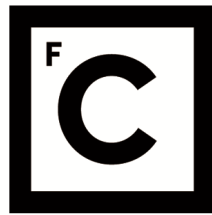


UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE BIOLOGIA VEGETAL



**Ciências
ULisboa**

**A Supervised Learning Approach for Prognostic
Prediction in ALS using Disease Progression
Groups and Patient Profiles**

Sofia Isabel Ferro Pires

Mestrado em Bioinformática e Biologia Computacional
Especialização em Bioinformática

Dissertação orientada por:
Prof^ª. Doutora Sara Alexandra Cordeiro Madeira

Resumo

A Esclerose Lateral Amiotrófica (ELA) é uma Doença Neurodegenerativa caracterizada pela perda progressiva de neurónios motores, que causam inervação e comprometimento muscular. Pacientes que sofrem de ELA não têm geralmente um prognóstico promissor, morrendo entre de 3 a 5 anos após o início da doença. A causa mais comum de morte é a insuficiência respiratória. Não havendo uma cura para a ELA, muitos esforços estão concentrados na elaboração de melhores tratamentos para prevenir a progressão da doença. Tem sido comprovado que a Ventilação Não Invásiva (VNI) melhora o prognóstico quando administrado atempadamente. Esta dissertação propõe abordagens de aprendizagem automática para criar modelos capazes de prever a necessidade de VNI em pacientes com ELA dentro de um intervalo de tempo de k dias, possibilitando assim aos médicos antecipar a prescrição de VNI. No entanto, a heterogeneidade da doença apresenta um desafio para encontrar tratamentos e soluções que possam ser utilizadas para todos os pacientes. Com isso em mente, propomos duas abordagens de estratificação de pacientes, com o objetivo de criar modelos especializados que possam prever melhor a necessidade de VNI para cada um dos grupos criados. A primeira abordagem consiste em criar grupos com base na taxa de progressão do paciente, e a segunda consiste em criar perfis de pacientes agrupando avaliações de pacientes mais semelhantes usando métodos de agrupamento e perfis clínicos baseados em subconjuntos de características (Geral, Prognóstico, Respiratório e Funcional). Também testamos um conjunto de seleção de atributos, para avaliar o valor preditivo dos mesmos, bem como uma abordagem de imputação de valores ausentes para lidar com a alta proporção dos mesmos, característica comum para dados clínicos. Os modelos prognósticos propostos mostraram ser uma boa solução para a previsão da necessidade do uso de NIV, apresentando resultados geralmente promissores. Além disso, mostramos que o uso de estratificação de pacientes para criar modelos especializados, melhorando assim o desempenho dos modelos prognósticos, pode contribuir para um acompanhamento mais personalizado de acordo com as necessidades de cada paciente, melhorando assim o seu prognóstico e qualidade de vida.

Palavras Chave: Esclerose Lateral Amiotrófica, Aprendizagem Supervisionada, Estratificação de Pacientes, Grupos de Progressão da Doença, Perfis de Pacientes, Ventilação Não-Invásiva

Abstract

Amyotrophic Lateral Sclerosis (ALS) is a Neurodegenerative Disease characterized by the progressive loss of motor neurons, which cause muscular innervation and impairment. Patients who suffer from ALS usually do not have a promising prognosis, dying within 3-5 years from the disease onset. The most common cause of death is respiratory failure. With the lack of a cure for ALS, many efforts are focused in designing better treatments to prevent disease progression. Non-Invasive Ventilation (NIV) has been proven to improve prognosis when administered earlier on. This dissertation proposes machine learning approaches to create learning models capable to predict the need for NIV in ALS patients within a time window of k days, enabling clinicians to anticipate NIV prescription beforehand. However, the heterogeneity of the disease presents as a challenge to find treatments and solutions that can be used for all patients. With that in mind, we proposed two patient stratification approaches, with the aim of creating specialized models that can better predict the need for NIV for each of the created groups. The first approach consists in creating groups based on the patient's progression rate, and the second approach consists in creating patient profiles by grouping patient evaluations that are more similar using clustering and clinical profiles based on subset of features (General, Prognostic, Respiratory, and Functional). We also tested a feature selection ensemble, to evaluate the predictive value of the features, as well as a Missing value imputation approach to deal with the high proportion of missing values, common characteristic for clinical data. The proposed prognostic models showed to be a good solution for prognostic prediction of NIV outcome, presenting overall promising results. Furthermore, we show that the use of patient stratification to create specialized models, thus improving performance in prognostic models that can contribute to a better-personalized care according to each patient needs, thus improving their prognostic and quality of life.

Keywords: Amyotrophic Lateral Sclerosis, Supervised Learning, Patient Stratification, Disease Progression Groups, Patient Profiles, Non-Invasive Ventilation

Resumo Alargado

A Esclerose Lateral Amiotrófica (ELA) é uma doença neurodegenerativa caracterizada pela morte dos neurónios motores que controlam os movimentos voluntários. Isto leva à perda progressiva de movimento nos pacientes, sendo que os primeiros sintomas são geralmente a falta de força nos membros inferiores ou superiores. Em relação a outras doenças neurodegenerativas, a progressão da ELA é geralmente mais rápida, resultando na morte dos pacientes num período entre 3 a 5 anos. É uma doença que se manifesta essencialmente em idades mais avançadas (58 a 65 anos), no entanto pacientes com histórico familiar de ELA têm um aparecimento da doença mais precoce (43-63 anos).

Dado não existir uma cura conhecida para ELA, o Riluzole é o único fármaco disponível para controlar a progressão da doença. Assim, o acompanhamento médico desta doença é geralmente baseado em tratamentos que aliviam os sintomas e tentam retardar a progressão da doença, de forma a melhorar a qualidade de vida dos pacientes e o seu prognóstico.

A causa mais comum de morte em ELA é a paragem respiratória, e por isso um dos tratamentos mais comuns em ELA é a Ventilação Não Invásiva (VNI), de forma a controlar os sintomas associados com a perda de função respiratória e evitar complicações. O uso desta terapêutica é no entanto mais eficaz quando colocado em estágios iniciais da doença. Quando aplicado atempadamente, o uso de VNI pode prolongar a sobrevivência dos pacientes de ELA em alguns casos por mais de um ano.

Tendo isto em conta, definimos então o principal objetivo desta dissertação: Criar modelos preditivos de prognóstico que nos permitam prever a necessidade de uso de VNI em pacientes com ELA. Para ir de encontro a esse objetivos, usámos dados de 1220 pacientes de ELA, seguidos no Hospital de Santa Maria, em Lisboa, que depois de processados são utilizados como input num conjunto de classificadores de forma a criar modelos preditivos capazes de prever se um paciente que chega à consulta vai ou não necessitar de VNI. Uma vez que a aplicação antecipada de VNI é benéfica para o prognóstico dos pacientes, e uma vez que os pacientes são geralmente observados a cada três meses, decidimos então usar janelas temporais que nos permitam também antecipar a necessidade desta terapêutica. As janelas temporais escolhidas foram então 90, 180, e 365 dias (3, 6, e 12 meses). Assim, no conjunto de dados usado ao longo desta dissertação, cada instância pode ser vista como um tuplo constituído por um vetor de atributos que descrevem a condição de um paciente num determinado ponto no tempo, e uma classe Evolução que expressa

a informação sobre a necessidade ou não do uso de VNI para aquele paciente, num intervalo de k dias entre a última consulta e esse ponto no tempo.

Nesta primeira abordagem foram obtidos modelos preditivos com resultados promissores, especialmente quando usadas as janelas temporais mais longas, uma vez que estas são geralmente mais balanceadas e o que contribui para uma melhor performance.

A ELA é uma doença complexa e altamente heterogênea e apesar da sobrevivência média ser de três a cinco anos, existem pacientes que cuja sobrevivência pode ser menos de um ano, e outros que podem viver mais de 10 anos com a doença. Um dos problemas comumente associados a estudos de ELA é a incapacidade de criar tratamentos e desenvolver medicamentos que sejam benéficos para todos estes doentes. Assim, a estratificação de pacientes tem sido uma ferramenta útil para tentar contornar este problema, promovendo o desenvolvimento de terapêuticas mais personalizadas e mais eficazes de acordo com as necessidades de cada paciente.

Nesta dissertação, propomos o uso de duas abordagens de estratificação em pacientes de ELA, de forma a criar modelos com maior nível de especialização. Na primeira estratificamos os pacientes em grupos de acordo com a sua progressão da doença e na segunda estratificamos os pacientes de acordo com a sua condição em cada consulta, de acordo com um dado perfil clínico, criando assim perfis de pacientes.

Para a primeira abordagem foi calculado o declínio na Escala de Classificação Funcional de ELA (ALS Functional Rating Scale) de cada doente, e a partir da distribuição conjunta de todos os pacientes foram criados 3 grupos de progressão: Lentos, Neutros e Rápidos. A informação sobre cada grupo foi depois usada para criar conjuntos de dados contendo apenas os doentes de cada grupo e novos modelos mais especializados foram treinados. À primeira vista os resultados obtidos nesta abordagem não são benéficos para os modelos e no caso dos progressores rápidos parecem mesmo ser prejudiciais. No entanto, para verificar a sua veracidade, voltámos a correr os classificadores com todos os pacientes, e analisando cada predição feita conseguimos obter uma noção de como se comporta o modelo geral a prever cada grupo. Com essa análise pudemos verificar que na verdade os modelos gerais apenas classificam com sucesso os progressores neutros (grupo mais representativo da população geral), e para os dois grupos mais extremos limita-se a prever pela classe maioritária desse grupo. Com isto conseguimos então provar que a utilização destes modelos especializados em cada grupo de progressão, são mais eficazes a prever do que os modelos que usam toda a população disponível.

Na segunda abordagem agrupamos as observações de pacientes mais similares, de acordo com um conjunto predefinido de features (perfil clínico), de forma a obter grupos de observações parecidas a que chamamos perfis de pacientes. Usámos 4 conjuntos diferentes de perfis clínicos: Geral, Prognóstico, Respiratório e Funcional, que diferem de acordo com o conjunto de atributos

usado para os criar. Depois dos perfis de pacientes gerados, vamos mais uma vez treinar novos modelos especializados em cada perfil. Os conjuntos de perfis que mostraram melhores resultados foram o Geral e o de Prognóstico, alcançando resultados melhores que os modelos base. Provamos novamente os potenciais da estratificação para criar modelos especializados mais capazes de prever a necessidade de uso de VNI, que por sua vez permitem um melhor acompanhamento do paciente.

Numa tentativa de melhorar os modelos, testámos o uso de um método de seleção de variáveis nos nossos modelos, que apesar de não mostrar melhorias em relação aos modelos anteriores, se tornou bastante útil por dele conseguirmos extrair a informação de que testes são mais importantes para a esta previsão. Ter esse conhecimento é uma mais valia para os clínicos, uma vez que permite fazer um melhor planeamento dos testes e exames a efectuar para grupos específicos de pacientes, o que resulta numa melhor gestão de tempo (vital quando falamos de pacientes com ELA).

Em cada abordagem presente neste documento testámos ainda um método de preenchimento de valores em falta, denominado de Última Observação Levada Adiante, onde os valores em falta são preenchidos de com o valor da observação anterior, caso esta esteja presente. Esta técnica permite-nos obter um conjunto de dados mais preenchidos, o que é benéfico para os modelos. De facto, o uso deste método, provou ser benéfico para todas as abordagens desta dissertação.

Para os modelos base e grupos de progressão experimentámos também criar modelos que usassem informação histórica do paciente (múltiplas observações) como input, no entanto, apesar de alguns modelos mostrarem resultados semelhantes aos modelos usando apenas a condição atual do paciente, na sua maioria estes modelos demonstraram ter pior performance em relação aos anteriores.

Por fim, o trabalho apresentado nesta dissertação resulta na proposta de duas abordagens de estratificação de pacientes para a criação de modelos personalizados a grupos de pacientes o que possibilita um melhor acompanhamento dos pacientes por parte dos clínicos, e por sua vez, melhora o prognóstico e a qualidade de vida.

Acknowledgements

First, I thank my parents, Edite and João, and brother João André, for the unconditional support they have given me, not only during this dissertation but in all my life. If not for them I would never be able to pursue this goal. To my boyfriend Leonel, a big thank you, for always being present and for listening to my everyday conquests and challenges. I am also grateful for my fellow masters' students and dear friend, for all the moments shared in the last two years, which contributed to me becoming a better person. I also thank the great people at Lasige, where I developed this dissertation, for always being available to share their knowledge and making Lasige a great environment to work. I also want to express my gratitude to Dr. Mamede de Carvalho and his team at Instituto de Medicina Molecular, for the precious clinical insight given during this dissertation and for all the advice given. Then, to Prof. Sara Madeira, my advisor, first for the opportunity to work in this project and then for the availability to answer my every question and for all the guidance provided in this last year. Last but not least, a formal acknowledgment to LASIGE Research Unit, (ref. UID/CEC/00408/2013) and Fundação para a Ciência e Tecnologia for the funding mainly through the Neuroclinomics2 project (PTDC/EEI-SII/1937/2014) and a bachelor grant.

"I am just a child who has never grown up. I still keep asking these 'how' and 'why' questions. Occasionally, I find an answer."

-Stephen Hawking

Contents

| | | |
|----------|--|-----------|
| 1 | Introduction | 1 |
| 1.1 | Motivation | 1 |
| 1.2 | Problem Formulation and Original Contributions | 2 |
| 1.3 | Thesis Outline | 3 |
| 2 | Background | 5 |
| 2.1 | Amyotrophic Lateral Sclerosis | 5 |
| 2.2 | Data Mining Techniques | 7 |
| 2.2.1 | Data Preprocessing | 7 |
| 2.2.1.1 | Feature Selection | 7 |
| 2.2.1.2 | Missing Value Imputation | 8 |
| 2.2.1.3 | Dealing with Imbalanced Data | 9 |
| 2.2.2 | Machine Learning | 9 |
| 2.2.2.1 | Supervised Learning | 9 |
| 2.2.2.2 | Unsupervised Learning | 14 |
| 2.2.2.3 | Model Evaluation and Selection | 15 |
| 2.3 | Prognostic Prediction in ALS | 19 |
| 2.3.1 | Patient Snapshots and Evolution Class | 19 |
| 2.4 | Patient Stratification in ALS | 21 |
| 2.4.1 | Disease Progression Groups | 22 |
| 2.4.2 | Patient Profiles | 23 |
| 2.5 | Portuguese ALS Dataset | 24 |
| 3 | Time Independent Prognostic Models | 27 |
| 3.1 | Single Snapshot Prediction | 28 |
| 3.1.1 | Creating Learning Instances | 28 |
| 3.1.2 | Learning the Predictive Models | 29 |

CONTENTS

| | | |
|----------|--|-----------|
| 3.1.3 | Results and Conclusions | 32 |
| 3.2 | Using a set of Snapshots | 35 |
| 3.2.1 | Creating Learning Instances | 35 |
| 3.2.2 | Learning the Predictive Models | 37 |
| 3.2.3 | Results and Conclusions | 38 |
| 4 | Progression Groups | 45 |
| 4.1 | Single Snapshot Prediction | 46 |
| 4.1.1 | Creating Learning Instances | 46 |
| 4.1.2 | Learning the predictive Models | 49 |
| 4.1.3 | Results and Conclusions | 51 |
| 4.2 | Using a Set of Snapshots | 56 |
| 4.2.1 | Creating Learning Instances | 56 |
| 4.2.2 | Learning the Predictive Models | 57 |
| 4.2.3 | Results and Conclusions | 58 |
| 5 | Patient Profiles | 65 |
| 5.1 | Creating Patient Profiles | 65 |
| 5.2 | Single Snapshot Prediction | 66 |
| 5.2.1 | Creating Learning Instances | 67 |
| 5.2.2 | Learning the predictive Models | 68 |
| 5.2.3 | Results and Conclusions | 71 |
| 6 | Conclusions and Future Work | 77 |
| 6.1 | Conclusions | 77 |
| 6.2 | Future Work | 78 |
| | References | 79 |
| | Appendices | |

List of Figures

- 2.1 Example of a Decision Tree for the following problem: "Will the customer buy or not buy a computer?". Rectangle boxes are the attributes, branches are the possible values and oval boxes the predictions. Adapted from Han *et al.* (2012) 10
- 2.2 Example of a k -Nearest Neighbor classifier using 3 neighbors ($k=3$). The blue rectangles are instances for one class and green circles instances for the other. The orange triangle is the new instance. 11
- 2.3 Example of a SVM. Figure adapted from Aha *et al.* (1991). 12
- 2.4 Example of a Agglomerative Hierarchical Clustering. 15
- 2.5 Example of a ROC curve. 18
- 2.6 Example of creating Snapshots. 20
- 2.7 Definition of the Evolution Class (E) according to the patient’s requirement of NIV in the interval of k days. i is the median date of the snapshot. $E=1$ means the patient requires NIV and $E=0$ means the patient does not require NIV. Adapted from Carreiro (2016). 21
- 2.8 Example of creating learning instances using a time window of 90 days. 22

- 3.1 Problem Formulation: Knowing the Patient current condition, can we predict the need for Non-Invasise Ventilation (NIV) within a time window of k days? 27
- 3.2 Workflow of the methodology for ALS prognostic prediction using patient snapshots 28
- 3.3 Problem Formulation: Given a set of N consecutive patient evaluations, can we predict the need for Non-Invasive Ventilation (NIV) k days after the last evaluation? 36
- 3.4 Example of learning examples using multiple snapshots. 36
- 3.5 Example of transformation using temporal aggregation. 38

- 4.1 Progression Rate Distribution among all patients. 46

LIST OF FIGURES

- 4.2 Problem Reformulation using Progression Groups: Knowing the Patient current state, as well as their progression group can we predict the need for Non-Invasive Ventilation k days after, using group specific models? 47
- 4.3 Workflow of the proposed methodology for ALS prognostic prediction using patient snapshots and progression groups. 48
- 4.4 Problem Reformulation using Progression Groups: Knowing the Patient clinical history (follow-up), as well as their progression group can we predict the need for Non-Invasive Ventilation within k days from the last evaluation, using specialized models for each group? 57

- 5.1 Problem reformulation using Patient Profiles: Knowing the Patient current state, as well as the attributed patient profile (for a given clinical profile) can we predict the need for NIV within a given time window? 66
- 5.2 Workflow of the proposed methodology for ALS prognostic prediction using patient snapshots and patient profiles. 67

List of Tables

- 2.1 Confusion Matrix. 16
- 2.2 Available Features in the ALS dataset. 25
- 3.1 Statistics and Class distribution for time windows of $k=90,180,365$ days. 29
- 3.2 Parameters and correspondent ranges tested for each classifier. 30
- 3.3 Selected Features by the Feature Selection Ensemble for each time window. 30
- 3.4 Proportion of Missing Data before and after Last Observation Carried Forward (LOCF) imputation. 32
- 3.5 AUC, Sensitivity and Specificity results for the prognostic models for 90, 180 and 365 days. 32
- 3.6 AUC, Sensitivity and Specificity results for the prognostic models for the 90, 180 and 365 days Time Windows using Feature Selection. 34
- 3.7 AUC, Sensitivity, and Specificity results for the prognostic models for the 90, 180 and 365 days Time Windows using Missing Value Imputation. 34
- 3.8 Best results achieved for baseline models using the patients current condition. 35
- 3.9 Statistics and Class distribution for time windows of $k=90,180,$ and 365 days. 37
- 3.10 AUC, Sensitivity and Specificity results for the prognostic models using 2 and 3 time points (TP) for each time windows of $k=90,180,365$ days. 39
- 3.11 AUC, Sensitivity and Specificity results for the prognostic models using 2 or 3 patient time points (TP) using Feature Selection 40
- 3.12 AUC, Sensitivity and Specificity results for the prognostic models using 2 or 3 time points (TP) for the 90, 180 and 365 days Time Windows using Missing Value Imputation. 41
- 3.13 AUC, Sensitivity and Specificity results for the prognostic models using temporal aggregation. 42
- 3.14 Best results obtained for the baseline models using 1 TP (current condition) as well as using 2 and 3 TP (clinical history). 43

LIST OF TABLES

| | | |
|------|---|----|
| 4.1 | Statistics and Class distribution for time windows of $k=90,180$, and 365 , for each progression group. | 47 |
| 4.2 | Selected Features for each progression groups and each time window. | 50 |
| 4.3 | AUC, Sensitivity and Specificity results for the prognostic models for the $90, 180$ and 365 days Time Windows, as well as for each Progression Group. | 51 |
| 4.4 | AUC, Sensitivity and Specificity results for the prognostic models for the $90, 180$ and 365 days Time Windows, as well as for each Progression Group using Feature Selection. | 53 |
| 4.5 | AUC, Sensitivity and Specificity results for the prognostic models for the $90, 180$ and 365 days Time Windows, as well as for each Progression Group using Missing Value Imputation. | 54 |
| 4.6 | AUC, Sensitivity and Specificity results for the prognostic models built without progression groups (baseline models) for the $90, 180$ and 365 days Time Windows, relative to each Progression Group | 55 |
| 4.7 | Best results obtained for the specialized models for each progression group, using the patients current condition. | 56 |
| 4.8 | Statistics and Class distribution for time windows of $k=90,180$, and 365 , for each progression group using 2 and 3 time points TP. | 58 |
| 4.9 | AUC, Sensitivity and Specificity results for the prognostic models for the $90, 180$ and 365 days Time Windows, as well as for each Progression Group using 2 and 3 Time Points. | 59 |
| 4.10 | AUC, Sensitivity and Specificity results for the prognostic models for the $90, 180$ and 365 days Time Windows, as well as for each Progression Group, using 2 Time Points (TP). | 60 |
| 4.11 | AUC, Sensitivity and Specificity results for the prognostic models for the $90, 180$ and 365 days Time Windows, as well as for each Progression Group, using 3 Time Points (TP). | 61 |
| 4.12 | Best results obtained for the specialized models for each disease progression group using 1 TP (current condition) as well as using 2 and 3 TP (clinical history). | 63 |
| 5.1 | Statistics and class distribution for each profile, across all time windows. | 69 |
| 5.2 | Selected features by the Feature Selection Ensemble (FSE) for each clinical profile and respective set of Patient Profiles. | 70 |

| | | |
|-----|--|----|
| 5.3 | AUC, Sensitivity and Specificity results for the prognostic models for the 90, 180 and 365 days Time Windows, as well as for each Clinical Profile and respective set of Patient Profiles. | 71 |
| 5.4 | AUC, Sensitivity and Specificity results for the prognostic models using Feature Selection for the 90, 180 and 365 days Time Windows, as well as for each Clinical Profile and respective set Patient Profiles. | 72 |
| 5.5 | AUC, Sensitivity and Specificity results for the prognostic models using Missing Value Imputation for the 90, 180 and 365 days Time Windows, as well as for each Clinical Profile and respective set Patient Profiles. | 73 |
| 5.6 | Best Results obtained for each Clinical Profile and respective set of patients profiles. | 75 |

Chapter 1

Introduction

1.1 Motivation

Amyotrophic Lateral Sclerosis (ALS) is a neurodegenerative disease characterized by the progressive loss of motor neurons in the brain and spinal cord, which leads to muscular weakness and ultimately ends in death (van Es *et al.* (2017)). The life expectancy of an ALS patient is 3 to 5 years after disease onset (Brown & Al-Chalabi (2017)). There is no cure or known causes for ALS, and the heterogeneity of the disease makes it difficult to understand its underlying mechanisms. Thus finding solutions to cure or slow disease progression is a challenge. Therefore, efforts must be taken to find solutions that can improve patient's prognosis and help to maintain the patients quality of life.

ALS studies focus mainly on patient survival (Georgouloupoulou *et al.*, 2013; Pastula *et al.*, 2009; Traynor *et al.*, 2003), exploring the impact of diagnostic delay (Gupta *et al.* (2012)), understanding and defining ALS sub-types (Chiò *et al.* (2011)), or finding relevant clinical features that can be used both as diagnostic or as prognostic predictors (Creemers *et al.* (2015)).

With the rapid advance in the fields of computer science, genetics, imaging, and other technologies came the promise of a new form of medicine, so-called precision medicine. It can be defined as targeted treatments for individual patients based on their genetic, phenotypic or psychological characteristics (Larry Jameson & Longo (2015)). Although this approach has already shown promising results in areas such as cancer, it is only now beginning to be used in ALS studies (Zou *et al.* (2016)).

