features than samples. This high dimensionality severely hampers the predictive power of the models[14]. Feature Selection by means of CFS was then tested and proved to be effective in improving the results. Another issue is, as we can see from Tables IV and Table II, creating these sets led to a significant depletion in instances. By using an oversampling technique, SMOTE, we can not only balance the classes but create synthetic instances that can be helpful for the model prediction.

The results obtained in this section show differences in how using a more extensive patient history, *i.e.* more snapshots T, affects the classification. Generally, using a higher number of snapshots results in better predictive performance for all models. Note that by higher T we mean not only using more consecutive snapshots ($FirstT$ datasets) but also using only the last snapshot of $FirstT$ (*Last* datasets). Regarding these, the models perform almost identically between the two but the AUC scores are slightly higher when using only the last snapshot. This indicates that the most recent information about the patient is more relevant to predict his NIV requirements. Despite this, it should be remarked that even with FS the $FirstT$ datasets have a worse ratio of features to instances than *Last*. This could negatively impact the predictive performance in the $FirstT$ scenario, especially for higher Ts which not only have less instances but also more features.

As done previously, we compared our results with those detailed in AV Carreiro PhD thesis[13]. Unfortunately, the version of the dataset with 758 patients used for his analysis of the single snapshot prediction was not applied in this setting. Instead, it was done using an earlier version containing 517 patients until 2011 using the same methodology applied in this section: concatenating the T individual snapshots. Regarding data preprocessing techniques, *sucRem* and SMOTE were used simultaneously. The difference between the best models (NB) is parallel to what was observed previously. While his performs better in the 90 days window ours does it for the larger one (365 days). As for predictions in 180 days, are again balanced in his favor. This strategy shows some drawbacks when compared to the single snapshot case to be used for NIV prognosis. For once, the building this dataset requires removing instances to be used as attributes which increases the dimensionality of the problem while reducing the amount of instances at the same time. A possible advantage is suggested by the of models trained with only the *T-th* snapshot which perform better than when using T consecutive snapshots, sometimes even better than the single snapshot setting.

Window	T	#Patients	E=0	E=1
90	1	654	501 (76.61%)	153 (23.39%)
	2	461	357 (77.44%)	104 (22.56%)
	3	336	272 (80.95%)	64 (19.05%)
	4	252	204 (80.95%)	48 (19.05%)
	5	190	154 (81.05%)	36 (18.95%)
	6	134	111 (82.84%)	23 (17.16%)
180	1	630	393 (62.38%)	237 (37.62%)
	2	447	290 (64.88%)	157 (35.12%)
	3	319	217 (68.03%)	102 (31.97%)
	4	230	155 (67.39%)	75 (32.61%)
	5	179	121 (67.6%)	58 (32.4%)
	6	122	87 (71.31%)	35 (28.69%)
365	1	586	260 (44.37%)	326 (55.63%)
	2	408	180 (44.12%)	228 (55.88%)
	3	289	135 (46.71%)	154 (53.29%)
	4	210	102 (48.57%)	108 (51.43%)
	5	160	80 (50.0%)	80 (50.0%)
	6	109	50 (45.87%)	59 (54.13%)

TABLE IV: Class distribution of all sets of snapshots with more than 100 points for k ={90,180,365} days.

[3] The tables detailing the results can be seen at section 3.2.3.2 of the complete dissertation.

C. Progression Groups

Single Snapshot Prediction

Below, in table V, we show a summary of the datasets used in this section along with that of the full dataset. We are losing approximately 200 snapshots per window, but the class distribution is maintained for most datasets. A good indicator that the stratification is meaningful is the percentage of those who have started NIV in the following k days, given by *E=1*. Analyzing this, the proportion of *E=1* in the Slow dataset is always lower than others. It would be expectable to see the opposite behavior for the Fast dataset, but there is little difference to the Neutral dataset, except in the 180 days window.

The training results of the best performing model are shown in Table VII. Comparing these with those obtained without progression groups, we observe a small decline in performance but overall the models are performing almost identically to those previously discussed. However, the confidence intervals given by the standard deviation are much wider which is probably due to the smaller amount of data available. Despite this, the *Slow* and *Fast* models are predicting the positive class much more accurately than the *Neutral* model and those obtained in the previous section.