In recent years there has been increasing attention to patient stratification in ALS. By grouping patients either by their progression level (Westeneng *et al.* (2018)) or by a set of prognostic features (Ganesalingam *et al.* (2009)), it has been shown to be possible to design new treatments

1. INTRODUCTION

or disease management strategies which are specialized for a specific group of patients that has something in common.

Since respiratory failure is responsible for the majority of deaths in ALS patients, there is a need to prevent the decline in respiratory capacity as earlier as possible. Non-invasive ventilation (NIV) is the standard treatment for respiratory impairment in ALS patients and has proven to prolong survival and improve quality of life, especially when administered in earlier stages (Georges *et al.* (2014)).

1.2 Problem Formulation and Original Contributions

This dissertation is the follow-up to the work developed in André Carreiro's PhD Thesis (Carreiro (2016)) in which he proposes an integrative approach combining supervised learning, clinical data of ALS patients and the insights from ALS experts, to develop prognostic models to predict changes in the clinical state of patients according to a given time window.

In this dissertation we will revisit and explore further some of the addressed questions, using an updated dataset containing more data and some approach alterations. Those questions being revisited are:

- Given a patient evaluation, can we predict if a given patient will require NIV within a certain time window?
- Given a set of consecutive patient evaluations (T_1, T_2, \dots, T_k) can we predict if the patient will require NIV within a certain time window after evaluation T_k ?

The dataset used to answer the questions above is the Portuguese ALS dataset presented in Section 2.5. Furthermore, to the questions above (also tackled by Carreiro et al), and whose results obtained with the updated data we use as baseline, we propose two different approaches to stratify patients: according to disease progression and patient profiles (Sections 2.4.1 and 2.4.2). We also investigate whether using these groups in specialized models for each group yields better results when answering the questions above.

Part of the contribution in this thesis we presented at The Sixth Workshop on Data Mining in Biomedical Informatics and Healthcare, held in conjunction with the IEEE International Conference on Data Mining (ICDM'18). The publication is presented in Appendix A.

1.3 Thesis Outline

Other than the current, this thesis is outlined over 5 additional chapters:

- Chapter 2 presents the background regarding ALS, Patient Stratification, Prognostic Prediction as well as concepts of Machine Learning and Data Mining techniques used in this dissertation. It also provides an overview of the dataset used in this work;
- Chapter 3 tackles the two questions proposed above, and presents the baseline results for this dissertation;
- Chapter 4 explores the first approach to patient stratification, presenting the results of the prognostic models using disease progression groups;
- Chapter 5 explores the second approach to patient stratification, presenting the results of the prognostic models using patient profiles;
- Chapter 6 presents the conclusions and future work.

Chapter 2

Background

2.1 Amyotrophic Lateral Sclerosis

Amyotrophic lateral sclerosis (ALS), also known as Motor Neuron Disease (MND) is a complex neurodegenerative disease characterized by the progressive loss of motor neurons in the brain and spinal cord (van Es *et al.* (2017)).

The prevalence of the disease is approximately 3-5 cases per 100.000 individuals. Although this is a seemingly low number for the general population, the risk of developing ALS increases in the latter years of life, reaching a risk of 1:300 at the age of 85 (Martin *et al.* (2017)).

Around 10% of ALS cases are familial (patients which have/had relatives with ALS) and the remaining 90% are considered sporadic. Although there are no major differences in the presentation and progression, familial ALS patients present an earlier disease onset than the ones with sporadic ALS (Brown & Al-Chalabi (2017)). The disease onset is within 58-63 years for sporadic ALS, and within 43-63 years for familial ALS (Andersen *et al.* (2012)).

With recent advances in genetics, more information about the disease is being discovered, providing new insights into what causes ALS and what are the risk factors. More than 20 genes related to ALS have been discovered in recent years, three of them seem to be more relevant than the others (Martin *et al.* (2017)). The *SOD1* gene was the first gene identified as being associated with ALS and for a long time the only gene known to be related with this disease (van Es *et al.* (2017)). *SOD1* mutations are present in approximately 20% of familial ALS patients and 5% of sporadic ALS patients. The second gene, *TARDBP*, represents approximately 5-10% of familial ALS mutations (Zarei *et al.* (2015)). The last gene, *C9orf72*, is responsible for 30% of familial ALS and up to 10% of sporadic ALS, being the gene with the biggest association to ALS (Martin *et al.* (2017)). Around 50-60% of the familial ALS patients have mutations in the described genes. However, there are still 40-50% (Zarei *et al.* (2015)) of the familial ALS cases

2. BACKGROUND

that are not linked to any gene, and an even greater percentage regarding sporadic ALS cases. The Mine project (Mine & Sequencing (2018)) launched a large-scale whole-genome sequencing study using 15 000 ALS patients and 7500 controls. The aim of this project is to discover new genetic risk factors and further elucidating the genetic basis of ALS. The ONWebDUALS project (ONtology-based Web Database for Understanding Amyotrophic Lateral Sclerosis) aims to create a standardized European database, with genetic and phenotypic information of ALS patients, in order to identify relevant risk and prognostic factors in ALS.

Non-genetic factors have also been linked to ALS, such as exposure to toxins, smoking, excessive physical activity, occupation, dietary factors and changes in immunity, especially regarding sporadic ALS patients (Zou *et al.* (2016)).

The initial symptoms of the disease are muscle weakness, twitching, and cramping, which can later lead to muscle impairment. These usually start in the limbs. However, a third of the ALS patients have a bulbar onset, characterized by difficulty in swallowing, chewing, and speaking (Brown & Al-Chalabi (2017)). Dyspnea and dysphagia are usually developed in more advanced stages of the disease (Zarei *et al.* (2015)). The eye and bladder muscles are usually the less affected, showing signs of impairment only in the latest stages of the disease (Brown & Al-Chalabi (2017)).

Regarding diagnosis, there is usually a delay of 13–18 months from the onset of a patient’s symptoms to confirmation of the diagnosis (Zarei *et al.* (2015)). This can be a consequence of the low prevalence of the disease, meaning it is not common for a primary care physician to have many patients with ALS. Moreover, the overlap with other neurodegenerative diseases may be difficult and delay the diagnosis (Hardiman *et al.* (2011)). These delays can worsen the prognosis of the patients since therapies have usually better outcomes when applied in the early stages of the disease. There is not yet a single test to directly diagnose ALS. Therefore, the initial steps towards diagnosis are the exclusion of other neurodegenerative diseases as well as other limb dysfunction causers. Then, Electrodiagnostic tests, neuroimaging, or laboratory tests can be used to find a final diagnosis (Zarei *et al.* (2015)).

There is no definitive cure for ALS. Thus, available treatments are focused on slowing the disease rather than stopping it. Riluzole is the only approved drug treatment, showing an increase in survival up to 14.8 months (van Es *et al.* (2017)). Other available treatments consist in symptom relief and progression. Symptomatic intervention and supporting care for ALS patients include the provision of ventilatory support, nasogastric feeding, and prevention of aspiration (Brown & Al-Chalabi (2017)). The latter consists in control of salivary secretions and use of cough-assist devices. Moreover, Nasogastric feeding helps preventing malnutrition, common in ALS patients, which improves survival and quality of life (van Es *et al.* (2017)).

The majority of ALS patients usually die of respiratory failure within 3-5 years from disease onset (Brown & Al-Chalabi (2017)). Thus, preventive treatments to maintain respiratory muscle function are vital for these patients. Non-Invasive Ventilation (NIV) is the only treatment to prevent respiratory failure in ALS patients (Georges *et al.* (2014)). The use of NIV, when administrated in earlier stages of the disease, has shown to improve survival in ALS patients and their quality of life.

With no definitive cure available, the focus in ALS patients' care is to slow disease progression, improve prognosis, and maintain the quality of life. To achieve these objectives, a multidisciplinary team of clinicians, together with the caregivers, is imperative to ensure the patient has the best care, resulting in increasing survival (Brown & Al-Chalabi (2017)). In this context, during clinical follow-up, the patient's condition is evaluated using ALS related functional scores, Respiratory and Neurophysiological tests, as well as other physical values. These Longitudinal data together with the static data collected at disease diagnosis is used for prognostic prediction.

2.2 Data Mining Techniques

Data Mining is a multidisciplinary subject that combines domains such as statistics, machine learning, pattern recognition, database and data warehouse systems, information retrieval, visualization, algorithms and high-performance computing (Han *et al.* (2012)). Its main purpose is to extract new and useful information from collections of data (Laxman & Sastry (2006)). Data Mining techniques allow us to find patterns and relationships in data which could go unnoticed to the human eye (Sharma *et al.* (2018)). It has applications in many fields such as science, marketing, finance, healthcare or retail (Fayyad *et al.* (1996)).

2.2.1 Data Preprocessing

2.2.1.1 Feature Selection

It is common in Machine Learning (ML) problems, to have a dataset that has a very high number of features (dimensions). With high dimensionality, the amount of data needed to build reliable models is also very high. Without enough learning instances the performance of the classifiers can be hindered (Bolón-Canedo *et al.* (2014)).

Having too many features can be detrimental for the model's performance, and this is a known problem, commonly called by "The curse of dimensionality" (Somorjai *et al.* (2003)). Dimensionality Reduction is one of the most popular techniques to deal with this issue. It can be divided into Feature Extraction and Feature selection. Feature extraction combines the original features into a new set of feature with reduced dimensionality (Tang *et al.* (2014)). As

2. BACKGROUND

for Feature Selection (FS), it selects a subset of features which better describe the data and reduce the effects of noisy and irrelevant features (Chandrashekar & Sahin (2014)).

Using FS to downsize the number of features in a dataset, can bring several benefits: data visualization becomes easier due to the lower dimensionality, reduces storage requirements as well as reduced training and utilization times (Guyon & Elisseeff (2003)).

For a recent survey on FS see (Li *et al.* (2017)). In this survey the authors divide the methods in filter, wrapper, and embedded methods.

2.2.1.2 Missing Value Imputation

Missing data is common in almost all real datasets. However, missing data means lack of information, which can be problematic as some data mining methods rely on complete datasets to work.

There are 3 types of missing data: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) (Newgard & Lewis (2015)). MCAR is the least common and also the least problematic since it yields less biased results. MAR is a more realistic version on MCAR, but can lead to biased results. Lastly, MNAR is the most problematic of the three, making almost impossible to find a statistically approach to deal with this type.

There are mainly two ways to deal with missing data: deletion and imputation (Cheema (2014)). Deletion methods consist in removing either instances or columns with missing values. However, in datasets with low quantities of data or with many missing data this can lead to a considerable loss of information. On the other hand, imputation methods do not remove any instances or columns, but rather replace the missing values with predicted values obtained from the study of the whole dataset.

Missing Value Imputation (MVI) methods can be divided into two groups: Single Imputation (SI) or Multiple Imputation (MI). Single imputation methods replace the missing value with plausible values by observing the characteristics of the population. The most common method is Mean Imputation which imputes missing values using the population mean for the variable. However, when performed in datasets with large quantities of missing data, it leads to a loss of feature variance and correlation distortion that can lead to a biased dataset (Josse & Husson (2012)). Last Observation Carried Forward (LOCF) presents as an alternative to Mean Imputation, by assuming that the value does not change from the last observation (Newgard & Lewis (2015)). This methodology is especially common in clinical datasets, where longitudinal data is available.

MI methods consist in imputing multiple versions of the same dataset, each with a different imputation methodology (Donders *et al.* (2006)).

2.2.1.3 Dealing with Imbalanced Data

Having an imbalanced data, means having a dataset where there are more instances for one of the class values than the other. In imbalanced data the number of available instances for each of the classes to be learned is not the same.

Dealing with imbalanced datasets presents a challenge for some machine learning problems, as most machine learning algorithms assume that classes are balanced. However, in real-life problems, this is seldom the case (Krawczyk (2016)).

Undersampling and Oversampling are two opposite alternatives to deal with class imbalance. The first consists in removing instances from the majority class and the latter on adding instances for the minority class, both of them ensuring a balanced dataset (Rahman & Davis (2013)). In undersampling techniques, instances of the majority class are randomly removed until the number of instances for each class is equalized. Problems with this method lie in the possible loss of information from the removed instances and the resulting low number of overall instances when the minority class has only a few instances (He & Garcia (2009)). Oversampling techniques overcome these problems by keeping all instances and resampling minority class instances until class balance is achieved. This can be done by simply resampling the minority instances, however, this easily can lead to a biased dataset (He & Garcia (2009)).

Synthetic Minority Over-sampling Technique (SMOTE) proposes an oversampling alternative, as it creates synthetic learning instances for the minority class by using k-Nearest Neighbors method to find similar k instances to one of the minority class examples, and use them to create a new instance (Chawla *et al.* (2002)).

2.2.2 Machine Learning

2.2.2.1 Supervised Learning

Supervised Learning algorithms consist in learning a function from a set of training data able to predict the desired output. Each training instance consists of a vector with information on each variable in the data, and a truth value for the outcome we are trying to predict (label/class). After training, the model should be able to receive the feature vector as input, and according to the function learned by the model, output a prediction for the outcome.

These algorithms can be divided into two groups: classification and regression. The first outputs the prediction in a discrete value, usually binary, and the latter outputs a continuous value for the prediction.

2. BACKGROUND

Decision Trees

Decision Tree (DT) is a simple and powerful statistical tool that can be used in classification, prediction or even data manipulation (Song & Lu (2015)). The features are represented as internal nodes of the tree, and each one of its resulting branches are possible values (discrete) or ranges (continuous) of each feature. The final nodes, or leaves, are the predictions (Han *et al.* (2012)). In Figure 2.1 we show a DT example used to predict if a customer will or will not buy a computer.

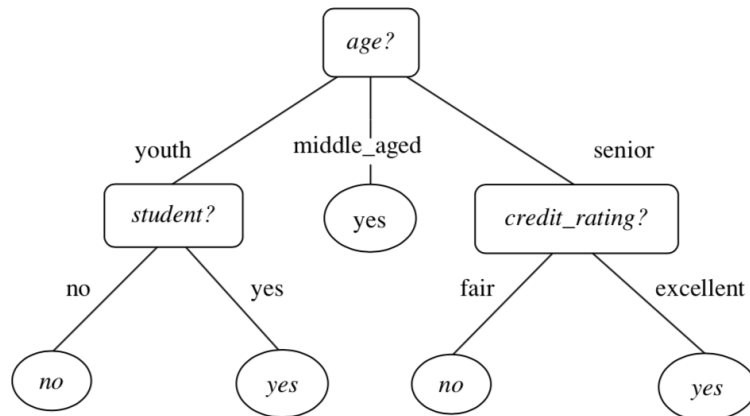


Figure 2.1: Example of a Decision Tree for the following problem: "Will the customer buy or not buy a computer?". Rectangle boxes are the attributes, branches are the possible values and oval boxes the predictions. Adapted from Han *et al.* (2012)

The top of the DT is the feature which better separates data, according to the class values of each instance. The next internal node for each branch will be the best describing feature for the data subset that follows each branch. This last step is repeated until either all features have been used, or all the final branches lead to a prediction. When we reach at a feature in which each value leads to a single prediction, then there is no need to look further in the remaining features and the branches of that feature will lead to the leaves (final predictions) of the tree. In situations where even with all available features, there is no combination that allows reaching single predictions, we have to resort to majority voting. This consists in choosing the prediction value for each branch according to the most popular value (Hu *et al.* (2012)).

The simplicity allied with easy understanding, interpretation, and visualization are some of the advantages of using this algorithm (Song & Lu (2015)). However, for datasets with high dimensionality, visualization and interpretation may be challenging when using DT.

Random Forests

Random Forests (RF) is an ensemble learning method that combines several Decision Trees to make a prediction. Each DT is trained with a random subset of features available.

This method has shown improvement in classification problems as the final prediction is decided by the majority vote of each singular tree prediction (Breiman (2001)).

K-Nearest Neighbors

The k-Nearest Neighbor (k NN) algorithm is a commonly used classifier among many classification problems, in which the output is computed by accessing the outcomes for a given number of learning instances (k) closer to the input instance. It is considered a lazy learner since the classifier is not really trained before its use, but rather trained at the moment of use (Han *et al.* (2012)).

When a new instance is fed to the classifier, the algorithm finds its k -nearest instances. The most common metric to compute the distances between instances is the Euclidean distance, however, other distances can be used (Liu *et al.* (2004)). In order to classify the new instance, the classifier looks at the outcome of each neighbor and chooses according to the majority class of the neighbors. When using k NN, it is advised to use an odd number of neighbors to prevent ties in classification. However, one solution to solve the tie would be using the majority class of the entire dataset or the class of the nearest neighbor. Moreover, to avoid overfitting its advised against using a larger k .

An example of the described algorithm can be seen in Figure 2.2

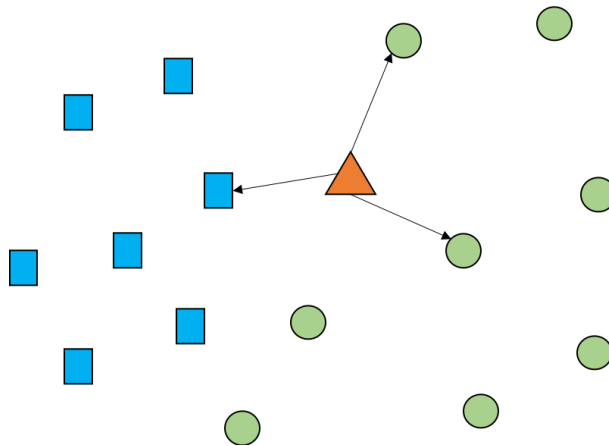


Figure 2.2: Example of a k -Nearest Neighbor classifier using 3 neighbors ($k=3$). The blue rectangles are instances for one class and green circles instances for the other. The orange triangle is the new instance.

Support Vector Machines

2. BACKGROUND

Support Vector Machines (SVM) are powerful models commonly used in nonlinear classification, regression and outliers detection (Aha *et al.* (1991)).

In this algorithm, data is mapped in a multidimensional space (one dimension per feature). Then, the SVM tries to find linear hyperplanes that separate the classes with maximal margins, called support vectors (Hsu *et al.* (2008)). Figure 2.3 shows an example of how support vectors are created. When linear separation is not possible, SVM uses the kernel technique to automatically release non-linear mappings in the feature space (Furey *et al.* (2000)). The most common kernels used are linear, polynomial, radial basis function (RBF), and sigmoid. All support vectors computed are then combined into a function that receives a feature vector as input and outputs a prediction for the outcome in the study.

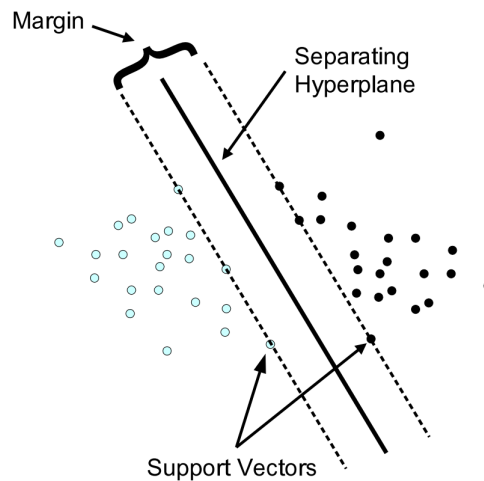


Figure 2.3: Example of a SVM. Figure adapted from Aha *et al.* (1991).

Although SVMs are very robust, they usually work best for problems with fewer features, since a higher number of features translates to a higher number of dimensions and a higher number of support vectors to be computed, which can be detrimental to the performance of the classifier. To use SVM in high dimensional problems it is advised to build the classifiers with only a subset of the features (Hsu *et al.* (2008)).

Naive Bayes

The Naive Bayes (NB) classifier is a simple, probabilistic and easy to use algorithm that has been showing promising results in several applications. It is called Naive because it is based on the assumption that the features are independent of the class (Rish (2001)).

Let $C = (c_1, \dots, c_k)$ be the class variable we are trying to predict and let X be a vector representing the attribute values of each instance, $X = (x_1, \dots, x_n)$. Given a random vector X ,

the classifier predicts its class according to the highest posterior probabilities that are conditioned on X (Han *et al.* (2012)). The posterior probability for each value $i = 1, \dots, k$ is obtained by:

$$p(C = c_i | X = x) = \frac{p(C = c_i)p(X = x | C = c_i)}{p(X = x)}. \quad (2.1)$$

As NB follows the assumption that all variables are independent, then, 2.1 can also be defined by the sum the conditional probabilities of each variable, according to a class value:

$$p(X | C = c_i) = \prod_n p(X = x_j | C = c_i). \quad (2.2)$$

Another assumption followed by the NB classifier is that all numeric attributes follow a Gaussian/Normal distribution. Thus there is a need to estimate a set of parameters from the training data (mean and standard deviation). Density estimation methods have explored ways to overcome the problems with this last assumption. These methods work by averaging over a set of Gaussian kernels:

$$p(X | C = c_i) = \frac{1}{n} \sum_i G(X, \mu, \sigma), \quad (2.3)$$

where i ranges each training point in class c_i , $\mu = x_i$ and $\sigma = \frac{1}{\sqrt{n_i}}$, where n_i is the number of instances with class value c_i .

Logistic Regression

Regression methods are popular in data analysis due to their capability to describe relationships between the target variable and one or more explanatory variables. When using a discrete target, Logistic Regression (LR) is the standard method used (Hosmer & Lemeshow (2000)) and is based on the following logistic function:

$$f(z) = \frac{e^z}{e^z + 1} = (1 + e^z)^{-1}. \quad (2.4)$$

One of the advantages of using LR is that the classifier outputs a real predicted value for each class value, between 0 and 1, that allows to look not only to the prediction but also to its probability (Naive Bayes also does that). The input of the classifier z , usually called *logit*, is a representation of the explanatory variables of the problem and can be defined as:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k, \quad (2.5)$$

2. BACKGROUND

where β_i is the regression coefficient for the variable x_i and β_0 is the probability of the outcome if all variables do not contribute to the problem.

The threshold to determine the predicted class is usually 0.5 in binary classes, meaning that if the probability for a given class value is superior to 0.5 the makes the prediction for that value, and if the probability is lower than 0.5, the classifier prediction is made for the other value.

2.2.2.2 Unsupervised Learning

As opposed to Supervised Learning, Unsupervised Learning methods learn from unlabeled data to find functions that describe it. These methods are usually used to find groups and stratify data and to find hidden patterns. Clustering, Anomaly Detection, Neural Networks and latent variable models are some of the different fields.

Clustering is one of the most popular fields, in which the algorithms partition the data in subsets (clusters) according to the similarities and dissimilarities between instances. The clusters created can be helpful to retrieve new information from the similarity in each cluster but also from the differences between the clusters.

k-Means

K-Means is the simplest and most popular among all clustering algorithms. In this method the number of clusters (k) is defined apriori (Krishna & Murty (1999)).

In order to create the clusters, k random instances of data are chosen to be the initial centroids (points in the middle of each cluster). Then, the remaining instances are individually compared to each of the centroids and assigned to the cluster of the closest one.

After the first iteration, new centroids are computed using the mean of all instances inside one cluster. Then, all instances are once more compared to each centroid and assigned to the closest centroid. This process repeats until the point of conversion (when there is no change in the composition of the centroids between iterations) or until they reach the maximum number of iterations.

Although being simple and presenting overall acceptable results, there are some disadvantages to this method. One, is the need to have prior knowledge of the number of groups to be created. Another disadvantage lies in the number of iterations that change with the number of instances, the number of clusters, and the complexity of the problem, which can make computationally expensive (Alsabti *et al.* (1997)).

Hierarchical Clustering

Hierarchical Clustering (HC) methods create clusters according to a hierarchy and can be divided in two groups: divisive and agglomerative, the latter being the most popular.

Agglomerative Hierarchical Clustering (AHC) methods, work by interactively combining the two closest objects or clusters until all data falls in the same cluster. An opposite approach is used when using the divisive methods.

In the first step of the AHC process, the number of clusters is equal to the number of instances. Then, a proximity matrix is computed to register the distances between each two points of data. The two closest points are then combined in the same cluster and the proximity matrix is recalculated swapping the two instances for the cluster centroid and computing its distance to each of the left over instances. This process repeats itself until achieving one cluster with all instances in it.

When dealing with a relative low number of instances, the results of the AHC can be shown by a dendrogram, which provides an easy visualization (see Figure 2.4). A different number of clusters can be derived by choosing different cut-off in the similarity value.

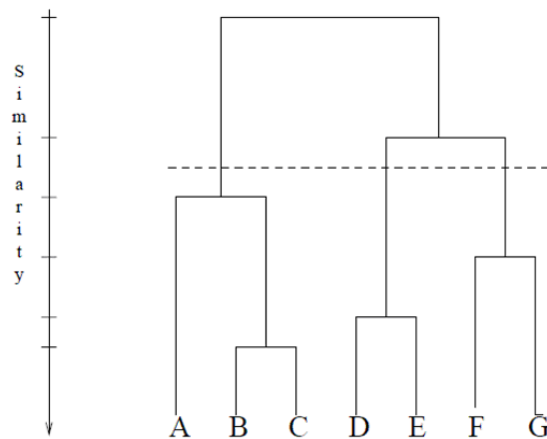


Figure 2.4: Example of a Agglomerative Hierarchical Clustering.

2.2.2.3 Model Evaluation and Selection

Cross-Validation

Cross-Validation (CV) is a popular method for model selection and parameter estimation in supervised learning. It works by splitting data a number of times, in order estimate the performance of each classifier. A subset of data, called training set, is used to train the classifiers, while the rest, testing set, is then used to assess its performance (Arlot & Celisse (2009)).

There are many ways of splitting the data, but the most popular are: Leave-one-out (LOO), Leave-p-out (LPO) and K-Fold (VF). The first two approaches are considered exhaustive splitters and the latter a partial splitter. In LOO, each instance is successively remove from the sample

2. BACKGROUND

and used for model validation, as for LPO, each possible subset of p instances is left out for validation.

In k -Fold CV data is partitioned in k subsets. In each iteration one subset (fold) is held out for test, while the other subsets are used for training purposes (Arlot & Celisse (2009)). This method is much less time consuming in comparison with the other methods as the number of iterations is much lower (one iteration for each fold). However, when splitting data in folds, the folds created can be very similar to each other. To overcome this problem, v -Fold CV is usually performed multiple times, each time generating different folds.

Performance Metrics for Supervised Learning

Confusion Matrix

A confusion matrix is a matrix of $c \times c$ dimensions, where c is the number of classes, usually used to compute performance metrics for model evaluation. The rows usually represent the predicted values and the columns the real values of the class. In binary classification we have a 2×2 table, as shown in the example in Table 2.1.

Table 2.1: Confusion Matrix.

| | | Predicted Class | |
|------------|-----|-----------------|-----|
| | | PC1 | PC2 |
| True Class | TC1 | TP | FN |
| | TC2 | FP | TN |

These tables allow us to evaluate where the classifiers are performing correct or wrong predictions. We can see this by the four indicators provided: Number of True Positives, True Negatives, False Positives and False Negatives, where:

- True Positive (TP): an instance from the positive class that is classified as positive;
- False Negative (FN): an instance from the positive class that is classified as negative;
- True Negative (TN): an instance from the negative class that is classified as negative;
- False Positive (FP): an instance from the negative class that is classified as positive.

The indicators are then used to compute metrics such as Accuracy, Sensitivity and Specificity.

Accuracy

Accuracy gives us the information about how many instances were correctly classified:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}. \quad (2.6)$$

However, when dealing with imbalanced classes, this metric can be biased by how the classifier performs in classifying the majority class.

Sensitivity

Sensitivity, also known as recall or true positive rate, shows the proportion of positive instances that were classified as such:

$$Sensitivity = \frac{TP}{TP + FN}. \quad (2.7)$$

Specificity

Specificity, selectivity or true negative rate shows the same information as the sensitivity metric, but for the negative class. Thus, it gives information about the proportion of negative instances that were correctly classified:

$$Specificity = \frac{TN}{TN + FP}. \quad (2.8)$$

ROC and AUC

When evaluating a classifier using the metrics described above, we have to be careful and take into account the proportions of each class value as to not be biased by the results.

The receiver operating characteristic (ROC) curve combines the Sensitivity and Specificity metrics in a graph for each classification threshold. This results in a graph that shows the performance of a classification model. Figure 2.5 shows an example of a ROC curve.

The closer the curve is to the upper left corner the better the performance of the classifier, since the sensitivity and specificity measures are maximized.

The ROC is also used to compute one of the most popular metrics in performance evaluation, the Area Under the ROC Curve, also known as AUC. As it says in the name, this metric measures the area under the ROC curve. The AUC metric can be defined as either the a representation

2. BACKGROUND

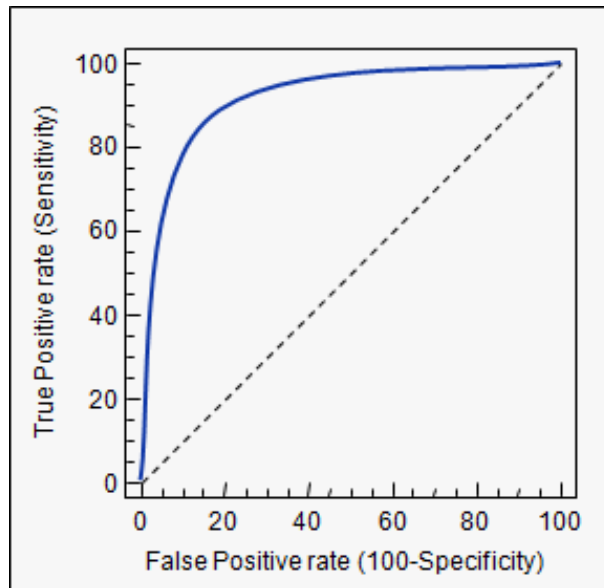


Figure 2.5: Example of a ROC curve.

of the classifier ability to separate the classes or the probability of an instance with a given class value being classified as such.

Cluster Validation

Determining the best number of clusters in a dataset is a common problem in when using clustering, especially when using k-means, where we need to tell the model how many cluster we want to create. There are some measures and scores that can be used to determine the optimal number of clusters to create for our data.

Silhouette Score

The Silhouette score for a given instance in our dataset gives us information on how close our instance is to the other instances in the same cluster. This score can vary between -1 and 1. Scores close to 1 mean the instance is in the right cluster, while scores close to -1 mean the instance is in the wrong cluster.

To determine the optimal number of clusters in a dataset, we can run the clustering algorithms with different number of clusters, and use the average Silhouette score to evaluate how good are the clusters. The clustering with higher Silhouette score, is the one with optimal number of clusters.

2.3 Prognostic Prediction in ALS

As stated previously, with no definitive cure for ALS, efforts are then focused on designing treatments that improve the prognosis of patients and their quality of life. However, many treatments are more effective when administered earlier on. Prognostic prediction can be a key tool in this context. By predicting a given prognosis beforehand, clinicians are then able to administer appropriate treatment before being too late. Nevertheless, the number of studies regarding prognostic prediction in ALS are scarce. Current studies usually focus in finding prognostic biomarkers associated with patient survival (Polkey *et al.*, 2017; Sato *et al.*, 2015).

In the work prior to this dissertation, André Carreiro (Carreiro (2016)) proposed an approach to predict the need of NIV in patients with ALS. He advanced the state of the art by, rather than predicting the immediate need, proposing the use of time windows. This allows to answer the following question: "Given a patient's current condition, will the patient need NIV within k days?". Longitudinal Data from a cohort of 758 patients from the ALS clinic of the Translational Clinical Physiology Unit, Hospital de Santa Maria, Lisbon was used to build the prognostic models.

To obtain learning instances comprising all information about a patient's condition and the need for NIV some preprocessing steps had to be taken. The first was to transform the demographic data and information of each clinical test into Patient Snapshots. Then, as the snapshot only accounts for the current condition, an Evolution Class (E) was created, to label each instance with the information about the patient's needed for NIV within a given time window. These preprocessing steps are detailed in Section 2.3.1. After, several classifiers were trained to predict this outcome. The results obtained were promising and helped to prove that prognostic prediction models can be useful tool in helping clinicians in their decision making process.

The DREAM-Phil Bowen ALS Prediction Prize4Life Challenge (Küffner *et al.* (2015)) encouraged researchers to use clinical trial data to predict disease progression in 3 to 12 months. In fact knowing how patients progress can be useful when designing clinical trials and help clinicians in better determining their patients prognosis. This led to several proposals being presented that helped validating various prognostic features described in the literature (Westeneng *et al.* (2018)).

2.3.1 Patient Snapshots and Evolution Class

In order for a classifier to be trained, it needs labeled learning instances. These instances are composed by a feature vector, which has the information about a patients current condition and

2. BACKGROUND

a class value, which has the information about the outcome on which the classifier will be trained to predict.

The original data used in André Carreiro’s work (Carreiro (2016)) was composed of demographic data about the patient, as well as a set of prescribed tests that are usually done periodically at each appointment. However, since many times the patient cannot perform all tests in the same day, but rather do it in a span of a few days or weeks, it becomes difficult to align all tests together to generate a snapshot that resembles that period. An approach to solve this problem is grouping the tests by their date. A good methodology to achieve this is using an Agglomerative Hierarchical Clustering scheme. This approach was also proposed by André Carreiro as an alternative to the standard approach based on pivot tables, which results in a greater number of instances, but with a higher proportion of missing values.

To build patient snapshots from the original data, the Hierarchical Clustering algorithm will group all the patient’s tests by date so that the tests performed at closer dates will end at the same group, thus creating a snapshot. However, there are some restrictions to the algorithm in order to have cohesive snapshots. First, two observations of the same test cannot be in the same group, and second, all observations in a group must have the same NIV status, meaning that there cannot be groups that have tests performed when the NIV status is 0 (the patient does not require NIV at current time) and tests where NIV status is 1 (the patient requires NIV at current time). The result from this approach is a dataset where each row is a patient observation, also called a patient snapshot, where each feature is the result from each test performed, and a NIV status class, with the information about the need or not for NIV at the time of said evaluation. A fictional example of the described process is illustrated in Figure 2.6.

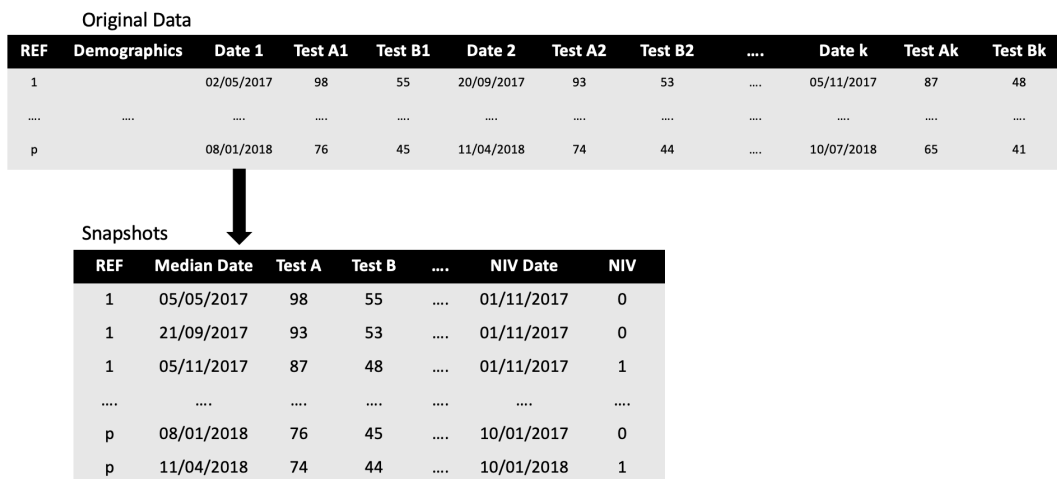


Figure 2.6: Example of creating Snapshots.

After this step, what we have is the information on the patient’s current condition and the current need or not need for NIV. However, since the goal is to predict the need for NIV beforehand, there is one more preprocessing step needed. Therefore, an Evolution class (E) is added in order to accommodate the temporal information regarding the NIV status. Essentially, if a patient needs NIV within a time window of k days from the current evaluation, then $E=1$ (the patient evolves to need NIV). If within the same k days, the patient does not need NIV, then $E=0$ (The patient does not evolve to need NIV).

The creation of this class has also some restrictions. They are as follows:

- Snapshots from patients who already require NIV at the first evaluation cannot be used as learning instances;
- Snapshots where there is no information about the NIV status of the patient after the time window used cannot be used as learning instances.

Figure 2.7 illustrates the possible cases for the creation of the Evolution class (E). Moreover, an illustration of this last preprocessing step is presented in Figure 2.8.

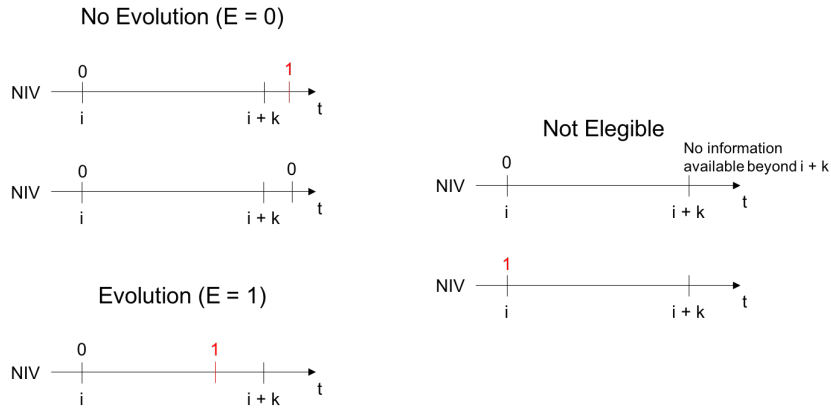


Figure 2.7: Definition of the Evolution Class (E) according to the patient’s requirement of NIV in the interval of k days. i is the median date of the snapshot. $E=1$ means the patient requires NIV and $E=0$ means the patient does not require NIV. Adapted from Carreiro (2016).

2.4 Patient Stratification in ALS

Due to ALS heterogeneous nature, special attention has been given in recent years to patient stratification. The idea is that designing specialized models using groups of patients stratified according to their progression (Westeneng *et al.* (2018)), or specific sets of prognostic biomarkers

2. BACKGROUND

Snapshots

| REF | Median Date | Test A | Test B | | NIV Date | NIV |
|------|-------------|--------|--------|------|------------|------|
| 1 | 05/05/2017 | | | | 01/11/2017 | 0 |
| 1 | 21/09/2017 | | | | 01/11/2017 | 0 |
| | | | | | | |
| p | 08/01/2018 | | | | 10/01/2018 | 0 |
| p | 11/04/2018 | | | | 10/01/2018 | 1 |

Learning Instances

| REF | Median Date | Test A | Test B | | NIV Date | Evolution |
|------|-------------|--------|--------|------|------------|-----------|
| 1 | 05/05/2017 | | | | 01/11/2017 | 0 |
| 1 | 21/09/2017 | | | | 01/11/2017 | 1 |
| | | | | | | |
| p | 08/01/2018 | | | | 10/01/2018 | 1 |
| p | 11/04/2018 | | | | 10/01/2018 | 1 |




Figure 2.8: Example of creating learning instances using a time window of 90 days.

(Ganesalingam *et al.* (2009)) may help understand the underlying mechanisms of the disease and provide a new perspective on how to plan clinical trials and better manage disease progression.

In this dissertation, we propose two approaches to perform patient stratification in ALS patients: 1) disease progression groups and 2) patient profiles. The two methodologies are further explained in Sections 2.4.1 and 2.4.2.

2.4.1 Disease Progression Groups

Although the average survival of an ALS patient is about 3-5 years, survival can vary between less than a year to over 10 years (Martin *et al.* (2017)). This shows that disease progression is not equal in all patients, thus making hindering to have treatments that perform well for all patients.

One way to analyze disease progression is by considering at the ALS Functional Rating Scale (ALSFRS) (Proudfoot *et al.* (2016)) decay in a period of time. The ALSFRS is a standard test used by physicians in practice that can be used to estimate the outcome of a treatment or the progression of the disease. Although very popular, this scale has only a small respiratory component. Given that respiratory failure is the most common cause of death in ALS patients, the ALS functional rating scale-revised (ALSFRS-R) was later proposed (Cedarbaum *et al.* (1999)). This new scale adds additional respiratory assessments and quickly became the preferred test to

quantify disease progression (Simon *et al.* (2014)). The test is composed of 13 questions, where each should be answered using a 5-point scale, ranging from 0 to 4, where 0 corresponds the worse condition and 4 is the best. The questions addressed by this scale are: 1) Speech, 2) salivation, 3) swallowing, 4) handwriting, 5) cutting and handling utensils, 6) dressing and hygiene, 7) turning in bed and adjusting bed clothes, 8) walking, 9) climbing stairs, 10) breathing, 11) dyspnea, 12) orthopnea, and 13) the need of respiratory support (Castrillo-Viguera *et al.* (2010)).

By measuring the change in ALSFRS-R over time, we can estimate how is the disease progressing and infer about the survival of the patient (Kimura *et al.* (2006)). By using the information about the time of first symptoms and the time of the first appointment we can compute its progression rate using the following equation:

$$ProgressionRate = \frac{48 - ALSFRSR_{1stVisit}}{\Delta t_{1stSymptoms;1stVisit}}, \quad (2.9)$$

where 48 is the maximum score of the ALSFRS-R scale (and the assumed score of a patient at the time of its first symptoms), $ALSFRSR_{1stVisit}$ is the ALSFRS-R score of a given patient at the beginning of the first appointment (diagnosis) and $\Delta t_{1stSymptoms;1stVisit}$ is the time in months between the time of first symptoms and the first visit.

By knowing each patient's progression rate we can then group them to build specialized models for each disease progression group.

2.4.2 Patient Profiles

A patient's condition at disease onset is usually more similar to other patients' condition in the same situation than to his/her own condition in the latter stages of the disease.

In this context, this second approach proposed consists in stratifying patients using patient profiles. Instead of grouping patients, we now group patients snapshots. This means that is not obligatory for all snapshots of a given patient to end up in the same group. The aim is to group snapshots that are more similar to each other, thus reducing the variability in the data and potentially enhance the classifiers' performance.

With the help and insight of the clinicians in our group, we propose to stratify the patient snapshots using four sets of patients profiles: General, Prognostic, Respiratory, and Functional. Each set of profiles uses a different subset of features from the original dataset. The subsets of features used for each profile are the following:

- General Profile - All features in the dataset;
- Prognostic Profile - Features described as good prognostic biomarkers in the literature;

2. BACKGROUND

- Respiratory Profile - Features associated with respiratory function;
- Functional Profile - Features associated with functional scores.

By creating several sets of patient profiles, using different sets of features, rather than just the one set with features specific to our problem, we are then able to use the different profiles to predict different outcomes.

2.5 Portuguese ALS Dataset

For this dissertation the dataset used is the Portuguese ALS dataset. It contains clinical data from respiratory tests and neurophysiological data, as well as some demographic factors, from ALS patients. All patients were followed in the ALS clinic of the Translational Clinic Physiology Unit, Hospital de Santa Maria, IMM, Lisbon. Evaluations of patients present in this dataset were made between 1995 and March 2018. It contains observations from a cohort of 1220 patients, resulting in 5553 records with 27 features. Since every patient can have multiple records over time (average 5.18 evaluations per patient), and appointments usually occur every 3 months, we have an average of 15,6 months of follow-up data for each patient.

The dataset has two subsets of features: the static subset (features that do not change over time), containing demographic information like gender and age at onset, medical and family history, onset evaluation and genetic biomarkers, and a temporal subset, with functional scores, respiratory tests and status, some neurophysiological values and the information of when and if the patient has Non-Invasive Ventilation. All temporal features can change in between observations.

The list of features available in the Portuguese ALS dataset are shown in Table 2.2.

Table 2.2: Available Features in the ALS dataset.

| Name | Temporal/Static | Type | SubGroup |
|--|-----------------|------------------|----------------------------|
| Gender | Static | Categorical | Demographics |
| Body Mass Index (BMI) | Static | Numerical | Demographics |
| Family History of Motor Neuron Disease (MND) | Static | Categorical | Medical and Family History |
| UMN vs LMN | Static | Categorical | Onset Evaluation |
| Age at Onset | Static | Numerical | Onset Evaluation |
| Onset Form | Static | Categorical | Onset Evaluation |
| Diagnostic Delay | Static | Numerical | Onset Evaluation |
| El Escorial Reviewed Criteria | Static | Categorical | Onset Evaluation |
| Expression of C9orf72 Mutations | Static | Categorical | Genetic |
| ALSFRS* | Temporal | Numerical | Functional Scores |
| ALSFRS-R* | Temporal | Numerical | Functional Scores |
| ALSFRSb* | Temporal | Numerical | Functional Scores |
| ALSFRSsUL* | Temporal | Numerical | Functional Scores |
| ALSFRSsLL* | Temporal | Numerical | Functional Scores |
| ALSFRSr* | Temporal | Numerical | Functional Scores |
| R* | Temporal | Numerical | Functional Scores |
| Vital Capacity (VC) | Temporal | Numerical | Respiratory Tests |
| Forced VC (FVC) | Temporal | Numerical | Respiratory Tests |
| Airway Occlusion Pressure (P0.1) | Temporal | Numerical | Respiratory Tests |
| Maximal Sniff nasal Inspiratory Pressure (SNIP) | Temporal | Numerical | Respiratory Tests |
| Maximal Inspiratory Pressure (MIP) | Temporal | Numerical | Respiratory Tests |
| Maximal Expiratory Pressure(MEP) | Temporal | Numerical | Respiratory Tests |
| Date of Non-Invasive Ventilation | Temporal | Date/Categorical | Respiratory Status |
| Phrenic Nerve Response amplitude (PhrenMeanAmpl) | Temporal | Numerical | Neurophysiological Tests |
| Phrenic Nerve Response latency (PhrenMeanLat) | Temporal | Numerical | Neurophysiological Tests |
| Cervical Extension | Temporal | Numerical | Other Physical Values |
| Cervical Flexion | Temporal | Numerical | Other Physical Values |

* Scores and Sub-scores of the ALS Functional Rating Scale

Chapter 3

Time Independent Prognostic Models

In this section, the goal is to predict if a patient will require Non-Invasive Ventilation (NIV) within a time window of k days. To do this we use data describing the patient's past evaluation (current condition), as depicted in Figure 3.1.

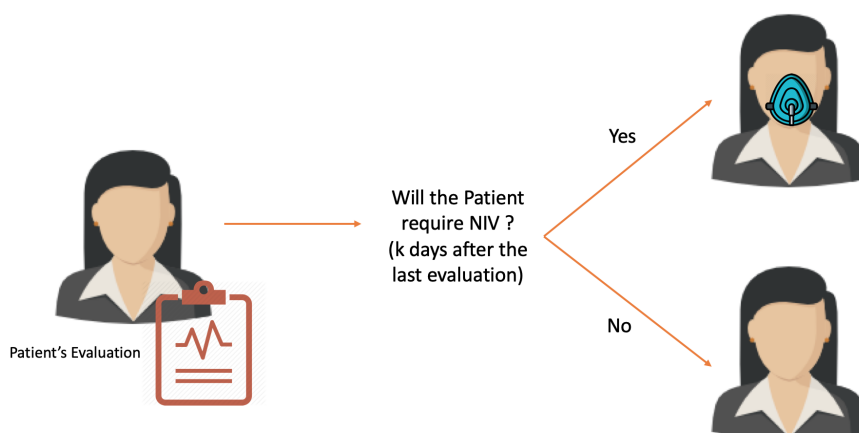


Figure 3.1: Problem Formulation: Knowing the Patient current condition, can we predict the need for Non-Invasive Ventilation (NIV) within a time window of k days?

Figure 3.2 presents the workflow used in this section. First, original data is preprocessed into patient snapshots and then into learning instances using time windows. These steps are followed by building the predictive models capable of predicting the need for NIV within a certain time window, given a patient's current condition. These first models will be used as baseline results for this dissertation. This scheme was proposed in Carreiro *et al.* (2015) with promising results. We use it as well (with a few alterations to the pipeline as well as an updated version of the

3. TIME INDEPENDENT PROGNOSTIC MODELS

dataset). The aim is to improve these results and use them to compare with the results of the two patient stratification approaches we proposed in this thesis.