To evaluate if it is worth having specific models to classify the patients' requirement of NIV in a time window after their current condition given their ALS progression rate, we have tested the datasets created for this chapter against the model shown in table III. Should these results be worse than those obtained in VII then having specific models for each progression group is beneficial. To do so, we trained a NB model al Model (GM))(Gener with the optimized parameters found in that section: Kernel, SMOTE with PSI of 50% and data preprocessing using MVI

In Table VI we have disclosed the AUC, Precision, Recall and Specificity for the three datasets across all time windows. Comparing with Table VII we find that the GM is performing better at predicting the NIV. However, the proportion of the positive class correctly predicted (Recall) by this model is lower than when using the specific models. Although there may not be a need to use specific models, these results do not invalidate the the proposed stratification method.

Window	Dataset	Snapshots	E=1	E=0
90	Slow	528	99 (18.75%)	429 (81.25%)
	Neutral	1133	241 (21.27%)	892 (78.73%)
	Fast	510	105 (20.59%)	405 (79.41%)
180	Slow	505	156 (30.89%)	349 (69.11%)
	Neutral	1078	362 (33.58%)	716 (66.42%)
	Fast	482	170 (35.27%)	312 (64.73%)
365	Slow	480	240 (50.0%)	240 (50.0%)
	Neutral	975	539 (55.28%)	436 (44.72%)
	Fast	426	230 (53.99%)	196 (46.01%)

TABLE V: Class distribution for k ={90,180,365} days when grouped by progression rate. The total number of snapshots for each window is: 90d - **2171**; 180d - **2065**; 365d - **1881**.

Window	Progression	AUC	Precision	Recall	Specificity
90	Slow	78.13	56.25	27.27	95.1
	Neutral	80.87	58.62	35.27	93.27
	Fast	82.12	60.81	42.86	92.84
180	Slow	79.76	66.97	46.79	89.68
	Neutral	82.6	64.24	58.56	83.52
	Fast	81.3	70.39	62.94	85.58
365	Slow	80.77	83.56	50.83	90.0
	Neutral	82.07	86.01	54.73	88.99
	Fast	83.27	82.32	58.7	85.2

TABLE VI: Performance of the NB model with the parameters and preprocessing techniques applied in table III when evaluating patients stratified in progression groups.

Set of Snapshots[4]

Finally, we are assessing how stratifying the patients in progression groups influences the strategy tested using the set of snapshots. In table VIII we have a summary of the datasets in question, for the number of snapshots T that ensures, as close as possible, 100 training instances.

Although we have few instances some remarks can be made when comparing to the class distribution of the dataset without the stratification into progression groups (table IV). When analyzing the positive class throughout the windows we see an increasing differentiation in its proportion in the different groups. When k equals 90 the different progression groups have the about the same distribution, but when k increases we see the percentage of patients who have started NIV is higher in the Fast dataset that in the Slow one, while those in the Neutral dataset lie in between. This is exactly the type of differentiating behavior equation 1 is proposing to model.

Regarding the training scheme, despite having used 5x10-FoldCV so far we have opted to use 5x5-FoldCV instead. Since we have less data, doing so will create training sets

[4] The tables detailing the results can be seen at section 4.3.2.1 of the complete dissertation.

Window	Progression	AUC	Precision	Recall	Specificity
90	Slow	71.56 ± 7.83	36.94 ± 11.01	61.35 ± 14.32	66.24 ± 18.84
	Neutral	75.07 ± 3.84	45.78 ± 8.23	42.91 ± 18.67	83.84 ± 9.12
	Fast	73.66 ± 6.67	40.82 ± 14.28	53.04 ± 14.55	74.29 ± 14.59
180	Slow	69.84 ± 5.01	44.19 ± 5.88	79.04 ± 7.97	49.63 ± 14.99
	Neutral	76.48 ± 2.73	56.76 ± 5.75	63.28 ± 11.34	70.11 ± 11.54
	Fast	73.15 ± 5.55	51.27 ± 6.49	76.23 ± 9.76	53.96 ± 16.2
365	Slow	72.89 ± 9.8	60.27 ± 5.94	87.92 ± 7.76	37.22 ± 15.66
	Neutral	77.33 ± 3.08	63.56 ± 5.84	74.28 ± 9.8	75.29 ± 9.63
	Fast	74.98 ± 6.42	72.63 ± 6.19	71.86 ± 6.83	64.75 ± 9.12

TABLE VII: Single Snapshot Prediction results using Progression Groups for three different windows (90, 180 and 365 days). These results were obtained in a 5x10-FoldCV using the NB model with Kernel with 50% of SMOTE instances being added on the 365 days window and 150% in the others. Regarding data preprocessing, a MVI filter using the mean/mode of the column was applied.

with higher counts of instances in each fold and provide more accurate results.

Window	Dataset	T	Patients	E=0	E=1
90	Slow	1	137	109 (79.56%)	28 (20.44%)
		2	94	69 (73.4%)	25 (26.6%)
	Neutral	1	319	240 (75.24%)	79 (24.76%)
		2	230	180 (78.26%)	50 (21.74%)
	Fast	1	144	108 (75.0%)	36 (25.0%)
		2	99	79 (79.8%)	20 (20.2%)
180	Slow	1	128	82 (64.06%)	46 (35.94%)
		2	90	58 (64.44%)	32 (35.56%)
	Neutral	1	309	194 (62.78%)	115 (37.22%)
		2	223	147 (65.92%)	76 (34.08%)
	Fast	1	140	82 (58.57%)	58 (41.43%)
		2	97	60 (61.86%)	37 (38.14%)
365	Slow	1	123	60 (48.78%)	63 (51.22%)
		2	88	44 (50.0%)	44 (50.0%)
	Neutral	1	285	124 (43.51%)	161 (56.49%)
		2	201	85 (42.29%)	116 (57.71%)
	Fast	1	128	48 (37.5%)	80 (62.5%)
		2	85	34 (40.0%)	51 (60.0%)

TABLE VIII: Class distribution for k ={90,180,365} days.

In order to estimate the quality of the following results we make a parallel with those obtained in the training phase of Section V B, specifically with the FirstT dataset since our Progression Groups and FirstT are essentially the same dataset with a different stratification strategy. The results show that there is no real trend as the results are varying considerably. It is clear that the predictions of the model trained with the *Neutral* patients is predicting better than when using the full dataset. However, the same cannot be said about the *Slow* and *Fast* models which are much more inconsistent. This could be caused by the small amount of training instances available for the second snapshot, given that the results when using the first are usually very similar. Another drawback here seems to be in these models' ability to predict $E = 1$, which is usually lower when using progression groups.

We then followed the same strategy used in the Single Snapshot case to evaluate the validity of the proposed strategy. That is, testing the best model (GM) obtained in section V B with the datasets used to create progression specific models. Then if the results succeeding from this evaluation are better than those obtained with the specific models, there is no benefit in using models targeting each progression group. To do so, we trained the a NB model optimized following the parameters that maximized the AUC results in section V B. As a training set we used the full datasets that represented the patient history, up to two consecutive snapshots.

In table IX we show the AUC, Precision, Recall and Specificity for the three datasets across the T snapshots tested. We found that, unlike the Single Snapshot case, this GM is not unequivocally better than the specific models. Specifically, the *Slow* and *Neutral* model perform better in some settings than the GM (highlighted below). Given this, and taking into account that the results obtained in table **??** had significantly less data, it is possible that using specialized models where patients are grouped by their progression at different rates is superior to using a more generalized approach.