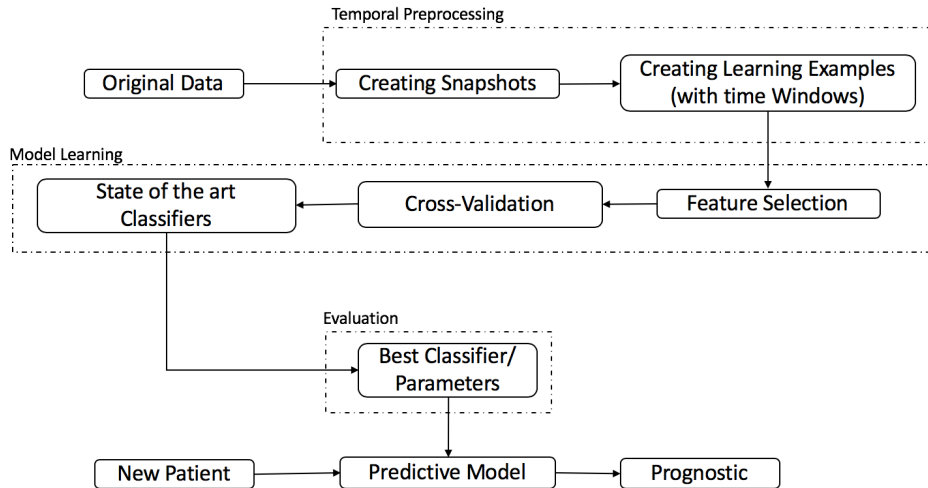


Figure 3.2: Workflow of the methodology for ALS prognostic prediction using patient snapshots (following Carreiro *et al.* (2015)). Original data is preprocessed in order to create patient snapshots that are then used to create the learning instances. The models are then built using a stratified 5 x 10-fold cross-validation scheme. The models are then evaluated and the best parameters are chosen. After the final model is learned, whenever a new patient arrives to the clinic, his data is fed to the predictive model that outputs a prediction: need or not need for NIV.

3.1 Single Snapshot Prediction

3.1.1 Creating Learning Instances

The clinical data used in this work consists in data from a cohort of 1220 ALS patients followed in the ALS clinic of the Translational Clinic Physiology Unit, Hospital de Santa Maria, IMM, Lisbon (a full description of this dataset is available in Section 2.5). To transform the original data into patient snapshots and learning instances, we followed an approach proposed by Carreiro *et al.* (2015) to create patient snapshots and learning instances using time windows (See Section 2.2.1). As appointments are usually every three months, we use time windows of 90, 180, and 365 days (3, 6, and 12 months respectively), as recommended by the clinicians in our group as well as in the literature (Andersen *et al.* (2012)).

Table 3.1 shows some results and statistics for these preprocessing steps, for each time window.

By analyzing the table we can see that the number of snapshots decreases with the increasing of k . This can be explained by the fact that for some snapshots we no longer have information

Table 3.1: Statistics and Class distribution for time windows of $k=90,180,365$ days.

| k | 90 | 180 | 365 |
|----------------------|---------------|---------------|---------------|
| Nr of Snapshots | 3178 | 3018 | 2762 |
| Nr of Patients | 861 | 823 | 775 |
| Snapshots p/ Patient | 3.69 | 3.67 | 3.56 |
| Evolution (E=1) | 559 (17.59%) | 906 (30.02%) | 1342 (48.59%) |
| No Evolution (E= 0) | 2619 (82.41%) | 2112 (69.98%) | 1420 (51.41%) |

about the NIV status after the time window. The same goes for the number of patients, since some patients will require NIV soon after their first appointment. Looking at the class distribution, we can see that for the first time window of 90 days, less than 20% of patients evolve to NIV. However, when we look at the window of 365 days, we observe that the number of NIV evolutions rises to more than 50%. This is to be expected since the probability of a patient requiring NIV increases when we consider longer periods of time. We believe that if we would continue extending k , the number of instances where E=1 would increase and instances with E=0 would continue to decrease. By extending k to a long enough period we believe we would possibly end up with a class distribution of 100%((E=1)/0%(E=0)).

3.1.2 Learning the Predictive Models

After creating the three datasets for the chosen time windows, the next step is to train the predictive models. Each dataset serves as input for 6 different classifiers, using a stratified 5 x 10-fold cross-validation (CV) scheme. The chosen classifiers are Decision Tree (DT), k-Nearest Neighbors (kNN), Support Vector Machines (SVM) with Polynomial (P) and Gaussian (G) kernels, Naïve Bayes (NB), Random Forest (RF), and Logistic Regression (LR). All classifiers used are available in Weka (Hall *et al.* (2009)). Moreover, to find the best parameters for each classifier, we perform a grid search using the range of parameters presented in Table 3.2. For model evaluation, we use the following metrics: Sensitivity, Specificity, and AUC. The AUC metric is used for model comparisons since it combines the results of the other metrics.

Regarding Feature selection (FS) we test the use of a Feature Selection Ensemble (FSE) proposed by Pereira *et al.* (2018). The proposed method combines both stability and predictability to chose the best features for prognostic prediction in Alzheimer’s Disease. This ensemble combines multiple FS algorithms to return a set of features that in general should be less biased by the characteristics of each FS algorithms. The FS algorithms used are ReliefF, Information Gain, Conditional Mutual Information Maximization, Minimum Redundancy Maximum Relevance, and Chi-Squared. This methodology is composed of two phases. First data is used by

3. TIME INDEPENDENT PROGNOSTIC MODELS

the FSE to select a reduced set of features. Then the feature set is optimized for stability and predictability. The selected features for each dataset are presented in Table 3.3.

Table 3.2: Parameters and correspondent ranges tested for each classifier.

| Classifier | Parameter | Range |
|------------|-------------------|---|
| DT | Confidence factor | {0.15,0.20,0.25,0.30} |
| kNN | Nr of Neighbours | {1,3,5,7,9,11} |
| SVM P/G | Complexity | { 10^{-2} , 10^{-1} , 10^1 , 10^2 } |
| SVM P | Polinomial Degree | {1,2,3} |
| SVM G | Gamma | { 10^{-3} , 10^{-2} , 10^{-1} , 10^1 , 10^2 , 10^3 } |
| NB | Kernel | {True,False} |
| RF | Nr of Trees | {5,10,15,20} |
| LR | Ridge Factor | { 10^{-9} , 10^{-8} , 10^{-7} , 10^{-6} , 10^{-5} , 10^{-4} } |

Table 3.3: Selected Features by the Feature Selection Ensemble for each time window.

| Features | 90 days | 180 days | 365 days |
|-------------------------------|---------|----------|----------|
| Gender | | | |
| Age at Onset | X | X | X |
| BMI | X | X | X |
| Family History MND | | | |
| Disease Duration | X | X | X |
| El Escorial Reviewed Criteria | | | |
| UMN vs LMN | | | |
| Onset form | | | |
| c9orf72 | | | |
| ALS-FRS | X | X | X |
| ALS-FRS-R | X | | X |
| ALS-FRSb | X | | |
| ALS-FRSsUL | | | |
| ALS-FRSsLL | | | |
| ALS-FRSr | | | |
| R | | | |
| VC | X | X | X |
| FVC | X | X | X |
| MIP | X | X | X |
| MEP | X | X | X |
| P0.1 | X | X | |
| SNIP | | | |
| PhrenMeanLat | X | | |
| PhrenMeanAmpl | X | | |
| Cervical Flexion | | | |
| Cervical Extension | | | |

Most of the selected features across all datasets are features recognized in the literature as good prognostic indicators in ALS patients. As we are predicting a respiratory target, it would be expected that respiratory features would be more prevalent than the other features regarding other aspects of the disease. Those expectations were somewhat confirmed, however, demographic features are also relevant to the prognostic models.

Although there are some differences, we can see that there is a subset of features that are selected in all datasets (Age at Onset, BMI, Disease Duration, ALS-FRS, VC, FVC, MIP, and MEP). These should be the most important features to predict the need for NIV in ALS patients.

To deal with class imbalance, especially for the 90 and 180 days time windows, we use a combination of undersampling and oversampling. We chose to use this combination instead of the singular use of one of the methods due to the fact that by using undersampling exclusively would lead to a great loss of data, which could lead to loss of performance. By only using oversampling, as the number of positive instances to be created would surpass the number of original instances for that class, it could lead to overfitting problems. Thus, in order to solve our problem, we first use random undersampling to remove instances from the majority class until we obtain a class distribution of 60%/40% (for the majority and minority classes respectively). Then we use SMOTE to oversample the minority class until we end up with a balanced dataset.

One common characteristic of real datasets, especially when dealing with clinical data is the existence of missing values, and although some classifiers are able to work with missing values, others require a complete dataset. In this work we compare two approaches regarding missing data: in the first approach, we use the most common and simple method, Mean Imputation. For the second approach, we use the patients own information to impute the missing data. For the most part, ALS patients do not improve condition between appointments, but rather stay stable or worsen their condition. Therefore, in cases where we have information in one appointment but not in the next, we assume that the patient remained stable and use the value of the last observation to impute the missing value. This methodology is called Last Observation Carried Forward (LOCF) and is commonly applied in clinical settings (Newgard & Lewis (2015)). As this approach can only be used for missing values of which we have information about the last observation, it could not impute the whole dataset. For the remaining missing values, Weka automatically imputes those using Mean imputation for the classifiers that need imputed dataset without missing values. Table 3.4 shows the proportion of missing values for each dataset before and after the application of the methodology, for each time window.

As described before, this approach does not account for all missing values, however, we can see that the proportion of missing data decreases over 10% in all cases, which can be helpful for classifier performance.

3. TIME INDEPENDENT PROGNOSTIC MODELS

Table 3.4: Proportion of Missing Data before and after Last Observation Carried Forward (LOCF) imputation.

| Time Window | Original Dataset | Imputed Dataset |
|-------------|------------------|-----------------|
| 90 days | 31.14% | 19.21% |
| 180 days | 31.06% | 19.25% |
| 365 days | 30.89% | 19.32% |

We would like to note that, the approaches regarding class imbalance and missing value imputation are only performed in each training sets of the 10-fold CV in order to avoid overfitting.

Finally to compare the different approaches in terms of AUC values we use the Wilcoxon Signed Rank Test for paired instances (Rey & Neuhäuser (2011)) to access whether or not the differences are statistically significant.

3.1.3 Results and Conclusions

We follow by presenting the results for the final models with optimized parameters. To evaluate them in terms of performance we use the AUC metric. We first look at the results for our baseline approach (without Feature Selection or Missing Value Imputation) for each of the selected time windows (90, 180, and 365 days). Sensitivity, Specificity, and AUC results for these models are presented in Table 3.5.

Table 3.5: AUC, Sensitivity and Specificity results for the prognostic models for 90, 180 and 365 days. The classifiers are: Decision Trees (DT), k-Nearest Neighbors, Support Vector Machine (SVM) with Polynomial (P) and Gaussian (G) kernels, Naïve Bayes (NB), Random Forest (RF) and Linear Regression (LR).

| | DT | kNN | SVM P | SVM G | NB | RF | LR |
|--------------------|-------|-------|-------|-------|--------------|--------------|-------|
| AUC | | | | | | | |
| 90d | 72.21 | 68.56 | 62.63 | 58.92 | 79.84 | 80.74 | 79.15 |
| 180d | 74.79 | 69.81 | 63.30 | 58.64 | 80.57 | 85.27 | 79.68 |
| 365d | 71.27 | 71.51 | 70.56 | 61.54 | 82.25 | 89.20 | 80.21 |
| Sensitivity | | | | | | | |
| 90d | 60.47 | 58.71 | 66.98 | 69.30 | 69.87 | 68.55 | 69.41 |
| 180d | 63.62 | 52.65 | 69.60 | 67.99 | 73.22 | 70.73 | 72.30 |
| 365d | 75.10 | 67.96 | 70.61 | 71.49 | 78.41 | 82.07 | 73.85 |
| Specificity | | | | | | | |
| 90d | 74.81 | 69.93 | 58.29 | 48.53 | 74.99 | 77.60 | 74.78 |
| 180d | 74.01 | 76.15 | 57.01 | 49.29 | 73.94 | 81.35 | 72.47 |
| 365d | 60.68 | 63.59 | 70.51 | 51.51 | 71.96 | 80.66 | 73.55 |

We can see that although all classifiers present promising results, some perform better than the others. In general, all classifiers show better results for the longer the time window. This can

be due to the fact that for those datasets there is less class imbalance. Moreover, sensitivity results also seem to improve for the longer time windows and the opposite happens for the specificity results. Thus, for longer time windows the classifier performance improves and sensitivity and specificity metrics tend to become more balanced.

The SVM G classifier seems to be the least capable to predict the need for NIV, having the lowest AUC results and specificity lower than 50% for the models regarding the 90 and 180 days time windows. Then follow the DT, kNN and SVM P classifiers, all presenting better results, as well as reaching AUC's over 70% for the 365 days time window. However, the best classifiers seem to be NB, RF and LR, especially RF reaching an AUC of approximately 90% for the last time window. As these last three classifiers stand out, for the next tests only these are used.

Comparing our baseline results with those obtained by the previous work done by Carreiro *et al.* (2015), we can see that there is a major improvement in the results. Their best results for each time windows were 81.36%, 78.93%, and 79.98%, while ours were 80.74%, 85.27%, and 89.20% respectively (for the 90, 180 and 365 days datasets). However, while there are significant changes in the AUC results, the greater differences are in the sensitivity and specificity metrics, where they had a greater imbalance between the two, achieving sensitivities as lower as 15% and specificities as high as 97%. This shows that the classifiers were very good at predicting the negative instances but performed poorly in predicting the positive instances. In our results, although not achieving as high results in specificity, our sensitivities are considerably higher. This results in a greater balance between the two, therefore meaning that the classifiers have a similar performance when predicting both classes.

After building the baseline models, we then follow by training new models using only the features selected by the FSE for each time window. Table 3.6 comprises the results for the baseline models and the results with feature selection (using the FSE approach).

By looking at the results we can see that the models that use all features available tend to have better results than the ones using the set of features selected by the FSE. The better performance by the baseline models can be explained by the fact that the number of observations is high enough to handle the number of features available in data. FS methods usually have better results in situations where we have a high number of features for a small number of observations.

Although the results using all features are better than the ones using the FSE approach, we still performed the Wilcoxon Signed-Ranks Test for Paired Samples with 0.05 significance to see whether or not these differences are statistically significant. The test yielded a p-value of 0.0284, meaning that the differences are indeed significant. Therefore, the models using FS do not improve baseline models and should not be used.

3. TIME INDEPENDENT PROGNOSTIC MODELS

Table 3.6: AUC, Sensitivity and Specificity results for the prognostic models for the 90, 180 and 365 days Time Windows using Feature Selection. Orig is the original dataset, FS is the dataset with features selected by Feature Selection Ensemble (FSE). The classifiers are: Naïve Bayes (NB), Random Forest (RF) and Linear Regression (LR).

| | | AUC | | | Sensitivity | | | Specificity | | |
|------|------|-------|--------------|-------|-------------|-------|-------|-------------|-------|-------|
| | | NB | RF | LR | NB | RF | LR | NB | RF | LR |
| 90d | Orig | 79.84 | 80.74 | 79.15 | 69.87 | 68.55 | 69.41 | 74.99 | 77.6 | 74.78 |
| | FS | 79.05 | 77.48 | 79.29 | 76.06 | 71.34 | 74.78 | 67.25 | 68.23 | 69.98 |
| 180d | Orig | 80.57 | 85.27 | 79.68 | 73.22 | 70.73 | 72.3 | 73.94 | 81.35 | 72.47 |
| | FS | 77.51 | 82.02 | 78.21 | 74.61 | 74.5 | 73.77 | 66.36 | 73.45 | 68.5 |
| 365d | Orig | 82.25 | 89.20 | 80.21 | 78.41 | 82.07 | 73.85 | 71.96 | 80.66 | 73.55 |
| | FS | 79.91 | 88.95 | 80.46 | 74.63 | 82.07 | 75.02 | 70.77 | 80.9 | 72.21 |

Regarding missing value imputation (MVI) we tested two approaches: using Mean Imputation (MI) and then using Last Information Carried Forward (LOCF). The results for these tests, together with the results for the baseline models are presented in Table 3.7.

Table 3.7: AUC, Sensitivity, and Specificity results for the prognostic models for the 90, 180 and 365 days Time Windows using Missing Value Imputation. Orig is the original dataset, Mean MVI is the dataset imputed with Mean Imputation and LOCF MVI is the dataset imputed with Last Observation Carried Forward Imputation. The classifiers are: Naïve Bayes (NB), Random Forest (RF) and Linear Regression (LR).

| | | AUC | | | Sensitivity | | | Specificity | | |
|------|----------|-------|--------------|-------|-------------|-------|-------|-------------|-------|-------|
| | | NB | RF | LR | NB | RF | LR | NB | RF | LR |
| 90d | Orig | 79.84 | 80.74 | 79.15 | 69.87 | 68.55 | 69.41 | 74.99 | 77.6 | 74.78 |
| | Mean MVI | 79.28 | 80.63 | 79.3 | 70.74 | 65.58 | 69.55 | 74.35 | 79.18 | 74.84 |
| | LOCF MVI | 80.38 | 81.98 | 79.61 | 72.27 | 69.33 | 69.37 | 73.55 | 79.38 | 74.44 |
| 180d | Orig | 80.57 | 85.27 | 79.68 | 73.22 | 70.73 | 72.30 | 73.94 | 81.35 | 72.47 |
| | Mean MVI | 80.10 | 84.05 | 79.83 | 74.28 | 67.57 | 72.45 | 71.36 | 82.4 | 72.79 |
| | LOCF MVI | 81.86 | 87.29 | 80.66 | 72.69 | 72.93 | 71.45 | 75.94 | 84.13 | 74.09 |
| 365d | Orig | 82.25 | 89.20 | 80.21 | 78.41 | 82.07 | 73.85 | 71.96 | 80.66 | 73.55 |
| | Mean MVI | 83.49 | 87.40 | 80.21 | 74.18 | 78.97 | 73.84 | 77.02 | 80.29 | 73.54 |
| | LOCF MVI | 83.93 | 91.50 | 82.70 | 78.00 | 83.41 | 75.05 | 74.54 | 84.63 | 74.95 |

When using Mean Imputation we can see that for the NB and RF models the AUC results are slightly worse than the ones in baseline models and the LR results are slightly higher. Regarding

the models using LOCF, we can see that they are always better than the baseline results. This goes with our expectations. However, to determine the statistical significance of these differences, we followed the same approach as with the feature selection tests and performed the Wilcoxon Signed-Ranks Test for Paired Samples with 0.05 significance.

We compared each of the imputation models with the baseline models. For the MI models, the test resulted in a p-value of 0.3743 meaning there is no statistical difference between the models. As for the LOCF models, the test yielded a p-value of 0.0077 meaning the differences between the classifiers are indeed significant. Therefore, the models using LOCF imputation method should be used instead of the baseline methods.

Finally, Table 3.8 sows the best results achieved for the baseline models among all the tests performed.

Table 3.8: Best results achieved for baseline models using the patients current condition.

| | Sensitivity | Specificity | AUC |
|------|-------------|-------------|-------|
| 90d | 69.33 | 79.38 | 81.98 |
| 180d | 72.93 | 84.13 | 87.29 |
| 365d | 83.41 | 84.63 | 91.50 |

3.2 Using a set of Snapshots

In this section we address the following problem: "Given a set of consecutive patient evaluations, can we predict the need for NIV, in a time window of k days after the last evaluation?". An illustration of this problem is illustrated in Figure 3.3.

We follow the same pipeline used in the last section. However, instead of using only the patient's current condition to predict NIV, we use information about the patient's medical history (follow-up), to study whether it helps the models performance. In theory, the more information we have the best should be the prediction although in practice this is not always the case.

3.2.1 Creating Learning Instances

In order to create the learning instances to address this problem we use the snapshots created in Section 3.1.1. We take N consecutive patient evaluations and repeat the temporal features of each snapshot along the columns. Then, we set the value of the Evolution class to be equal

3. TIME INDEPENDENT PROGNOSTIC MODELS

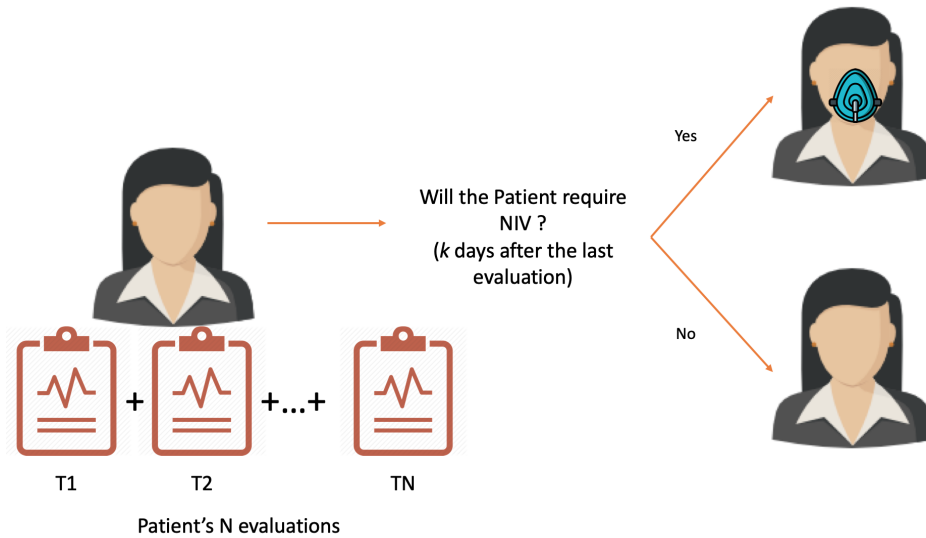


Figure 3.3: Problem Formulation: Given a set of N consecutive patient evaluations, can we predict the need for Non-Invasive Ventilation (NIV) k days after the last evaluation?

to the one in the last observation. Figure 3.4 shows the format of the learning examples using multiple snapshots.

| REF | Att1 T1 | Att2 T1 | ... | Att1 Tn | Att2 Tn | Evolution |
|-----|---------|---------|-----|---------|---------|-----------|
| 1 | 98 | 54 | ... | 93 | 48 | 0 |
| 2 | 76 | 47 | ... | 75 | 44 | 1 |
| ... | ... | ... | ... | ... | ... | ... |
| p | 91 | 50 | ... | 87 | 47 | 0 |

Figure 3.4: Example of learning examples using multiple snapshots.

Given that the number of snapshots per patient is less than four for all time windows, we chose to only create learning instances using two and three consecutive evaluations, as using more than that would lead to datasets with a small number of instances. Statistics and Class distribution for these datasets are presented in Table 3.9.

As in Section 3.1.1, the number of snapshots, as well as the number of patients, are increasingly lower as we increase the time windows. The number of snapshots per patient also decreases as before. However, we can see that for the same time window, the higher the number of evaluations we use, the higher the number of snapshots per patient. This is due to the fact that patients with fewer observations will usually be progressively left out as they do not have the

Table 3.9: Statistics and Class distribution for time windows of k=90,180, and 365 days.

| | 90 days | 180 days | 365 days |
|----------------------|---------------|---------------|---------------|
| 2 TP | | | |
| Nr of Snapshots | 2312 | 2191 | 1983 |
| Nr of Patients | 607 | 591 | 545 |
| Snapshots p/ Patient | 3.81 | 3.7 | 3.65 |
| Evolution (E = 1) | 357 (15.43%) | 588 (26.84%) | 902 (45.49%) |
| No Evolution (E= 0) | 1956 (84.57%) | 1603 (73.16) | 1081 (54.51%) |
| 3 TP | | | |
| Nr of Snapshots | 1706 | 1600 | 1438 |
| Nr of Patients | 434 | 415 | 381 |
| Snapshots p/ Patient | 3.93 | 3.86 | 3.77 |
| Evolution (E = 1) | 227 (13.31%) | 389 (24.31%) | 614 (42.70%) |
| No Evolution (E= 0) | 1439 (86.69%) | 1211 (75.68%) | 824 (57.30%) |

necessary number of evaluations. Thus the patients that stay in the analysis are the ones with a higher number of observations. Regarding class distribution, similar to the other datasets, we see that the proportion of positive instances increases for the longer time windows. Moreover, the higher the number of snapshots we use to build learning instances, the more imbalanced the dataset is, as the patients with more observations tend to progress slower, thus evolving to need NIV later.

3.2.2 Learning the Predictive Models

To learn the predictive models we followed the same pipeline used in Section 3.1.2. However, rather than using datasets with only the current condition of the patient we use datasets containing the patient’s clinical history, using consecutive observations of the same patient for that effect (patient follow-up).

We also tested the effects of FS and MVI on the performance of the classifiers. For the FS models, we created datasets using only the features selected by the FSE for the baseline models for each time point. As for MVI, we used only the LOCF approach, since it showed to improve results on the baseline models.

The use of multiple time points in learning instances also presents an opportunity to study if the models have better performance when using the value of the tests in each appointment or when using the differences between appointments. Thus, we decided to transform our data using temporal aggregation in order to create a set of temporal features that represent the differences between the two appointments. For the aggregation we use two approaches: the first consists in computing the numerical difference between the first and second appointments, while

3. TIME INDEPENDENT PROGNOSTIC MODELS

the second consists in analyzing the nominal variation between appointments to then create a categorical feature that can have three possible values: U - Up, D - Down, and N - No Change. The transformation process is presented in Figure 3.5. The new sets of features (numerical and categorical) can then be used to create new learning instances to be feed to the classifiers.

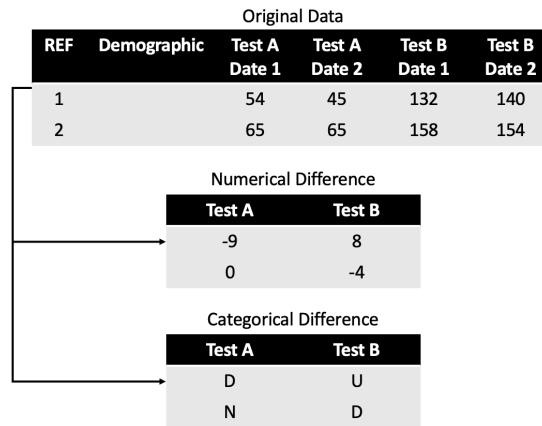


Figure 3.5: Example of transformation using temporal aggregation.

We choosed to proceed with NB, RF, and LR as they proved to be the best classifiers to predict NIV in the baseline tests. We trained the classifiers using a 5 x 10-fold CV scheme and performed a grid search to find the best parameters for the selected classifiers.

3.2.3 Results and Conclusions

We trained the models with data containing information of 2 and 3 time points (clinical history) with the intent to predict the need for NIV within a time window of $k = 90, 180,$ and 365 days. The results for the baseline models using a set of consecutive observations are presented in Table 3.10.

Looking at the table we can see that the models using two or three time points (TP) are very similar in performance, meaning that there is no improvement in performance in using three rather than two TP. Moreover, in comparison with the models using only the patient's current condition, we can see that although the results are similar, the models using only the current condition seem to perform better. Therefore, we can assume that the models did not benefit using the patient's clinical history over the patient's current condition.

We then built a new set of models to check if FS was beneficial to the classifier performance. We used the set of selected features obtained for the baseline models in Section 3.1.2 for each time windows and repeated those features for each time point (the ones selected as the best to

Table 3.10: AUC, Sensitivity and Specificity results for the prognostic models using 2 and 3 time points (TP) for each time windows of k=90,180,365 days. The classifiers are: Naïve Bayes (NB), Random Forest (RF) and Linear Regression (LR).

| | | AUC | | | Sensitivity | | | Specificity | | |
|------|------|-------|--------------|-------|-------------|-------|-------|-------------|-------|-------|
| | | NB | RF | LR | NB | RF | LR | NB | RF | LR |
| 90d | | | | | | | | | | |
| | 2 TP | 78.39 | 78.94 | 76.93 | 76.36 | 70.64 | 70.76 | 67.09 | 73.22 | 70.72 |
| | 3 TP | 77.88 | 79.42 | 76.97 | 72.33 | 67.58 | 66.7 | 70.63 | 74.79 | 74.06 |
| 180d | | | | | | | | | | |
| | 2 TP | 79.46 | 83.84 | 79.88 | 78.84 | 71.53 | 72.62 | 66.54 | 78.69 | 73.64 |
| | 3 TP | 79.34 | 83.45 | 80.03 | 76.86 | 71.41 | 71.36 | 68.46 | 78.25 | 75.26 |
| 365d | | | | | | | | | | |
| | 2 TP | 82.22 | 88.33 | 79.56 | 80.89 | 79.47 | 73.06 | 68.53 | 79.74 | 74.23 |
| | 3 TP | 82.38 | 88.57 | 79.08 | 79.58 | 78.99 | 71.04 | 70.73 | 80.92 | 76.14 |

predict, using the patient’s current condition). The results for these models are presented in Table 3.11.

The results show that in most cases FS does not improve the classifier’s performance for these cases but rather worsens it. The same outcome was obtained for the models using the patient’s current condition, and the explanation for those results is the same regarding these last ones: the number of instances from which the classifiers are trained is sufficient to handle the number of features available in the dataset with all features. We can see however that the FS models seem to have a greater balance between sensitivity and specificity than the models without FS.

As before, as the differences between the models are not substantial we performed the Wilcoxon Signed Rank Test for paired samples with 0.05 significance to assess the statistical significance of those differences. The test resulted in a p-value of 0.0311, meaning the differences are statistically significant. Therefore, the models using feature selection should not be used.

In the last section two MVI approaches were tested: Mean Imputation and LOCF imputation. The first showed no improvements in classifier performance, however, the latter did. Thus, in this section, we decided to test only the last approach to see if it also improves performance for these new models. The results for this test can be found in Table 3.12.

Once more we can see that this imputation method usually improves the models’ performance. The WSRT supports this claim since it yielding a p-value of 0.0033 (statistically significant). These models show even more promising results than the baseline, reaching an AUC of 91.40% for the time window of 365 days using the RF classifier. This results highly approximate the results obtained for the same model using only the current condition. Therefore, using one or the other should result in similar outcomes.

3. TIME INDEPENDENT PROGNOSTIC MODELS

Table 3.11: AUC, Sensitivity and Specificity results for the prognostic models using 2 or 3 patient time points (TP) for the 90, 180 and 365 days Time Windows using Feature Selection. Orig is the original dataset, FS is the dataset with features selected by Feature Selection Ensemble (FSE). The classifiers are: Naïve Bayes (NB), Random Forest (RF) and Linear Regression (LR).

| | | | AUC | | | Sensitivity | | | Specificity | | |
|------|------|--|-------|--------------|-------|-------------|-------|-------|-------------|-------|-------|
| | | | NB | RF | LR | NB | RF | LR | NB | RF | LR |
| 90d | | | | | | | | | | | |
| 2 TP | Orig | | 78.39 | 78.94 | 76.93 | 76.36 | 70.64 | 70.76 | 67.09 | 73.22 | 70.72 |
| | FS | | 77.52 | 77.16 | 77.84 | 73.89 | 71.32 | 73.56 | 67.57 | 67.79 | 69.73 |
| 3 TP | Orig | | 77.88 | 79.42 | 76.97 | 72.33 | 67.58 | 66.70 | 70.63 | 74.79 | 74.06 |
| | FS | | 78.28 | 77.95 | 79.26 | 76.3 | 69.34 | 71.89 | 65.99 | 70.37 | 72.06 |
| 180d | | | | | | | | | | | |
| 2 TP | Orig | | 79.46 | 83.84 | 79.88 | 78.84 | 71.53 | 72.62 | 66.54 | 78.69 | 73.64 |
| | FS | | 77.86 | 81.59 | 79.10 | 74.18 | 74.56 | 74.22 | 67.05 | 71.75 | 70.53 |
| 3 TP | Orig | | 79.34 | 83.45 | 80.03 | 76.86 | 71.41 | 71.36 | 68.46 | 78.25 | 75.26 |
| | FS | | 78.36 | 82.20 | 79.84 | 75.73 | 75.53 | 71.98 | 67.23 | 72.65 | 71.89 |
| 365d | | | | | | | | | | | |
| 2 TP | Orig | | 82.22 | 88.33 | 79.56 | 80.89 | 79.47 | 73.06 | 68.53 | 79.74 | 74.23 |
| | FS | | 80.55 | 88.06 | 81.60 | 75.54 | 81.44 | 74.50 | 71.49 | 77.65 | 73.47 |
| 3 TP | Orig | | 82.38 | 88.57 | 79.08 | 79.58 | 78.99 | 71.04 | 70.73 | 80.92 | 76.14 |
| | FS | | 81.03 | 87.43 | 81.90 | 78.50 | 80.39 | 73.13 | 70.07 | 76.21 | 73.83 |

Finally, we test the use of temporal aggregation of features using the datasets containing two time points. The results are presented in Table 3.13.

Table 3.12: AUC, Sensitivity and Specificity results for the prognostic models using 2 or 3 time points (TP) for the 90, 180 and 365 days Time Windows using Missing Value Imputation. Orig is the original dataset and MVI is the dataset imputed with Last Observation Carried Forward Imputation. The classifiers are: Naïve Bayes (NB), Random Forest (RF) and Linear Regression (LR).

| | | | AUC | | | Sensitivity | | | Specificity | | |
|------|------|--|-------|--------------|-------|-------------|-------|-------|-------------|-------|-------|
| | | | NB | RF | LR | NB | RF | LR | NB | RF | LR |
| 90d | | | | | | | | | | | |
| 2 TP | Orig | | 78.39 | 78.94 | 76.93 | 76.36 | 70.64 | 70.76 | 67.09 | 73.22 | 70.72 |
| | MVI | | 78.87 | 80.00 | 78.17 | 75.69 | 67.90 | 69.47 | 67.65 | 76.20 | 73 |
| 3 TP | Orig | | 77.88 | 79.42 | 76.97 | 72.33 | 67.58 | 66.70 | 70.63 | 74.79 | 74.06 |
| | MVI | | 78.30 | 79.38 | 78.69 | 75.07 | 66.61 | 68.72 | 68.26 | 76.21 | 75.66 |
| 180d | | | | | | | | | | | |
| 2 TP | Orig | | 79.46 | 83.84 | 79.88 | 78.84 | 71.53 | 72.62 | 66.54 | 78.69 | 73.64 |
| | MVI | | 80.34 | 85.76 | 80.62 | 77.93 | 72.55 | 70.99 | 68.51 | 81.91 | 74.87 |
| 3 TP | Orig | | 79.34 | 83.45 | 80.03 | 76.86 | 71.41 | 71.36 | 68.46 | 78.25 | 75.26 |
| | MVI | | 80.58 | 85.19 | 81.43 | 76.14 | 70.54 | 72.13 | 70.45 | 80.74 | 76.3 |
| 365d | | | | | | | | | | | |
| 2 TP | Orig | | 82.22 | 88.33 | 79.56 | 80.89 | 79.47 | 73.06 | 68.53 | 79.74 | 74.23 |
| | MVI | | 84.30 | 91.51 | 82.00 | 79.73 | 81.75 | 74.83 | 73.78 | 84.81 | 76.1 |
| 3 TP | Orig | | 82.38 | 88.57 | 79.08 | 79.58 | 78.99 | 71.04 | 70.73 | 80.92 | 76.14 |
| | MVI | | 84.96 | 91.40 | 81.77 | 78.01 | 81.11 | 73.19 | 76.70 | 84.05 | 78.25 |

Overall, the models using temporal aggregation do not show great improvements in performance when compared to the baseline models, using two TP. For the 180 and 365 days time windows, the models using NB and LR classifiers, seem to benefit from using temporal aggregation, especially when adding the categorical features. The exclusive use of the new features is, however, not advised, since the models show a high decay in the results. Still, the best results are held by the RF classifier, which does not benefit from the use of the new features. Therefore, as the models do not show considerable improvements, we chose to use the baseline models instead, as they have fewer features and therefore end up in simpler models, which should always be preferred.

To close this chapter, Table 3.14 presents the best possible results in our analysis for the baseline models both using the current condition of the patient (1 TP) and using the clinical history of the patient (2 and 3 TP).

3. TIME INDEPENDENT PROGNOSTIC MODELS

Table 3.13: AUC, Sensitivity and Specificity results for the prognostic models using temporal aggregation. Legend: Orig \rightarrow Original dataset (Temporal Features), Dyn \rightarrow Temporal + Categorical Numerical Features, DynCat \rightarrow Temporal + Categorical Features, DynNum \rightarrow Temporal + Numerical Features, DynOnly \rightarrow Categorical Numerical Features, DynCatOnly \rightarrow Categorical Features, DynNumOnly \rightarrow Numerical Features. The classifiers are: Naïve Bayes (NB), Random Forest (RF) and Linear Regression (LR).

| | | AUC | | | Sensitivity | | | Specificity | | |
|----------|------------|-------|--------------|-------|-------------|-------|-------|-------------|-------|-------|
| | | NB | RF | LR | NB | RF | LR | NB | RF | LR |
| 90 days | Orig | 78.39 | 78.94 | 76.93 | 76.36 | 70.64 | 70.76 | 67.09 | 73.22 | 70.72 |
| | Dyn | 77.75 | 76.76 | 75.51 | 67.06 | 61.96 | 64.37 | 72.91 | 74.55 | 74.06 |
| | DynCat | 78.21 | 77.12 | 75.25 | 73.50 | 65.21 | 64.31 | 68.24 | 73.55 | 73.25 |
| | DynNum | 77.50 | 77.34 | 76.67 | 70.20 | 66.44 | 69.30 | 70.25 | 73.12 | 72.28 |
| | DynOnly | 68.74 | 70.25 | 67.83 | 56.53 | 54.62 | 62.52 | 70.87 | 73.19 | 63.72 |
| | DynCatOnly | 67.37 | 70.32 | 67.85 | 68.24 | 56.08 | 65.77 | 55.45 | 71.04 | 60.82 |
| | DynNumOnly | 67.56 | 71.45 | 69.34 | 55.69 | 56.41 | 64.09 | 66.62 | 72.56 | 63.42 |
| 180 days | Orig | 79.46 | 83.84 | 79.88 | 78.84 | 71.53 | 72.62 | 66.54 | 78.69 | 73.64 |
| | Dyn | 79.65 | 80.18 | 78.32 | 73.91 | 64.46 | 65.85 | 70.14 | 78.08 | 76.28 |
| | DynCat | 80.07 | 80.89 | 78.50 | 76.50 | 66.39 | 66.39 | 68.68 | 78.23 | 75.32 |
| | DynNum | 79.74 | 81.24 | 80.06 | 76.70 | 68.44 | 71.60 | 68.62 | 76.99 | 74.17 |
| | DynOnly | 71.50 | 74.07 | 71.42 | 58.37 | 59.18 | 65.82 | 72.89 | 75.18 | 66.06 |
| | DynCatOnly | 71.14 | 74.28 | 71.09 | 70.48 | 61.05 | 68.10 | 60.61 | 73.10 | 63.97 |
| | DynNumOnly | 71.24 | 75.92 | 72.70 | 65.00 | 60.48 | 67.21 | 64.19 | 75.38 | 66.61 |
| 365 days | Orig | 82.22 | 88.33 | 79.56 | 80.89 | 79.47 | 73.06 | 68.53 | 79.74 | 74.23 |
| | Dyn | 82.75 | 83.95 | 82.65 | 72.42 | 73.28 | 73.41 | 76.45 | 77.22 | 76.15 |
| | DynCat | 83.08 | 85.23 | 82.86 | 80.60 | 75.79 | 73.66 | 69.34 | 77.76 | 76.24 |
| | DynNum | 83.21 | 85.71 | 83.34 | 74.41 | 76.08 | 74.39 | 76.26 | 77.58 | 76.21 |
| | DynOnly | 74.53 | 78.57 | 75.77 | 56.43 | 68.16 | 68.94 | 77.93 | 72.75 | 68.70 |
| | DynCatOnly | 76.34 | 79.92 | 75.79 | 74.48 | 70.89 | 69.96 | 62.65 | 73.77 | 67.46 |
| | DynNumOnly | 74.38 | 81.14 | 75.83 | 51.09 | 71.93 | 69.65 | 79.80 | 73.95 | 67.70 |

The aim of Section 3.2 was to check whether or not the use of clinical history (follow-up) was better to predict the need for NIV within a predefined time window compared to the use of the patient’s current condition. By analyzing Table 3.14 we observe that overall there is no clear benefits in using models which are trained with the patients’ clinical history.

Table 3.14: Best results obtained for the baseline models using 1 TP (current condition) as well as using 2 and 3 TP (clinical history).

| | | Sensitivity | Specificity | AUC |
|------|-----|--------------------|--------------------|------------|
| 90d | | | | |
| | 1TP | 69.33 | 79.38 | 81.98 |
| | 2TP | 67.90 | 76.20 | 80.00 |
| | 3TP | 67.58 | 74.79 | 79.43 |
| 180d | | | | |
| | 1TP | 72.93 | 84.13 | 87.29 |
| | 2TP | 72.55 | 81.91 | 85.76 |
| | 3TP | 70.54 | 80.74 | 85.19 |
| 365d | | | | |
| | 1TP | 83.41 | 84.63 | 91.50 |
| | 2TP | 81.75 | 84.81 | 91.51 |
| | 3TP | 81.11 | 84.05 | 91.40 |

Chapter 4

Progression Groups

Given the heterogeneous nature of ALS, the progression rate is highly variable across all patients. Moreover, patients with different progression rates usually have different prognosis. In this context, when creating prognostic models with such different patients, we risk that our models learn to predict well a subset of patients and for the others, only be guesses the outcome.

In this section, we explore our first patient stratification approach, in which we stratify the patients according to their progression rate (see Section 2.4.1). We create three Progression Groups: Slow, Neutral and Fast, and use patients in these groups to build predictive models specialized for each group. The groups are created from a cohort of 1220 patients using the information at the time of disease onset and the ALS-FRS-R scale at first appointment. With that, we compute the progression rate of the patient using the formula presented in Section 2.4.1.

Only 1093 of 1220 patients (89.6%) could be used for analysis, as the other 127 patients lacked at least one of the informations needed to compute the progression rate. Using the progression rate of the selected patients, we obtained the distribution presented in Figure 4.1.

The higher patients' progression rate are, the faster the patients progress, while lower progression rates are usually associated with a slower disease progression. Following consensual clinical insight, we decided to stratify the patients in three disease progression groups. The 25% of the patients with higher progression rates were grouped together and labeled as Fast Progressors. The 25% of the patients with lower progression rates were also grouped together to create the Slow Progressors group. The remaining 50%, with an average progression, were grouped together and called Neutral Progressors. This totalizes 271 Slow progressors, 552 Neutral progressors and 270 Fast progressors.

4. PROGRESSION GROUPS

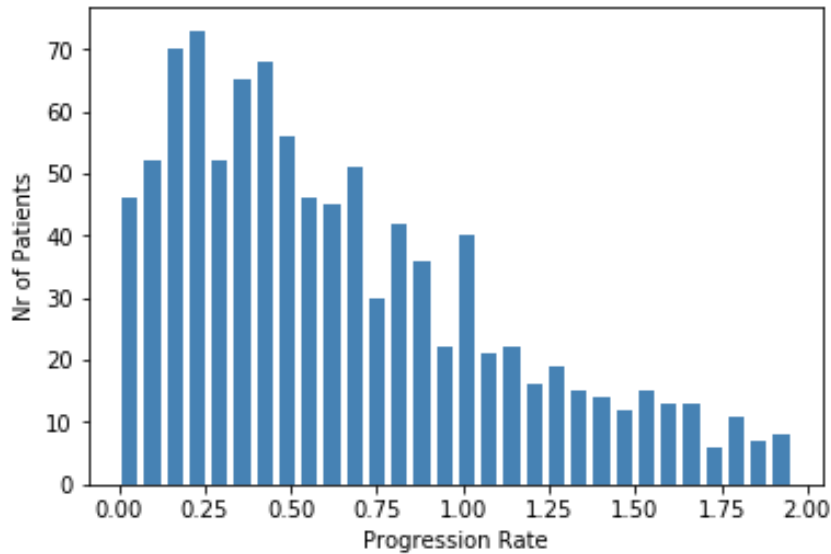


Figure 4.1: Progression Rate Distribution among all patients.

4.1 Single Snapshot Prediction

In this section, we revisit the problem addressed in Chapter 3, adding information about the patient's progression groups. Our aim is to create specialized models for each progression group to investigate if by stratifying the patients according to this criteria, we can improve performance. Figure 4.2 shows the new problem formulation. The new specialized models enable us to answer the following question: "Given that we know the patient's progression group, as well as their current condition, can we predict the patient's need for NIV, k days after his/her last appointment?".

Regarding to the pipeline used, we adapted the pipeline used in the previous approach, to accommodate the progression group information. Figure 4.3 shows the updated pipeline.

4.1.1 Creating Learning Instances

The process of creating learning instances for the specialized models consists in extracting all snapshots of each patient and using them to create three new datasets, one for each progression group. Thus, each new dataset is a subset of the original dataset containing the snapshots corresponding to the subset of patients in each group. Table 4.1 presents Statistics and the Class distribution for each dataset.

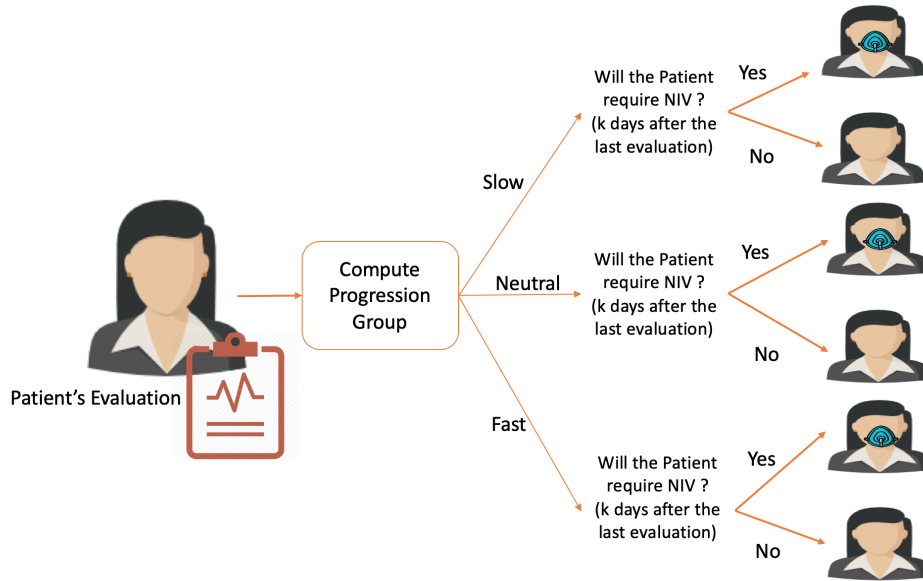


Figure 4.2: Problem Reformulation using Progression Groups: Knowing the Patient current state, as well as their progression group can we predict the need for Non-Invasive Ventilation k days after, using group specific models?