Window	T	Progression	AUC	Precision	Recall	Specificity
90	1	Slow	**80.28**	41.67	71.43	74.31
		Neutral	**78.05**	46.15	60.76	76.67
		Fast	88.81	57.14	88.89	77.78
	2	Slow	**23.54**	53.85	20.29	52.0
		Neutral	78.62	47.69	62.0	81.11
		Fast	72.22	40.0	50.0	81.01
180	1	Slow	80.91	56.9	71.74	69.51
		Neutral	77.92	57.82	73.91	68.04
		Fast	80.09	61.43	74.14	67.07
	2	Slow	**23.17**	50.0	36.21	34.38
		Neutral	82.12	50.79	84.21	57.82
		Fast	75.36	51.85	75.68	56.67
365	1	Slow	78.86	73.47	57.14	78.33
		Neutral	**22.82**	22.22	24.19	34.78
		Fast	81.72	85.29	72.5	79.17
	2	Slow	84.4	82.93	77.27	84.09
		Neutral	77.72	74.42	82.76	61.18
		Fast	87.43	85.42	80.39	79.41

TABLE IX: Performance of the NB model with Kernel trained with the *FirstT* dataset when evaluating patients stratified in progression groups. The model was trained with FS and SMOTE with 50% more instances added to the 365 days window and 150% to 90 and 180 days.

VI. CONCLUSIONS

ALS is a progressive disorder defined by a rapid degeneration of motor neurons in the brain and the spinal cord. It is a devastating condition that quickly spreads, leading to severe muscular deterioration and causing death due to respiratory paralysis in 3 to 5 years from disease onset. Since there is no cure available, the focus is in symptom control and prolonging the quality of the patient, to which the early administration of Non Invasive Ventilation (NIV) has been associated with. Recently a new calculated measurement of the patient rate of progression has been suggested, as a function of one of the most widely used functional tests (ALSFRS). Therefore, this dissertation tackled two problems: The first is the application of a novel patient stratification technique using constrained hierarchical clustering to the most recent iteration of the Portuguese ALS dataset (2017) and from it create learning examples from patient snapshots based on time windows. These learning examples are used for prognosis of the patients' NIV requirement using state of the art Machine Learning models. The second is partitioning the patients according to their disease progression rate and examine how it affects the prognosis models built before.

The first proposed problem can be divided into two subproblems: using a single snapshot and using a set of consecutive snapshots to make the prognosis. Using learning examples in time windows of 90, 180 and 365 days we built a preprocessing pipeline to train and tune the predictions of three supervised models (NB, SVM and RF). The preprocessing steps usually involved missing value imputation, feature selection and oversampling which were evaluated in a 5x10-Fold Cross Validation setting. The overall results for the first problem show that this data benefits of MVI, having achieved an AUC score of 75.89, 77.2 and 78.41 for windows of 90, 180 and 365 days, respectively, for the NB model. Following this, by concatenating the individual T snapshots we created a representation of the patient history, in which each patient is now defined by a single instance. In this setting we used the same pipeline to evaluate if the complete patient account has more predictive power than the patients' most recent snapshot, yielding two datasets: *FirstT* and *Last*. Since this concatenation significantly increased the dimensionality of the problem, using FS proved to help the predictive ability of the models. Unfortunately, the results from this section were inferior to those obtained in single snapshot prediction although we found that, in most cases, the most recent snapshot was more informative since the *Last* dataset scored higher AUC scores.

Finally, we have used the ALSFRS progression rate criteria to divide the patients' snapshots into three different groups: *Slow*, *Neutral* and *Fast*. By examining the class distribution of each window, it was noticeable that the criteria used was meaningful as the proportion of those who required NIV for larger windows, was higher or lower in the *Slow* or *Fast* datasets, respectively. Although we results were acceptable, they were still not as good as those obtained using the full dataset.

In conclusion from all the procedures tested, using individual snapshots to predict the NIV requirements in a time frame still outperforms the other applied strategies. However, it is known that patients progress at different rates and the proposed strategy showed to be clinically meaningful. The downside of the patient partition is that it could be, in itself, limiting. There was a noticeable lack of training instances, specially in the set of snapshots context, so new strategies could address this in the future. Since it was shown that the class distributions changes with the different progression groups, one possible solution would be to include the $\Delta ALSFRS$ result as an extra feature of the dataset. More generally, the disadvantage of these approaches is that we are using the snapshots independently, even though several belong to the same patients. Taking into account the temporal dependencies between them could be advantageous, although it has been shown that it might not be crucial[13]. If anything, the extensive analysis made in this dissertation proves that predicting the NIV in advance is a great and important challenge, and can have a major impact in prolonging and improving the quality of life of those who suffer from ALS.