Table 4.1: Statistics and Class distribution for time windows of $k=90,180$, and 365 , for each progression group.

| k | 90 | 180 | 365 |
|----------------------|---------------|---------------|--------------|
| Slow Progressors | | | |
| Nr of Snapshots | 1242 | 1191 | 1090 |
| Nr of Patients | 215 | 210 | 200 |
| Snapshots p/ Patient | 5.78 | 5.67 | 5.45 |
| Evolution (E=1) | 88 (7.09%) | 163 (13.69%) | 269 (24.68%) |
| No Evolution (E= 0) | 1154 (92.91%) | 1028 (86.31%) | 821 (75.32%) |
| Neutral Progressors | | | |
| Nr of Snapshots | 1459 | 1390 | 1278 |
| Nr of Patients | 441 | 425 | 399 |
| Snapshots p/ Patient | 3.31 | 3.27 | 3.20 |
| Evolution (E=1) | 328 (22.48%) | 527 (37.91%) | 801 (62.68%) |
| No Evolution (E=0) | 1131 (77.52%) | 863 (62.09%) | 477 (37.32%) |
| Fast Progressors | | | |
| Nr of Snapshots | 384 | 348 | 311 |
| Nr of Patients | 171 | 158 | 148 |
| Snapshots p/ Patient | 2.24 | 2.20 | 2.10 |
| Evolution (E=1) | 131 (34.11%) | 193 (55.46%) | 238 (76.53%) |
| No Evolution (E=0) | 253 (65.89%) | 155 (44.54%) | 73 (23.47%) |

4. PROGRESSION GROUPS

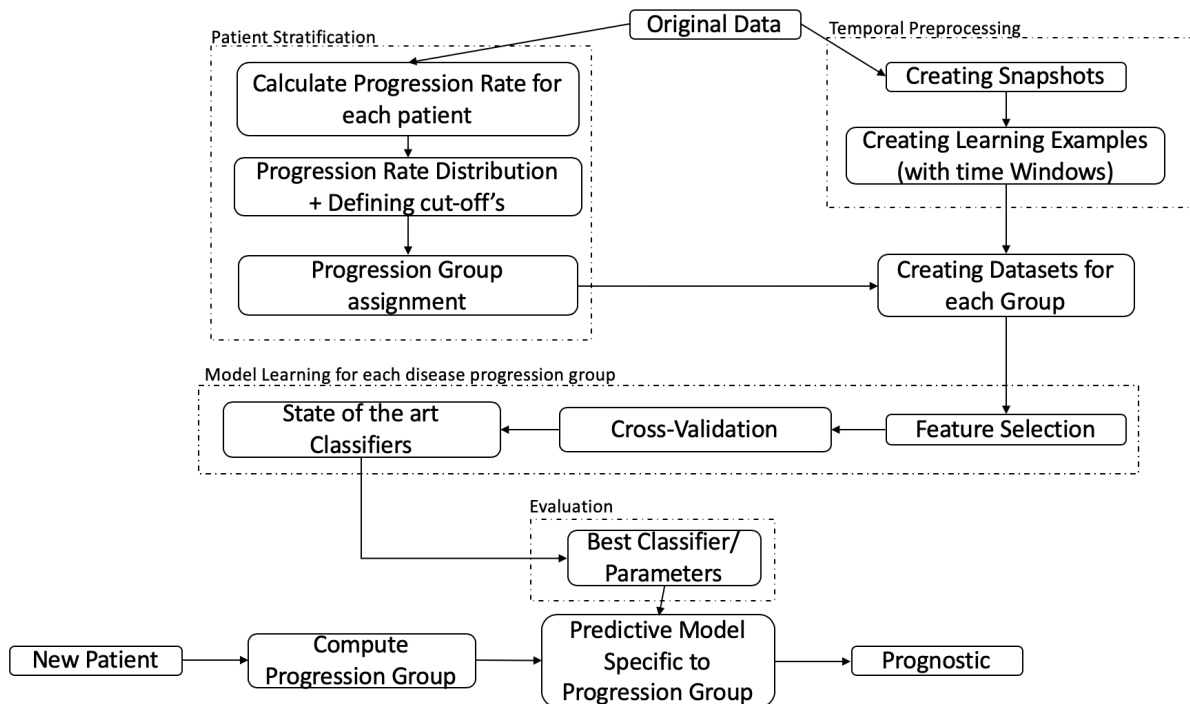


Figure 4.3: Workflow of the proposed methodology for ALS prognostic prediction using patient snapshots and progression groups. Original data is preprocessed in order to create patient snapshots, that are then used to create the learning instances. At the same time, the original data is used to create the progression groups. The progression groups and learning instances are then merged to create separate sets of data for each group. The models are then built using a stratified 5 x 10-fold cross-validation scheme. The models are then evaluated and the best parameters are chosen. After the final model is complete when a new patient arrives at the consult, their progression group is computed and their data is given to the specific predictive model that should output a prediction for the problem being addressed.

We can see that similarly to the original dataset, the number of patients, the number of snapshots as well as the number of snapshots per patient, all decrease as the time windows increase. However, these number differ considerably for each of the groups. The major differences are between the Slow and Fast progressors. This is expected given that they should be the most dissimilar groups. The Slow progressors show a higher number of Snapshots per patient, more than the double regarding the Fast progressors. This can be explained by the fact that these patients have a slower progression, thus usually having a better prognosis and consequently a higher survival. In this context, it is expected that they attend a higher number of appointments, having a longer clinical history, thus having a greater contribute in terms of patient Snapshots for our dataset. In the case of Fast progressors, their disease progression is so fast that they only have two or three appointments, therefore, the number of snapshots and even the number of patients

that have enough information for this analysis tend to be lower. Finally, the Neutral progressors are more representative of the average of the patients, and thus are those with statistics closer to the original dataset.

Regarding to class distribution, we can also see differences between the groups. Although they all follow the same pattern across the time windows (the longer the selected time window, the higher the number of positive instances), there are clear differences between them. The Slow progressors present the lower proportion of positive instances in all time windows, and the fast show higher proportions of negative instances than the other groups as well as the baseline datasets. Thus, it seems that the faster the disease progresses, the more urgent the need for NIV is (which actually makes sense). Moreover, we can see that while in the original dataset the number of positive instances was always lower than the number of negative instances, this does not hold true for three of these datasets: Fast progressors in time windows of 180 and 365 days, and Neutral progressors in the 365 days dataset. As for the class imbalance it becomes opposite than the distribution for the baseline datasets.

4.1.2 Learning the predictive Models

After creating the datasets for each group according to the three selected time windows (90, 180, and 365 days), we then use each of them as input to learn the specialized models for each group. Each model is built using a 5 x 10-fold CV scheme as before and the metric selected to evaluate the models' performance is AUC.

Similarly to the analysis using the baseline dataset, we performed FS for each of the datasets using the FSE proposed by Pereira *et al.* (2018), as described in the Chapter 3. The goal is not only to study if using only the selected features improves the models' performance but also to investigate differences between the sets of features selected for each progression group. Table 4.2 presents the sets of selected features for each dataset.

4. PROGRESSION GROUPS

Table 4.2: Selected Features for each progression groups and each time window.

| Features | Slow Progressors | | | Neutral Progressors | | | Fast Progressors | | |
|-------------------------------|------------------|-------|-------|---------------------|-------|-------|------------------|-------|-------|
| | 90 d | 180 d | 365 d | 90 d | 180 d | 365 d | 90 d | 180 d | 365 d |
| Gender | | | | | | | | | |
| Age at Onset | X | X | X | X | X | X | X | | X |
| BMI | X | X | X | X | X | X | X | X | X |
| Family History MND | | | | | | | | | |
| Disease Duration | X | X | X | X | X | X | | | |
| El Escorial Reviewed Criteria | | | | | | | | | |
| UMN vs LMN | | | | | | | | | |
| Onset form | X | | | | | | | | |
| c9orf72 | X | X | | | | | | | |
| ALS-FRS | X | X | X | X | X | X | | | |
| ALS-FRS-R | X | X | X | X | X | X | | | |
| ALS-FRSb | X | X | | X | X | X | | | |
| ALS-FRSsUL | X | X | | X | X | X | | | |
| ALS-FRSsLL | X | X | | X | X | X | | | |
| ALS-FRSr | X | X | | | | | | | |
| R | X | X | | X | X | X | | | |
| VC | X | X | X | X | X | X | X | X | X |
| FVC | X | X | X | X | X | X | X | X | X |
| MIP | X | X | X | X | X | X | | | X |
| MEP | X | X | X | X | X | X | | | X |
| P0.1 | X | X | | X | X | X | | | |
| SNIP | | | | | | | | | |
| PhrenMeanLat | X | X | | X | X | X | | | |
| PhrenMeanAmpl | X | X | | X | X | X | | | |
| Cervical Flexion | | | | | | | | | |
| Cervical Extension | | | | | | | | | |

There are clear differences between the selected features for each group. Slow progressors tend to need more features to build good prognostic models, while Fast progressors seem to rely on few features. Moreover, while for Slow progressors in longer time windows the number of selected features diminishes, in Fast progressors, this is reversed. In Neutral progressors, the set of selected features is the same for all time windows. The differences in the selected features between the progression groups can also be important in the clinician’s point of view, since knowing which tests and exams are more important for each type of patient, can save time and resources that can lead to a better prognosis.

In addition to learning models using the set of features selected by the FSE for each dataset, we also build models using datasets for each group using the features selected for the baseline dataset presented in the Chapter 3 to check which set of features is more helpful. Our intuition

is that the important features are group specific but it is worth checking.

For dealing with missing data we follow the LOCF methodology described before, as it has shown improvements on classifier performance in the models for the baseline datasets.

Finally, we compare the performance of specialized models to that of the baseline models using all patients to see how the models differ in predicting each of the progression groups. We look at how the baseline models predicts each instance from each group and compare them to the predictions using the specialized models for the group. What we want to know is: When we look at the patients in a specific group, how good is the baseline model?

The classifiers used were NB, LR and RF and the metric used to assess classifier performance is AUC.

4.1.3 Results and Conclusions

Table 4.3 presents the results for the specialized models for each progression group, across all the selected time windows of 90, 180 and 365 days.

Table 4.3: AUC, Sensitivity and Specificity results for the prognostic models for the 90, 180 and 365 days Time Windows, as well as for each Progression Group. The selected classifiers are: Naive Bayes (NB), Random Forest (RF) and Linear Regression (LR).

| | AUC | | | Sensitivity | | | Specificity | | |
|---------|--------------|--------------|-------|-------------|-------|-------|-------------|-------|-------|
| | NB | RF | LR | NB | RF | LR | NB | RF | LR |
| Slow | | | | | | | | | |
| 90d | 81.11 | 81.11 | 74.15 | 70.91 | 69.55 | 64.09 | 79.41 | 75.58 | 73.99 |
| 180d | 85.21 | 86.51 | 80.94 | 77.42 | 72.27 | 72.88 | 80.88 | 83.29 | 77.67 |
| 365d | 84.4 | 90.58 | 80.22 | 73.16 | 78.29 | 69.52 | 80.56 | 85.87 | 77.47 |
| Neutral | | | | | | | | | |
| 90d | 77.29 | 76.2 | 74.88 | 67.74 | 58.17 | 64.88 | 74.06 | 77.84 | 71.71 |
| 180d | 75.32 | 81.61 | 74.68 | 60.91 | 62.01 | 63.76 | 77.27 | 82.92 | 72.84 |
| 365d | 74.4 | 85.71 | 73.54 | 70.64 | 78.8 | 68.76 | 67.04 | 77.06 | 66.5 |
| Fast | | | | | | | | | |
| 90d | 72.69 | 71.82 | 74.94 | 63.66 | 51.3 | 61.37 | 74.39 | 76.68 | 77.39 |
| 180d | 71.48 | 81.23 | 70.56 | 65.28 | 70.57 | 68.81 | 68 | 75.61 | 61.55 |
| 365d | 65.6 | 79.41 | 65.02 | 64.96 | 74.37 | 68.15 | 57.26 | 71.23 | 51.23 |

By analyzing Table 4.3 we can see that the specialized models for the Slow Progressors show better results than the models for the other groups. Moreover, these results surpass the ones of baseline models using all patients. Contrary to this, the models for Fast progressors seem to be the worse. These results can be explained by the reduced number of learning instances in

4. PROGRESSION GROUPS

this group, which can be hampering for the models' performance. The results for the Neutral Progressors are the most similar to the baseline models, probably due to their greater contribution for those models. Our intuition that baseline models are specialized in Neutral Progressors and are not good at the others.

Table 4.4 shows the results for the models using feature selection. For most cases, FS does not improve the models' performance. Moreover, the sets of features selected specifically for each dataset seem to have more predictive power than the ones selected for the original datasets using all patients. The differences are however not major. Therefore, similarly to previous analysis, we performed the Wilcoxon Signed Rank Test for paired samples statistic test to check the statistical significance of the differences. When comparing the models without feature selection and the models using the features selected for the baseline models the test yielded a p-value of 0.0014 meaning that the differences are significant. Thus, the set of features selected for the baseline models should not be used to build the specialized models for each progression group. The test comparing the specialized models without feature selection and the specialized models with feature selection specific to each group yielded a p-value of 0.8288, meaning the differences are not statistically significant. Therefore, the features selected by for each specific progression group should be used in the creation of specialized models as the use of a smaller set of features usually results in simpler models, which are generally better.

Table 4.5 shows the results for the models learnt using the LOCF imputation method. As in previous results, the models using the selected imputation method seem to improve the models' performance. The Wilcoxon Signed Rank Test for paired samples for these models resulted in a p-value of 0.0001 meaning that the differences between approaches are high enough to justify the use of this imputation method, in comparison to the non-imputed models, when building these specialized models.

To compare how the baseline models behave when classifying each specific progression group, we built the classifiers using all instances labeled with the progression group (not used in the classifiers). Then we retrieved the predictions for each instance and computed the confusion matrix and AUC to retrieve the necessary information to better compare the two approaches. To lower the number of presented results, this comparison was only made for NB (Table 4.6).

4.1 Single Snapshot Prediction

Table 4.4: AUC, Sensitivity and Specificity results for the prognostic models for the 90, 180 and 365 days Time Windows, as well as for each Progression Group using Feature Selection. Orig is the original dataset, FS is the dataset with selected features for each progression group by the Feature Selection Ensemble (FSE), and FS Orig is the dataset using the features selected for the main models, by the FSE. The selected classifiers are: Naive Bayes (NB), Random Forest (RF) and Linear Regression (LR).

| | | AUC | | | Sensitivity | | | Specificity | | | |
|----------------|------|---------|--------------|--------------|--------------|-------|-------|-------------|-------|-------|-------|
| | | NB | RF | LR | NB | RF | LR | NB | RF | LR | |
| Slow | 90d | Orig | 81.11 | 81.11 | 74.15 | 70.91 | 69.55 | 64.09 | 79.41 | 75.58 | 73.99 |
| | | FS | 82.24 | 79.77 | 79.17 | 72.27 | 68.86 | 70.45 | 76.79 | 73.97 | 75.06 |
| | | Orig Fs | 80.60 | 76.48 | 77.91 | 77.05 | 70.91 | 71.59 | 73.29 | 69.84 | 72.69 |
| | 180d | Orig | 85.21 | 86.51 | 80.94 | 77.42 | 72.27 | 72.88 | 80.88 | 83.29 | 77.67 |
| | | FS | 84.78 | 85.72 | 82.50 | 75.58 | 74.11 | 73.50 | 79.73 | 80.64 | 78.79 |
| | | Orig Fs | 81.07 | 82.11 | 80.64 | 79.14 | 72.76 | 75.21 | 70.04 | 75.86 | 73.99 |
| | 365d | Orig | 84.40 | 90.58 | 80.22 | 73.16 | 78.29 | 69.52 | 80.56 | 85.87 | 77.47 |
| | | FS | 82.27 | 88.92 | 82.71 | 75.54 | 80.00 | 73.01 | 75.44 | 82.39 | 75.98 |
| | | Orig Fs | 82.24 | 88.66 | 82.77 | 75.61 | 79.33 | 73.38 | 75.42 | 81.78 | 76.22 |
| Neutral | 90d | Orig | 77.29 | 76.20 | 74.88 | 67.74 | 58.17 | 64.88 | 74.06 | 77.84 | 71.71 |
| | | FS | 76.34 | 74.94 | 77.43 | 68.66 | 62.87 | 69.57 | 71.18 | 72.25 | 72.36 |
| | | Orig Fs | 75.75 | 72.74 | 76.06 | 70.30 | 60.24 | 68.84 | 67.06 | 70.47 | 69.05 |
| | 180d | Orig | 75.32 | 81.61 | 74.68 | 60.91 | 62.01 | 63.76 | 77.27 | 82.92 | 72.84 |
| | | FS | 75.78 | 80.24 | 76.62 | 66.72 | 65.24 | 66.22 | 73.02 | 78.61 | 73.42 |
| | | Orig Fs | 72.42 | 77.78 | 71.89 | 67.13 | 60.46 | 65.88 | 64.26 | 77.71 | 66.56 |
| | 365d | Orig | 74.40 | 85.71 | 73.54 | 70.64 | 78.80 | 68.76 | 67.04 | 77.06 | 66.50 |
| | | FS | 75.27 | 83.93 | 75.43 | 59.48 | 68.39 | 68.29 | 79.45 | 82.01 | 71.78 |
| | | Orig Fs | 71.74 | 84.10 | 72.22 | 60.85 | 66.77 | 65.74 | 73.04 | 82.39 | 68.89 |
| Fast | 90d | Orig | 72.69 | 71.82 | 74.94 | 63.66 | 51.30 | 61.37 | 74.39 | 76.68 | 77.39 |
| | | FS | 72.62 | 72.43 | 70.45 | 75.88 | 62.44 | 64.12 | 58.58 | 70.12 | 65.06 |
| | | Orig Fs | 72.97 | 70.37 | 75.04 | 68.40 | 64.12 | 68.09 | 66.64 | 65.30 | 69.96 |
| | 180d | Orig | 71.48 | 81.23 | 70.56 | 65.28 | 70.57 | 68.81 | 68.00 | 75.61 | 61.55 |
| | | FS | 70.62 | 81.26 | 69.14 | 61.14 | 66.22 | 60.10 | 67.35 | 76.26 | 63.23 |
| | | Orig Fs | 70.47 | 79.73 | 70.02 | 59.59 | 67.56 | 62.80 | 71.61 | 74.58 | 65.29 |
| | 365d | Orig | 65.60 | 79.41 | 65.02 | 64.96 | 74.37 | 68.15 | 57.26 | 71.23 | 51.23 |
| | | FS | 69.57 | 77.52 | 63.00 | 56.13 | 65.71 | 59.08 | 72.60 | 72.33 | 55.07 |
| | | Orig Fs | 66.97 | 75.54 | 60.51 | 54.54 | 64.03 | 57.06 | 67.40 | 70.14 | 55.34 |

4. PROGRESSION GROUPS

Table 4.5: AUC, Sensitivity and Specificity results for the prognostic models for the 90, 180 and 365 days Time Windows, as well as for each Progression Group using Missing Value Imputation. Orig is the original dataset and MVI is the dataset with LOCF imputation. The selected classifiers are: Naive Bayes (NB), Random Forest (RF) and Linear Regression (LR).

| | | AUC | | | Sensitivity | | | Specificity | | | |
|----------------|------|------|--------------|--------------|--------------|-------|-------|-------------|-------|-------|-------|
| | | NB | RF | LR | NB | RF | LR | NB | RF | LR | |
| Slow | 90d | Orig | 81.11 | 81.11 | 74.15 | 70.91 | 69.55 | 64.09 | 79.41 | 75.58 | 73.99 |
| | | MVI | 84.19 | 83.66 | 79.63 | 72.27 | 69.32 | 72.50 | 80.59 | 79.17 | 75.53 |
| | 180d | Orig | 85.21 | 86.51 | 80.94 | 77.42 | 72.27 | 72.88 | 80.88 | 83.29 | 77.67 |
| | | MVI | 87.24 | 88.45 | 84.74 | 75.21 | 74.85 | 75.71 | 83.23 | 85.45 | 79.63 |
| | 365d | Orig | 84.4 | 90.58 | 80.22 | 73.16 | 78.29 | 69.52 | 80.56 | 85.87 | 77.47 |
| | | MVI | 87.70 | 92.30 | 85.93 | 72.34 | 78.81 | 75.32 | 84.90 | 87.67 | 82.00 |
| Neutral | 90d | Orig | 77.29 | 76.20 | 74.88 | 67.74 | 58.17 | 64.88 | 74.06 | 77.84 | 71.71 |
| | | MVI | 77.64 | 78.05 | 75.54 | 66.46 | 63.23 | 66.89 | 74.45 | 78.94 | 72.73 |
| | 180d | Orig | 75.32 | 81.61 | 74.68 | 60.91 | 62.01 | 63.76 | 77.27 | 82.92 | 72.84 |
| | | MVI | 76.96 | 83.72 | 76.75 | 61.40 | 63.11 | 65.16 | 79.33 | 86.05 | 74.81 |
| | 365d | Orig | 74.40 | 85.71 | 73.54 | 70.64 | 78.8 | 68.76 | 67.04 | 77.06 | 66.50 |
| | | MVI | 77.24 | 89.54 | 75.96 | 69.51 | 78.70 | 69.61 | 72.33 | 84.49 | 68.47 |
| Fast | 90d | Orig | 72.69 | 71.82 | 74.94 | 63.66 | 51.30 | 61.37 | 74.39 | 76.68 | 77.39 |
| | | MVI | 70.69 | 70.71 | 72.97 | 59.08 | 48.70 | 57.56 | 70.91 | 76.76 | 74.62 |
| | 180d | Orig | 71.48 | 81.23 | 70.56 | 65.28 | 70.57 | 68.81 | 68.00 | 75.61 | 61.55 |
| | | MVI | 72.41 | 82.09 | 71.48 | 63.01 | 70.16 | 69.74 | 69.16 | 79.10 | 62.58 |
| | 365d | Orig | 65.60 | 79.41 | 65.02 | 64.96 | 74.37 | 68.15 | 57.26 | 71.23 | 51.23 |
| | | MVI | 68.96 | 80.33 | 70.08 | 64.03 | 70.50 | 68.99 | 62.47 | 75.34 | 61.37 |

Table 4.6: AUC, Sensitivity and Specificity results for the prognostic models built without progression groups (baseline models) for the 90, 180 and 365 days Time Windows, relative to each Progression Group. The selected classifiers is Naive Bayes (NB).

| | | Sensitivity | Specificity | AUC |
|------|---------|-------------|-------------|-------|
| 90d | Overall | 69.87 | 74.99 | 79.84 |
| | Slow | 33.41 | 94.52 | 81.88 |
| | Neutral | 68.54 | 71.03 | 76.48 |
| | Fast | 85.04 | 34.70 | 68.36 |
| 180d | Overall | 73.22 | 73.94 | 80.57 |
| | Slow | 40.86 | 95.91 | 85.31 |
| | Neutral | 69.15 | 68.76 | 75.22 |
| | Fast | 84.87 | 31.87 | 66.11 |
| 365d | Overall | 78.41 | 71.96 | 82.25 |
| | Slow | 51.45 | 92.98 | 84.30 |
| | Neutral | 78.72 | 55.81 | 74.44 |
| | Fast | 90.84 | 31.87 | 66.37 |

By comparing the AUC between the overall population and each specific group in Table 4.6 we can see that they are similar. However, when looking at Sensitivity and Specificity, we see that the results are quite different and the specialized models are indeed more effective for NIV prediction.

Regarding Slow progressors, Sensitivity and Specificity measures are highly imbalanced, with a very high specificity and a very low sensitivity, meaning that the baseline model is correctly classifying almost all the negative instances and poorly classifying the positive instances. The opposite can be seen with Fast progressors, where the baseline model correctly predicts almost all positive instances but incorrectly classifies the majority of negative instances. Moreover, results for Fast progressors show that the AUC difference between approaches is higher than that of the other groups. Once again this can be due to the reduced number of instances hampering the performance of the classifiers. Neutral progressors are those with closer results between approaches. This is probably due to the fact that Neutral progressors give a higher contribution in terms of patient and learning instances to the baseline model, as well as the fact that they are more representative of the average of the population. This means that the results for the baseline model, generalize around the Neutral progressors, thus they predict better the instances from the Neutral group or those from the other groups that are closer to them. For the instances

4. PROGRESSION GROUPS

that are more dissimilar, the model tries to predict the outcome but generally fails. Using the specialized models for the groups ensures that each model learns with a subset of patients that are similar to each other and generalize around a less heterogeneous set of data.

In this context, the results show that there are clear benefits in using the specialized models for the disease progression groups rather than the baseline ones, provided we are able to compute the progression group of the patient.

To close this Section, Table 4.7 presents the best results obtained among all tests for the specialized models for each disease progression group.

Table 4.7: Best results obtained for the specialized models for each progression group, using the patients current condition.

| | | Sensitivity | Specificity | AUC |
|---------|---------|--------------------|--------------------|------------|
| Slow | 90d | 72.27 | 80.59 | 84.19 |
| | 180d | 74.85 | 85.45 | 88.45 |
| | 365d | 78.81 | 87.67 | 92.30 |
| | Neutral | | | |
| Neutral | 90d | 63.23 | 78.94 | 78.05 |
| | 180d | 63.11 | 86.05 | 83.72 |
| | 365d | 78.70 | 84.49 | 89.54 |
| Fast | 90d | 68.09 | 69.96 | 75.04 |
| | 180d | 70.16 | 79.10 | 82.09 |
| | 365d | 70.50 | 75.34 | 80.33 |

4.2 Using a Set of Snapshots

For this section, we use the datasets from our first stratification approach using Progression Groups to build specialized models capable to answer the question previously addressed in Section 3.2: "Given a patient's N consecutive appointments and the information about its progression group, will the patients evolve or not evolve to need NIV, k days from the date of the last appointment". Figure 4.4 presents an illustration of the problem addressed.

4.2.1 Creating Learning Instances

In order to create the learning instances to train the classifiers, we use the methodology used in Subsection 3.2.1 for each of the progression groups. This results in a set of new datasets whose statistics and class distribution are presented in Table 4.8.