[1] Martin R Turner, Jakub Scaber, John A Goodfellow, Melanie E Lord, Rachael Marsden, and Kevin Talbot. The diagnostic pathway and prognosis in bulbar-onset amyotrophic lateral sclerosis. *J Neurol Sci*, 294(1-2):81–85, Jul 2010.

[2] Mamede de Carvalho, Reinhard Dengler, Andrew Eisen, John D England, Ryuji Kaji, Jun Kimura, Kerry Mills, Hiroshi Mitsumoto, Hiroyuki Nodera, Jeremy Shefner, and Michael Swash. Electrodiagnostic criteria for diagnosis of ALS. *Clin Neurophysiol*, 119(3):497–503, Mar 2008.

[3] J M Cedarbaum, N Stambler, E Malta, C Fuller, D Hilt, B Thurmond, and A Nakanishi. The ALSFRS-R: a revised ALS functional rating scale that incorporates assessments of respiratory function. BDNF ALS Study Group (Phase III). *J Neurol Sci*, 169(1-2):13–21, Oct 1999.

[4] Fusun Baumann, Robert D Henderson, Stephen C Morrison, Michael Brown, N Hutchinson, James A Douglas, Peter J Robinson, and Pamela A McCombe. Use of respiratory function tests to predict survival in amyotrophic lateral sclerosis. *Amyotroph Lateral Scler*, 11(1-2):194–

202, 2010.

[5] Susana Pinto, Antonia Turkman, Anabela Pinto, Michael Swash, and Mamede de Carvalho. Predicting respiratory insufficiency in amyotrophic lateral sclerosis: the role of phrenic nerve studies. *Clin Neurophysiol*, 120(5):941–946, May 2009.

[6] Stephen C Bourke, Mark Tomlinson, Tim L Williams, Robert E Bullock, Pamela J Shaw, and G John Gibson. Effects of non-invasive ventilation on survival and quality of life in patients with amyotrophic lateral sclerosis: a randomised controlled trial. *Lancet Neurol*, 5(2):140–147, feb 2006.

[7] The EFNS Task Force on Diagnosis, Management of Amyotrophic Lateral Sclerosis:, Peter M. Andersen, Sharon Abrahams, Gian D. Borasio, Mamede de Carvalho, Adriano Chio, Philip Van Damme, Orla Hardiman, Katja Kollewe, Karen E. Morrison, Susanne Petri, Pierre-Francois Pradat, Vincenzo Silani, Barbara Tomik, Maria Wasner, and Markus Weber. EFNS guidelines on the Clinical Management of Amyotrophic Lateral Sclerosis (MALS) – revised report of an EFNS task force. *European Journal of Neurology*, 19(3):360–375, 2012.

[8] André V Carreiro, Pedro M T Amaral, Susana Pinto, Pedro Tomás, Mamede de Carvalho, and Sara 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(Supplement C):133–144, 2015.

[9] Malcolm Proudfoot, Ashley Jones, Kevin Talbot, Ammar Al-Chalabi, and Martin R Turner. The ALSFRS as an outcome measure in therapeutic trials and its relationship to symptom onset. *Amyotroph Lateral Scler Frontotemporal Degener*, 17(5-6):414–425, Jul-Aug 2016.

[10] F Kimura, C Fujimura, S Ishida, H Nakajima, D Furutama, H Uehara, K Shinoda, M Sugino, and T Hanafusa. Progression rate of ALSFRS-R at time of diagnosis predicts survival time in ALS. *Neurology*, 66(2):265–267, Jan 2006.

[11] Eibe Frank, Mark A Hall, and Ian H Witten. The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques". 2016.

[12] Peter Reutemann. *Python Weka Wrapper Documentation*, 2017.

[13] André Valério Raposo Carreiro. *An integrative mining approach for prognostic prediction in neurodegenerative diseases.* PhD thesis, Instituto Superior Técnico - Universidade de Lisboa, 2016.

[14] G Hughes. On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory*, 14(1):55–63, jan 1968.