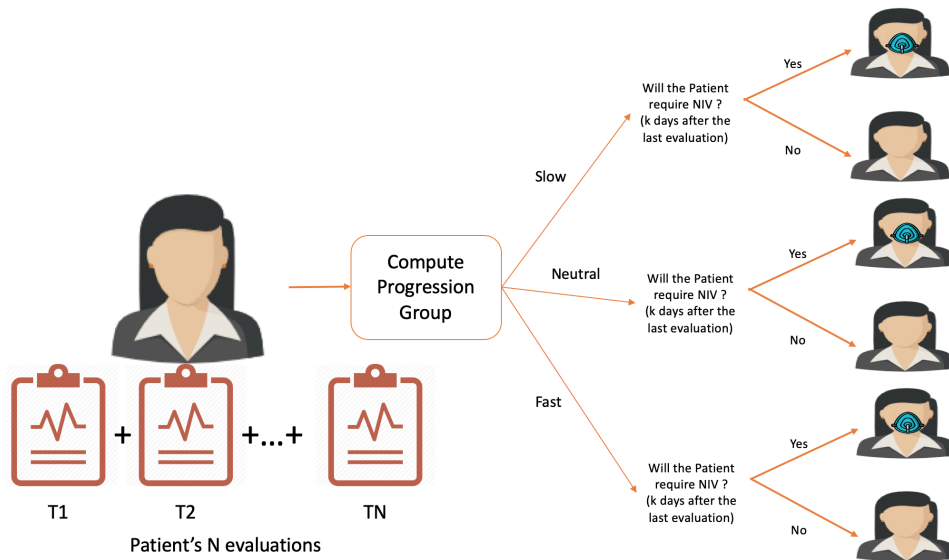


Figure 4.4: Problem Reformulation using Progression Groups: Knowing the Patient clinical history (follow-up), as well as their progression group can we predict the need for Non-Invasive Ventilation within k days from the last evaluation, using specialized models for each group?

In this analysis we can observe that the number of snapshots, the number of patients and thus the number of snapshots per patient all decrease. Similarly to Section 4.1.1, the Slow progressors have a large number of snapshots for the number of patients and the opposite happens to Fast progressors. In this case, the differences between each group are even more noticeable. For example, even though the Slow and Fast progression groups were initially created with approximately the same number of patients we can see how that the number of Slow progressors more than doubles the number of Fast progressors, as the patients with a faster progression unfortunately tend to die in a shorter period of time, thus typically not having long clinical histories. Moreover, the conclusions regarding the class distribution are the same as the ones in the Section 4.1.1, where Slow progressors have smaller proportions of positive instances, Neutral progressors are the ones with a distribution closer to the original datasets and Fast progressors show the greater proportions of positive instances.

4.2.2 Learning the Predictive Models

We build specialized models for each of the progression groups using the information of its patients' clinical history (two or three appointments) to predict the need for NIV, k days after the last appointment. The models are trained using a 5 x 10-fold CV for the 3 selected classifiers: NB, RF, and LR. After, we test the use of an FS and LOCF methodologies to check if they improve prediction.

4. PROGRESSION GROUPS

Table 4.8: Statistics and Class distribution for time windows of k=90,180, and 365, for each progression group using 2 and 3 time points TP.

| | | Nr of Snapshots | Nr of Patients | Snapshots p/ Patient | Evolution (E = 1) | No Evolution (E= 0) |
|----------------|----------|--------------------|-------------------|----------------------------|----------------------|------------------------|
| Slow | | | | | | |
| 2 TP | 90 days | 1023 | 186 | 5.5 | 63 (6.16%) | 960 (93.84%) |
| | 180 days | 977 | 183 | 5.34 | 125 (12.79%) | 852 (87.21%) |
| | 365 days | 886 | 173 | 5.12 | 216 (24.38%) | 670 (75.62%) |
| 3 TP | 90 days | 837 | 154 | 5.44 | 43 (5.14%) | 794 (94.86%) |
| | 180 days | 794 | 151 | 5.26 | 95 (11.96%) | 699 (88.04%) |
| | 365 days | 713 | 138 | 5.17 | 173 (24.26%) | 540 (75.74%) |
| Neutral | | | | | | |
| 2 TP | 90 days | 1017 | 308 | 3.3 | 219 (21.53%) | 798 (78.47%) |
| | 180 days | 964 | 298 | 3.23 | 356 (36.93%) | 608 (63.07%) |
| | 365 days | 878 | 275 | 3.19 | 544 (61.96%) | 334 (38.04%) |
| 3 TP | 90 days | 708 | 215 | 3.29 | 148 (20.90%) | 798 (78.47%) |
| | 180 days | 665 | 207 | 3.21 | 240 (36.09%) | 425 (63.91%) |
| | 365 days | 602 | 192 | 3.14 | 368 (61.13%) | 234 (38.87%) |
| Fast | | | | | | |
| 2 TP | 90 days | 211 | 91 | 2.32 | 66 (31.28%) | 145 (68.72%) |
| | 180 days | 188 | 88 | 2.14 | 93 (49.47%) | 95 (50.53%) |
| | 365 days | 161 | 77 | 2.09 | 118 (73.29%) | 43 (26.71%) |
| 3 TP | 90 days | 119 | 53 | 2.25 | 31 (26.05%) | 88 (73.95%) |
| | 180 days | 99 | 45 | 2.2 | 45 (45.45%) | 54 (54.55%) |
| | 365 days | 83 | 39 | 2.13 | 57 (68.67%) | 26 (31.33%) |

4.2.3 Results and Conclusions

Table 4.9 shows the results for this section without FS or MVI. We can see that in the majority of the cases using two time points (TP) is better than using 3 TP. However, comparing to previous models we can see that for the Slow progressors the models using 2 TP perform slightly better than the ones using only the current condition of the patient. The major difference between the two approaches concerns Fast progressors, where using multiple observations seems to have a greater negative impact in the performance of classifiers. However, this is probably due to the reduced number of training instances.

Finally, we present the results for the models using FSE and LOCF. Regarding FS, we chose to use only the set of features selected for each dataset since in the last section the models using the features selected for the original datasets did not show any improvement to the prediction. Tables 4.10 and 4.11 shows the results for these methodologies using 2 and 3 TP, respectively.

Table 4.9: AUC, Sensitivity and Specificity results for the prognostic models for the 90, 180 and 365 days Time Windows, as well as for each Progression Group using 2 and 3 Time Points. The selected classifiers are: Naive Bayes (NB), Random Forest (RF) and Linear Regression (LR).

| | | AUC | | | Sensitivity | | | Specificity | | | |
|----------------|------|-------|--------------|-------|-------------|-------|-------|-------------|-------|-------|--|
| | | NB | RF | LR | NB | RF | LR | NB | RF | LR | |
| Slow | 90d | | | | | | | | | | |
| | 2 TP | 81.18 | 78.91 | 69.46 | 64.13 | 67.62 | 59.37 | 79.54 | 75.08 | 72.65 | |
| | 3 TP | 77.52 | 77.96 | 68.64 | 66.05 | 68.37 | 60.00 | 76.85 | 71.56 | 69.82 | |
| | 180d | | | | | | | | | | |
| | 2 TP | 85.48 | 86.99 | 82.25 | 75.52 | 75.20 | 70.40 | 79.91 | 81.17 | 77.91 | |
| | 3 TP | 83.14 | 84.12 | 77.32 | 72.84 | 69.05 | 68.21 | 78.00 | 79.11 | 73.91 | |
| | 365d | | | | | | | | | | |
| | 2 TP | 84.22 | 90.47 | 79.67 | 72.22 | 76.57 | 69.07 | 79.61 | 84.96 | 77.43 | |
| | 3 TP | 82.95 | 89.25 | 80.12 | 70.64 | 77.34 | 70.06 | 80.33 | 84.22 | 78.52 | |
| Neutral | 90d | | | | | | | | | | |
| | 2 TP | 72.48 | 72.68 | 70.20 | 64.29 | 58.36 | 60.73 | 71.95 | 74.89 | 70.00 | |
| | 3 TP | 69.27 | 71.00 | 66.72 | 58.51 | 52.43 | 57.70 | 69.64 | 74.00 | 69.46 | |
| | 180d | | | | | | | | | | |
| | 2 TP | 72.96 | 78.91 | 72.04 | 58.93 | 59.21 | 60.67 | 75.36 | 80.82 | 71.64 | |
| | 3 TP | 70.73 | 77.05 | 70.74 | 55.92 | 56.00 | 59.17 | 73.84 | 81.04 | 72.80 | |
| | 365d | | | | | | | | | | |
| | 2 TP | 75.00 | 82.98 | 73.16 | 69.08 | 74.45 | 69.01 | 68.20 | 75.63 | 64.85 | |
| | 3 TP | 75.99 | 84.03 | 71.66 | 69.35 | 81.20 | 69.57 | 70.60 | 71.03 | 61.62 | |
| Fast | 90d | | | | | | | | | | |
| | 2 TP | 61.81 | 67.55 | 63.55 | 58.48 | 57.88 | 56.67 | 60.00 | 66.07 | 66.34 | |
| | 3 TP | 61.9 | 62.48 | 57.24 | 52.9 | 50.97 | 46.45 | 63.86 | 64.09 | 65.00 | |
| | 180d | | | | | | | | | | |
| | 2 TP | 64.4 | 71.95 | 65.81 | 69.89 | 67.96 | 60.43 | 46.95 | 66.74 | 62.95 | |
| | 3 TP | 59.12 | 74.22 | 59.47 | 54.67 | 62.67 | 51.56 | 55.93 | 75.93 | 60.37 | |
| | 365d | | | | | | | | | | |
| | 2 TP | 54.93 | 74.02 | 60.82 | 67.12 | 71.02 | 62.54 | 37.67 | 64.19 | 53.02 | |
| | 3 TP | 51.04 | 70.94 | 53.13 | 55.79 | 65.26 | 56.49 | 49.23 | 70.77 | 50.77 | |

4. PROGRESSION GROUPS

Table 4.10: AUC, Sensitivity and Specificity results for the prognostic models for the 90, 180 and 365 days Time Windows, as well as for each Progression Group, using 2 Time Points (TP). Orig is the original dataset, FS is the dataset with selected features for each progression group by the Feature Selection Ensemble (FSE), and MVI is the dataset imputed with LOCF imputation. The selected classifiers are: Naive Bayes (NB), Random Forest (RF) and Linear Regression (LR).

| | | AUC | | | Sensitivity | | | Specificity | | |
|----------------|------|-------|--------------|-------|-------------|-------|-------|-------------|-------|-------|
| | | NB | RF | LR | NB | RF | LR | NB | RF | LR |
| Slow | | | | | | | | | | |
| 90d | Orig | 81.18 | 78.91 | 69.46 | 64.13 | 67.62 | 59.37 | 79.54 | 75.08 | 72.65 |
| | FS | 80.52 | 77.95 | 73.21 | 66.03 | 65.71 | 60.00 | 78.31 | 72.73 | 74.38 |
| | MVI | 81.68 | 82.05 | 72.84 | 63.81 | 73.65 | 60.00 | 82.06 | 77.77 | 73.17 |
| 180d | Orig | 85.48 | 86.99 | 82.25 | 75.52 | 75.20 | 70.40 | 79.91 | 81.17 | 77.91 |
| | FS | 83.98 | 86.08 | 83.23 | 75.36 | 76.16 | 72.32 | 78.31 | 78.15 | 79.53 |
| | MVI | 85.37 | 88.70 | 83.56 | 71.20 | 76.80 | 72.16 | 82.09 | 82.09 | 78.97 |
| 365d | Orig | 84.22 | 90.47 | 79.67 | 72.22 | 76.57 | 69.07 | 79.61 | 84.96 | 77.43 |
| | FS | 82.86 | 88.64 | 84.02 | 72.04 | 82.69 | 73.61 | 79.10 | 79.67 | 77.40 |
| | MVI | 86.98 | 92.54 | 85.86 | 72.04 | 79.54 | 73.98 | 83.97 | 88.18 | 81.88 |
| Neutral | | | | | | | | | | |
| 90d | Orig | 72.48 | 72.68 | 70.20 | 64.29 | 58.36 | 60.73 | 71.95 | 74.89 | 70.00 |
| | FS | 72.69 | 72.88 | 73.22 | 66.30 | 63.20 | 65.30 | 68.82 | 69.77 | 70.10 |
| | MVI | 73.82 | 74.54 | 72.09 | 62.83 | 55.34 | 61.46 | 72.53 | 79.57 | 72.73 |
| 180d | Orig | 72.96 | 78.91 | 72.04 | 58.93 | 59.21 | 60.67 | 75.36 | 80.82 | 71.64 |
| | FS | 72.67 | 76.75 | 74.47 | 62.30 | 64.04 | 66.52 | 72.76 | 74.84 | 71.48 |
| | MVI | 74.47 | 81.27 | 73.76 | 58.93 | 61.35 | 61.74 | 76.22 | 83.68 | 73.98 |
| 365d | Orig | 75.00 | 82.98 | 73.16 | 69.08 | 74.45 | 69.01 | 68.20 | 75.63 | 64.85 |
| | FS | 75.72 | 80.09 | 74.87 | 61.36 | 64.08 | 68.42 | 78.32 | 79.88 | 68.56 |
| | MVI | 77.97 | 87.76 | 75.75 | 67.35 | 75.04 | 70.48 | 73.05 | 84.07 | 68.92 |
| Fast | | | | | | | | | | |
| 90d | Orig | 61.81 | 67.55 | 63.55 | 58.48 | 57.88 | 56.67 | 60.00 | 66.07 | 66.34 |
| | FS | 66.91 | 73.05 | 66.93 | 60.61 | 63.33 | 60.30 | 61.66 | 70.21 | 64.00 |
| | MVI | 62.08 | 70.29 | 63.79 | 61.82 | 57.58 | 55.45 | 54.21 | 69.66 | 67.31 |
| 180d | Orig | 64.40 | 71.95 | 65.81 | 69.89 | 67.96 | 60.43 | 46.95 | 66.74 | 62.95 |
| | FS | 65.33 | 69.93 | 65.47 | 53.55 | 59.57 | 62.58 | 68.63 | 66.53 | 61.47 |
| | MVI | 65.89 | 71.27 | 69.39 | 66.88 | 65.59 | 65.16 | 53.47 | 68.42 | 70.11 |
| 365d | Orig | 54.93 | 74.02 | 60.82 | 67.12 | 71.02 | 62.54 | 37.67 | 64.19 | 53.02 |
| | FS | 66.16 | 80.70 | 54.41 | 54.75 | 61.53 | 54.07 | 68.37 | 82.79 | 47.44 |
| | MVI | 60.38 | 76.22 | 62.18 | 61.02 | 71.19 | 64.41 | 50.70 | 67.91 | 55.35 |

Table 4.11: AUC, Sensitivity and Specificity results for the prognostic models for the 90, 180 and 365 days Time Windows, as well as for each Progression Group, using 3 Time Points (TP). Orig is the original dataset, FS is the dataset with selected features for each progression group by the Feature Selection Ensemble (FSE), and MVI is the dataset imputed with LOCF imputation. The selected classifiers are: Naive Bayes (NB), Random Forest (RF) and Linear Regression (LR).

| | | AUC | | | Sensitivity | | | Specificity | | | |
|----------------|------|------|--------------|--------------|-------------|-------|-------|-------------|-------|-------|-------|
| | | NB | RF | LR | NB | RF | LR | NB | RF | LR | |
| Slow | 90d | Orig | 77.52 | 77.96 | 68.64 | 66.05 | 68.37 | 60.00 | 76.85 | 71.56 | 69.82 |
| | | FS | 79.61 | 78.37 | 72.22 | 65.12 | 70.23 | 65.12 | 76.93 | 71.81 | 72.82 |
| | | MVI | 78.43 | 78.59 | 70.52 | 65.58 | 66.98 | 60.47 | 77.71 | 76.10 | 70.91 |
| | 180d | Orig | 83.14 | 84.12 | 77.32 | 72.84 | 69.05 | 68.21 | 78.00 | 79.11 | 73.91 |
| | | FS | 83.69 | 83.63 | 79.20 | 73.68 | 69.89 | 67.79 | 77.40 | 78.28 | 75.77 |
| | | MVI | 82.97 | 87.37 | 77.86 | 69.89 | 74.53 | 65.26 | 81.23 | 82.58 | 75.74 |
| | 365d | Orig | 82.95 | 89.25 | 80.12 | 70.64 | 77.34 | 70.06 | 80.33 | 84.22 | 78.52 |
| | | FS | 82.95 | 87.86 | 83.63 | 76.07 | 80.35 | 72.95 | 76.44 | 78.67 | 78.41 |
| | | MVI | 86.87 | 92.23 | 86.30 | 71.91 | 80.23 | 73.53 | 85.15 | 86.74 | 82.85 |
| Neutral | 90d | Orig | 69.27 | 71.00 | 66.72 | 58.51 | 52.43 | 57.70 | 69.64 | 74.00 | 69.46 |
| | | FS | 70.17 | 70.77 | 72.04 | 62.97 | 57.16 | 64.73 | 67.75 | 71.11 | 70.96 |
| | | MVI | 71.62 | 73.53 | 71.92 | 61.76 | 56.35 | 61.22 | 70.07 | 78.32 | 72.96 |
| | 180d | Orig | 70.73 | 77.05 | 70.74 | 55.92 | 56.00 | 59.17 | 73.84 | 81.04 | 72.80 |
| | | FS | 71.17 | 74.68 | 72.19 | 58.25 | 59 | 62.33 | 71.44 | 75.20 | 71.58 |
| | | MVI | 72.95 | 81.66 | 74.17 | 59.67 | 61.08 | 63.67 | 74.92 | 84.00 | 75.58 |
| | 365d | Orig | 75.99 | 84.03 | 71.66 | 69.35 | 81.20 | 69.57 | 70.60 | 71.03 | 61.62 |
| | | FS | 75.81 | 80.38 | 73.57 | 61.14 | 72.39 | 68.26 | 76.07 | 73.42 | 66.41 |
| | | MVI | 79.76 | 88.68 | 74.37 | 68.26 | 79.13 | 70.11 | 74.79 | 81.79 | 64.96 |
| Fast | 90d | Orig | 61.90 | 62.48 | 57.24 | 52.90 | 50.97 | 46.45 | 63.86 | 64.09 | 65.00 |
| | | FS | 59.06 | 66.23 | 53.53 | 63.87 | 52.26 | 53.55 | 49.09 | 68.86 | 54.32 |
| | | MVI | 59.32 | 60.04 | 63.32 | 48.39 | 45.81 | 49.03 | 68.18 | 68.41 | 70.23 |
| | 180d | Orig | 59.12 | 74.22 | 59.47 | 54.67 | 62.67 | 51.56 | 55.93 | 75.93 | 60.37 |
| | | FS | 61.73 | 73.89 | 51.44 | 49.33 | 64.44 | 51.11 | 75.19 | 77.04 | 52.59 |
| | | MVI | 68.07 | 74.68 | 56.85 | 57.78 | 60.89 | 54.67 | 64.07 | 74.81 | 60.74 |
| | 365d | Orig | 51.04 | 70.94 | 53.13 | 55.79 | 65.26 | 56.49 | 49.23 | 70.77 | 50.77 |
| | | FS | 63.43 | 78.09 | 38.25 | 54.73 | 59.29 | 48.07 | 60.76 | 80.77 | 30.77 |
| | | MVI | 63.18 | 77.8 | 75.21 | 46.67 | 68.42 | 66.32 | 69.23 | 71.54 | 70.77 |

4. PROGRESSION GROUPS

Once again we can see that using FS does not seem to improve the models' performance. Moreover, the Wilcoxon Signed Ranked Test for paired samples resulted in a p-value of 0.007, meaning that in this case, we should use the models that use all features available in lieu of the set chosen by the FSE. However, for Fast Progressors the models using FS present better results. This is probably due to the fact that they use less features, which is important since Fast Progressors have a low number of instances from which the models can learn.

As for LOCF, the results showed to improve the models' performance. To support this conclusion, the statistical test resulted in a p-value of approximately 0, thus reinforcing the use of this methodology for our models.

Table 4.12 presents the best results obtained for the specialized models for the progression groups using both the patients current condition (1TP) and the patients clinical history (2 and 3 TP) among all analysis performed in this Chapter.

Finally, the results using multiple evaluations (clinical history) do not show improvements in comparison to the ones using the patients current condition. However, we note that by using multiple patient observations, the number of learning instances that we can create is very limited. As the performance of the classifiers tends to improve with the increase of training data, it is then expected that the models using only the current condition of the patient would perform better as they are trained with larger quantities of information. We also expect that by using classifiers that can take advantage of the temporal nature of data by looking beyond independent variable, follow-up turns out to be important.

Table 4.12: Best results obtained for the specialized models for each disease progression group using 1 TP (current condition) as well as using 2 and 3 TP (clinical history).

| | | Sensitivity | Specificity | AUC |
|---------|------|-------------|-------------|-------|
| Slow | 90d | | | |
| | 1TP | 72.27 | 80.59 | 84.19 |
| | 2TP | 73.65 | 77.77 | 82.05 |
| | 3TP | 65.12 | 72.82 | 79.61 |
| | 180d | | | |
| | 1TP | 74.85 | 85.45 | 88.45 |
| | 2TP | 76.8 | 82.09 | 88.7 |
| | 3TP | 74.53 | 82.58 | 87.37 |
| | 365d | | | |
| | 1TP | 78.81 | 87.67 | 92.3 |
| | 2TP | 79.54 | 88.18 | 92.54 |
| | 3TP | 80.23 | 86.74 | 92.23 |
| Neutral | 90d | | | |
| | 1TP | 63.23 | 78.94 | 78.05 |
| | 2TP | 55.34 | 79.57 | 74.54 |
| | 3TP | 56.35 | 78.32 | 73.53 |
| | 180d | | | |
| | 1TP | 63.11 | 86.05 | 83.72 |
| | 2TP | 61.35 | 83.68 | 81.27 |
| | 3TP | 61.08 | 84 | 81.66 |
| | 365d | | | |
| | 1TP | 78.7 | 84.49 | 89.54 |
| | 2TP | 75.04 | 84.07 | 87.76 |
| | 3TP | 79.13 | 81.79 | 88.68 |
| Fast | 90d | | | |
| | 1TP | 68.09 | 69.96 | 75.04 |
| | 2TP | 63.33 | 70.21 | 73.05 |
| | 3TP | 52.26 | 68.86 | 66.23 |
| | 180d | | | |
| | 1TP | 70.16 | 79.1 | 82.09 |
| | 2TP | 67.96 | 66.74 | 71.95 |
| | 3TP | 60.89 | 74.81 | 74.68 |
| | 365d | | | |
| | 1TP | 70.5 | 75.34 | 80.33 |
| | 2TP | 61.53 | 82.79 | 80.7 |
| | 3TP | 59.29 | 80.77 | 78.09 |

Chapter 5

Patient Profiles

In this section we present our second patient stratification approach, using Clinical Profiles: General, Prognostic, Respiratory and Functional. These profiles are then used in the creation of patients profiles (groups of patient evaluations that are closer to each other). We create four different sets of patient profiles, each using a different set of features, according to the clinical profiles. The aim is to create specialized models for each profile, that can be used to create better models when compared to the baseline models, which use all snapshots from all patients. Moreover, by having several sets of profiles based of different subsets of features rather then just one using a set of features specific to our problem, we can then select the most fitting for other outcomes.

5.1 Creating Patient Profiles

Each set of patient profiles was obtained by clustering the patient snapshots using for each a set of features from each snapshot. We created 4 sets of profiles: General, Prognostic Respiratory, and Functional. The choice of features for each profile was performed with the help of the clinicians involved in this project. The features used to create each set of clinical profiles were:

- General Profile – All features available in the dataset;
- Prognostic Profile – Gender, BMI, Family History MND, Age at Onset, Disease Duration, El Escorial reviewed criteria, UMN vs LMN, Onset form, c9orf72, ALS-FRS, ALS-FRS-R and FVC;
- Respiratory Profile – FVC, PhrenMeanAmpl, ALS-FRSr, R;
- Functional Profile – ALS-FRS, ALS-FRS-R, ALS-FRSb, ALS-FRSsUL, ALS-FRSsLL, ALS-FRSr, R.

5. PATIENT PROFILES

For each clinical profile, we create a set of patient profiles. First, we select the features from each clinical profile from the datasets for each time window. Then, we perform clustering of all the available snapshots according to the features selected. The resulting clusters are then called patient profiles. The clustering method used was k -Means, which has a requirement that the number of clusters to create is known apriori. However, as we do not know the optimal number of groups that should be created for each profile, we performed clustering for a range of 2 to 10 clusters and then used the silhouette score to determine the optimum number to use. The clustering is also performed for each of the selected time windows since the dataset for each window is different, thus the clustering can have different results.

5.2 Single Snapshot Prediction

In this section, revisit the problem presented in Section 3.2, but this time we build specialized models for each of the patient profiles in order to predict whether or not given a patient's current condition and given a specific patient profile (computed using a clinical profile), the patient will need NIV in a time windows of k days. This problem is presented in Figure 5.1.

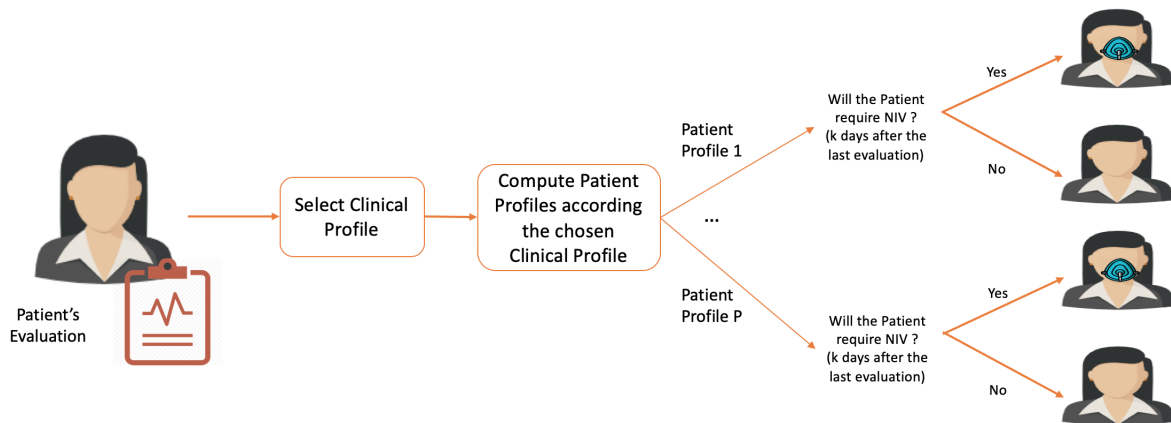


Figure 5.1: Problem reformulation using Patient Profiles: Knowing the Patient current state, as well as the attributed patient profile (for a given clinical profile) can we predict the need for NIV within a given time window?

Regarding the pipeline, we adapted the pipeline used in Chapter 3, in order to add the information on patient profiles. Figure 5.2 shows the updated pipeline.

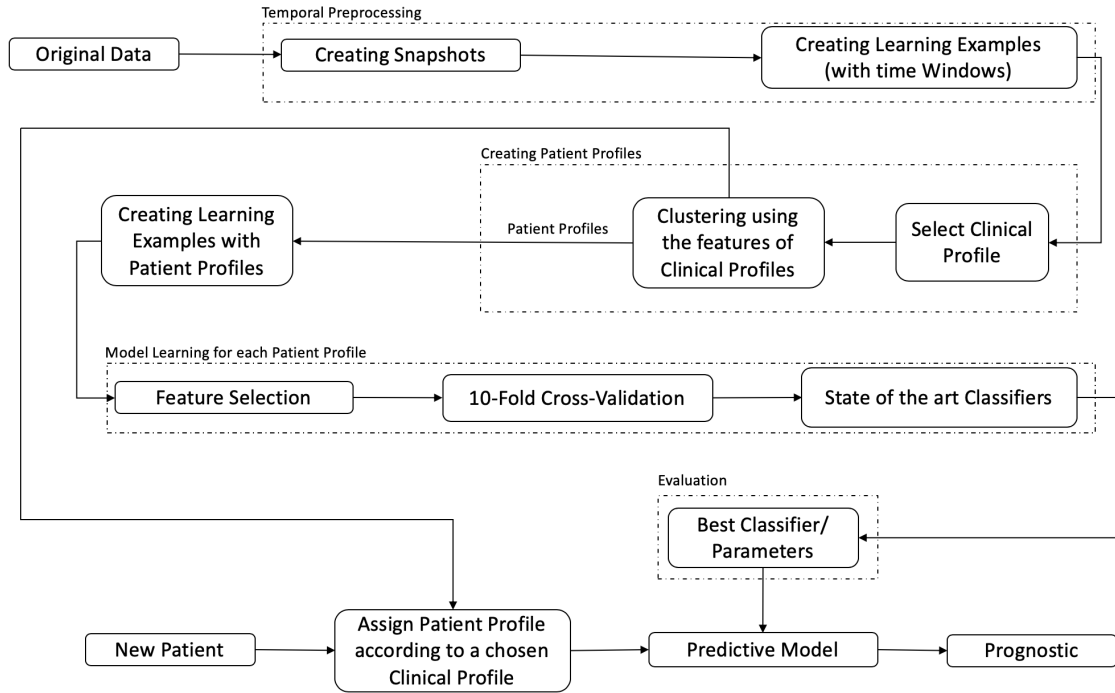


Figure 5.2: Workflow of the proposed methodology for ALS prognostic prediction using patient snapshots and patient profiles. Original data is preprocessed in order to create patient snapshots, that are then used to create the learning instances. These instances are then clustered to create different profiles. Then, the models are built using a stratified 5 x 10-fold cross-validation scheme. The models are then evaluated and the best parameters are chosen. After the final model is complete when a new patient arrives at the consult, their progression group is computed and their data is given to the specific predictive model that should output a prediction for the problem being addressed.

5.2.1 Creating Learning Instances

For all sets of patient profiles (one for each clinical profile) created in this section the highest Silhouette score was obtained when using $k=2$, thus, each clustering was performed to create two groups/ patient profiles. The groups were labeled as p1 and p2 for each dataset. However, we note that for two different datasets, the p1 or p2 groups of each are not the same. The instances belonging to each patient profile are then separated into different datasets that are used to train the specialized models. Table 5.1 presents the number of instances in each dataset and their class distribution. By analyzing the table we see that the groups obtained for the General and Prognostic profiles are very similar, not only in number of snapshots but also in class distribution. This can be explained by the fact that the prognostic profiles are created from a subset of features available in the general profile, that could be the most differentiating features, thus the clustering algorithm ends up creating almost the same groups. The groups created for each patient profile,

5. PATIENT PROFILES

seem to always consist in a larger cluster, containing the majority of instances, and a smaller one with the leftover instances. Moreover, in these cases, the smaller group seems to always have a more imbalanced distribution of NIV, when compared with the larger group. The Respiratory and Functional profiles show different results, showing less discrepancy in the number of instances in each group, and in these cases the smaller groups present more balanced class distribution than the larger.

5.2.2 Learning the predictive Models

After creating the datasets for each clinical profile and respective patient profiles, we use them as input to learn specialized models to predict the need for NIV. The models are trained using a 5 x 10-fold CV and AUC is the metric used to assess models' performance. The classifiers used are NB, RF, and LR. A grid search is also performed to determine the best parameters for each classifier.

Similarly to the previous analysis, we test the effects of FS and MVI in the models' performance. The methodologies used for this consist in FSE for the FS test and LOCF for MVI. Table 5.2 shows the set of features selected by the FSE for each dataset.

5.2 Single Snapshot Prediction

Table 5.1: Statistics and class distribution for each profile, across all time windows.

| | Nr of Snapshots | Evolution ($E = 1$) | No Evolution ($E = 0$) |
|--------------------|-----------------|-----------------------|--------------------------|
| 90 days | | | |
| General | | | |
| p1 | 2897 | 550 (18.99%) | 2347 (81.01%) |
| p2 | 281 | 9 (3.20%) | 272 (96.80%) |
| Prognostic | | | |
| p1 | 2895 | 549 (18.96%) | 2346 (81.04%) |
| p2 | 283 | 10 (3.53%) | 273 (96.47%) |
| Respiratory | | | |
| p1 | 453 | 204 (45.03%) | 249 (54.97%) |
| p2 | 2725 | 355 (13.03%) | 2370 (86.97%) |
| Functional | | | |
| p1 | 2348 | 334 (14.22%) | 2014 (85.78%) |
| p2 | 830 | 225 (27.11%) | 605 (72.89%) |
| 180 days | | | |
| General | | | |
| p1 | 2750 | 888 (32.29%) | 1862 (67.71%) |
| p2 | 268 | 18 (6.72%) | 250 (93.28%) |
| Prognostic | | | |
| p1 | 2748 | 887 (32.28%) | 1861 (67.72%) |
| p2 | 270 | 19 (7.04%) | 251 (92.96%) |
| Respiratory | | | |
| p1 | 433 | 278 (64.20%) | 155 (35.80%) |
| p2 | 2585 | 628 (24.29%) | 1957 (75.71%) |
| Functional | | | |
| p1 | 2224 | 548 (24.64%) | 1676 (75.36%) |
| p2 | 794 | 358 (45.09%) | 436 (54.91%) |
| 365 days | | | |
| General | | | |
| p1 | 2516 | 1314 (52.22%) | 1202 (47.78%) |
| p2 | 246 | 28 (11.38%) | 218 (88.61%) |
| Prognostic | | | |
| p1 | 2515 | 1314 (52.25%) | 1201 (47.75%) |
| p2 | 247 | 28 (11.34%) | 219 (88.66%) |
| Respiratory | | | |
| p1 | 410 | 337 (82.20%) | 73 (17.80%) |
| p2 | 2352 | 1005 (42.73%) | 1347 (57.27%) |
| Functional | | | |
| p1 | 2058 | 851 (41.35%) | 1027 (50.65%) |
| p2 | 704 | 491 (69.74%) | 213 (30.26%) |

5. PATIENT PROFILES

Table 5.2: Selected features by the Feature Selection Ensemble (FSE) for each clinical profile and respective set of Patient Profiles.

| Features | General Profiles | | | | | | Prognostic Profiles | | | | | | Respiratory Profiles | | | | | | Functional Profiles | | | | | |
|-------------------------------|------------------|----|-------|----|-------|----|---------------------|----|-------|----|-------|----|----------------------|----|-------|----|------|----|---------------------|----|-------|----|---|--|
| | 90 d | | 180 d | | 365 d | | 90 d | | 180 d | | 365 d | | 90 d | | 180 d | | 90 d | | 180 d | | 365 d | | | |
| | P1 | P2 | P1 | P2 | P1 | P2 | P1 | P2 | P1 | P2 | P1 | P2 | P1 | P2 | P1 | P2 | P1 | P2 | P1 | P2 | P1 | P2 | | |
| Gender | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | |
| Age at Onset | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | |
| BMI | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | |
| Family History MND | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | |
| Disease Duration | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | |
| El Escorial Reviewed Criteria | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | |
| UMN vs LMN | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | |
| Onset form | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | |
| c9orf72 | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | |
| ALS-FRS | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | |
| ALS-FRS-R | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | |
| ALS-FRSb | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | |
| ALS-FRSsUL | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | |
| ALS-FRSsLL | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | |
| ALS-FRSr | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | |
| R | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | |
| VC | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | |
| FVC | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | |
| MIP | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | |
| MEP | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | |
| P0.1 | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | |
| SNIP | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | |
| PhrenMeanLat | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | |
| PhrenMeanAmpl | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | |
| Cervical Flexion | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | |
| Cervical Extension | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | X | |

5.2.3 Results and Conclusions

Table 5.3 shows the results without FS or MVI for each dataset, in the three chosen time windows (90, 180, and 365 days).

Table 5.3: AUC, Sensitivity and Specificity results for the prognostic models for the 90, 180 and 365 days Time Windows, as well as for each Clinical Profile and respective set of Patient Profiles. The selected classifiers are: Naive Bayes (NB), Random Forest (RF) and Linear Regression (LR).

| | | AUC | | | Sensitivity | | | Specificity | | |
|--------------------|----------|--------------|--------------|-------|-------------|-------|-------|-------------|-------|-------|
| | | NB | RF | LR | NB | RF | LR | NB | RF | LR |
| General | | | | | | | | | | |
| p1 | 90 days | 78.86 | 79.75 | 78.25 | 68.07 | 64.40 | 67.35 | 75.59 | 78.46 | 74.89 |
| | 180 days | 79.15 | 84.32 | 79.02 | 69.80 | 68.27 | 70.05 | 74.20 | 81.57 | 74.59 |
| | 365 days | 80.11 | 87.57 | 77.95 | 76.85 | 79.59 | 72.60 | 69.53 | 79.32 | 70.95 |
| p2 | 90 days | 60.23 | 72.74 | 72.44 | 66.67 | 68.89 | 73.33 | 77.94 | 78.60 | 78.97 |
| | 180 days | 86.25 | 90.98 | 83.68 | 85.56 | 77.78 | 80.00 | 84.80 | 86.56 | 77.84 |
| | 365 days | 87.62 | 94.45 | 88.70 | 85.71 | 90.71 | 83.57 | 86.06 | 84.95 | 82.75 |
| Prognostic | | | | | | | | | | |
| p1 | 90 days | 78.63 | 79.55 | 77.92 | 68.63 | 63.90 | 68.01 | 74.69 | 78.58 | 75.12 |
| | 180 days | 79.11 | 84.26 | 78.97 | 69.33 | 67.03 | 69.29 | 74.78 | 82.47 | 74.54 |
| | 365 days | 80.09 | 87.93 | 78.23 | 76.18 | 79.13 | 72.89 | 70.11 | 80.37 | 71.64 |
| p2 | 90 days | 71.44 | 84.81 | 83.52 | 78.00 | 80.00 | 74.00 | 72.01 | 76.85 | 77.73 |
| | 180 days | 87.82 | 90.81 | 87.20 | 87.37 | 81.05 | 85.26 | 82.95 | 84.46 | 80.88 |
| | 365 days | 90.52 | 94.62 | 89.51 | 90.00 | 85.00 | 87.14 | 84.20 | 88.68 | 82.19 |
| Respiratory | | | | | | | | | | |
| p1 | 90 days | 70.54 | 73.53 | 70.9 | 71.96 | 66.08 | 64.22 | 59.04 | 67.95 | 69.64 |
| | 180 days | 68.36 | 73.11 | 68.14 | 71.65 | 70.65 | 69.21 | 56.00 | 61.29 | 58.58 |
| | 365 days | 71.29 | 79.65 | 69.98 | 72.52 | 76.44 | 70.56 | 61.37 | 70.41 | 59.45 |
| p2 | 90 days | 77.19 | 77.85 | 76.71 | 69.58 | 65.46 | 67.04 | 72.23 | 73.97 | 72.62 |
| | 180 days | 78.72 | 82.68 | 77.86 | 70.19 | 68.73 | 69.08 | 73.38 | 78.11 | 72.42 |
| | 365 days | 80.78 | 88.83 | 80.64 | 77.99 | 79.00 | 71.76 | 69.68 | 81.71 | 74.22 |
| Functional | | | | | | | | | | |
| p1 | 90 days | 80.82 | 81.41 | 79.83 | 71.98 | 69.58 | 71.26 | 75.68 | 77.33 | 74.58 |
| | 180 days | 81.27 | 85.28 | 81.33 | 72.55 | 71.50 | 73.61 | 76.31 | 80.85 | 74.71 |
| | 365 days | 81.82 | 88.82 | 80.34 | 77.74 | 76.64 | 73.21 | 69.81 | 83.60 | 74.40 |
| p2 | 90 days | 72.36 | 71.02 | 71.79 | 60.53 | 49.96 | 61.42 | 72.07 | 77.62 | 70.41 |
| | 180 days | 74.09 | 78.38 | 70.44 | 62.23 | 62.91 | 59.27 | 74.68 | 77.52 | 70.69 |
| | 365 days | 74.13 | 82.12 | 71.53 | 74.62 | 79.47 | 72.26 | 60.56 | 69.39 | 57.28 |

As in previous tests, the models for the longer time windows perform better. The models for General and Prognostic profiles behave overall better than the Respiratory and Functional profiles. Nonetheless, the results for the patient profiles approach are very promising. In fact, for the General and Prognostic profiles, in the 365 days window, the models reach AUC's higher

5. PATIENT PROFILES

than 94% which are so far the highest AUC scores obtained for the problem at hand.

Table 5.4 shows the effects of feature selection in the models.

Table 5.4: AUC, Sensitivity and Specificity results for the prognostic models using Feature Selection for the 90, 180 and 365 days Time Windows, as well as for each Clinical Profile and respective set Patient Profiles. The selected classifiers are: Naive Bayes (NB), Random Forest (RF) and Linear Regression (LR).

| | | AUC | | | Sensitivity | | | Specificity | | |
|--------------------|----------|--------------|--------------|--------------|-------------|-------|-------|-------------|-------|-------|
| | | NB | RF | LR | NB | RF | LR | NB | RF | LR |
| General | | | | | | | | | | |
| p1 | 90 days | 78.12 | 75.32 | 77.97 | 73.27 | 68.76 | 71.96 | 68.10 | 68.76 | 69.72 |
| | 180 days | 78.63 | 81.43 | 78.60 | 72.97 | 73.31 | 72.32 | 70.74 | 73.63 | 71.17 |
| | 365 days | 77.95 | 87.76 | 78.15 | 71.55 | 79.06 | 71.45 | 70.38 | 81.18 | 72.13 |
| p2 | 90 days | 67.90 | 71.24 | 69.26 | 66.67 | 73.33 | 75.56 | 80.88 | 76.54 | 76.40 |
| | 180 days | 90.32 | 87.38 | 90.08 | 65.56 | 76.67 | 82.22 | 88.16 | 81.60 | 78.64 |
| | 365 days | 92.38 | 92.26 | 86.17 | 90.71 | 87.86 | 81.43 | 82.75 | 82.02 | 80.28 |
| Prognostic | | | | | | | | | | |
| p1 | 90 days | 78.90 | 77.47 | 79.30 | 73.81 | 67.47 | 70.71 | 70.10 | 73.36 | 74.65 |
| | 180 days | 78.74 | 81.44 | 78.70 | 72.58 | 70.76 | 72.27 | 71.02 | 75.27 | 71.70 |
| | 365 days | 77.34 | 88.38 | 77.95 | 71.02 | 79.18 | 71.10 | 70.27 | 82.18 | 71.96 |
| p2 | 90 days | 75.55 | 81.97 | 74.13 | 80.00 | 78.00 | 64.00 | 79.41 | 75.24 | 80.95 |
| | 180 days | 71.38 | 74.87 | 69.14 | 72.63 | 69.47 | 57.89 | 52.83 | 68.61 | 66.29 |
| | 365 days | 92.09 | 92.90 | 88.45 | 85.71 | 90.00 | 86.43 | 83.20 | 82.92 | 83.20 |
| Respiratory | | | | | | | | | | |
| p1 | 90 days | 71.52 | 73.10 | 72.41 | 67.45 | 66.18 | 68.33 | 64.90 | 67.95 | 67.15 |
| | 180 days | 70.76 | 70.44 | 70.15 | 71.22 | 60.14 | 67.19 | 60.39 | 68.65 | 63.23 |
| | 365 days | 75.41 | 79.80 | 74.95 | 75.13 | 71.22 | 69.91 | 63.29 | 73.42 | 67.12 |
| p2 | 90 days | 72.21 | 72.24 | 73.59 | 76.56 | 68.00 | 68.96 | 58.53 | 64.95 | 65.92 |
| | 180 days | 77.31 | 80.57 | 77.82 | 73.57 | 76.31 | 73.44 | 68.58 | 70.18 | 69.77 |
| | 365 days | 78.22 | 89.18 | 78.51 | 77.35 | 81.61 | 73.45 | 65.27 | 82.21 | 70.38 |
| Functional | | | | | | | | | | |
| p1 | 90 days | 81.36 | 80.42 | 81.31 | 77.54 | 71.62 | 74.37 | 70.31 | 73.11 | 73.86 |
| | 180 days | 81.24 | 83.65 | 82.15 | 75.99 | 77.92 | 75.99 | 72.58 | 73.54 | 73.90 |
| | 365 days | 76.95 | 88.75 | 77.95 | 73.00 | 77.39 | 74.55 | 65.12 | 82.88 | 64.84 |
| p2 | 90 days | 71.50 | 67.63 | 71.38 | 76.71 | 53.07 | 68.09 | 48.93 | 71.37 | 62.02 |
| | 180 days | 68.17 | 76.03 | 67.80 | 73.69 | 62.96 | 67.32 | 49.86 | 74.17 | 58.26 |
| | 365 days | 70.97 | 82.08 | 68.64 | 75.40 | 69.86 | 73.40 | 52.02 | 77.93 | 48.26 |

Comparing to the models using no FS, these models show an overall loss in performance, although usually very small. This was already observed in previous tests. As the differences are not major, we performed a Wilcoxon Signed Rank Test for paired samples to check if they are statistically significant. The test resulted in a p-value of 0.0114, therefore the differences are statistically significant and we should not use the models using FS to predict the need for NIV.

5.2 Single Snapshot Prediction

However, we would like to note that, the set of features selected, can still be relevant for the clinicians to understand which tests are more important for each clinical profile.

Finally, Table 5.5 shows the effects of MVI, using LOCF as described before.

Table 5.5: AUC, Sensitivity and Specificity results for the prognostic models using Missing Value Imputation for the 90, 180 and 365 days Time Windows, as well as for each Clinical Profile and respective set Patient Profiles. The selected classifiers are: Naive Bayes (NB), Random Forest (RF) and Linear Regression (LR).

| | | AUC | | | Sensitivity | | | Specificity | | |
|--------------------|----------|--------------|--------------|-------|-------------|-------|-------|-------------|-------|-------|
| | | NB | RF | LR | NB | RF | LR | NB | RF | LR |
| General | | | | | | | | | | |
| p1 | 90 days | 79.60 | 80.86 | 78.15 | 70.22 | 65.24 | 66.55 | 73.75 | 79.41 | 74.41 |
| | 180 days | 80.50 | 85.93 | 79.61 | 72.03 | 70.25 | 69.50 | 74.14 | 83.37 | 74.64 |
| | 365 days | 82.21 | 90.75 | 80.96 | 76.06 | 81.48 | 74.93 | 72.65 | 84.38 | 74.08 |
| p2 | 90 days | 80.00 | 82.50 | 77.12 | 80.00 | 83.97 | 72.79 | 73.33 | 82.65 | 75.03 |
| | 180 days | 91.69 | 91.68 | 86.82 | 88.89 | 85.56 | 75.55 | 87.76 | 86.64 | 84.08 |
| | 365 days | 95.64 | 96.46 | 95.19 | 95.00 | 92.14 | 89.29 | 89.17 | 91.10 | 89.72 |
| Prognostic | | | | | | | | | | |
| p1 | 90 days | 79.41 | 80.56 | 78.24 | 72.09 | 65.36 | 66.74 | 72.71 | 79.97 | 74.96 |
| | 180 days | 80.48 | 86.35 | 79.67 | 70.51 | 69.79 | 69.09 | 75.52 | 84.52 | 74.95 |
| | 365 days | 82.17 | 90.62 | 81.15 | 75.71 | 81.64 | 74.93 | 73.09 | 84.18 | 74.45 |
| p2 | 90 days | 87.47 | 86.97 | 87.19 | 88.00 | 70.00 | 78.00 | 81.10 | 83.08 | 83.66 |
| | 180 days | 91.43 | 92.95 | 88.72 | 89.47 | 85.26 | 78.95 | 86.22 | 87.09 | 83.90 |
| | 365 days | 94.43 | 96.32 | 94.44 | 93.57 | 90.71 | 90.71 | 88.95 | 89.13 | 89.22 |
| Respiratory | | | | | | | | | | |
| p1 | 90 days | 70.08 | 74.11 | 70.42 | 71.76 | 66.67 | 63.73 | 57.59 | 68.59 | 69.56 |
| | 180 days | 68.83 | 73.59 | 67.69 | 70.65 | 69.78 | 67.99 | 56.26 | 65.68 | 56.52 |
| | 365 days | 73.96 | 79.01 | 68.45 | 74.48 | 76.85 | 70.21 | 59.45 | 66.85 | 56.16 |
| p2 | 90 days | 78.03 | 78.27 | 76.74 | 69.35 | 66.08 | 66.42 | 71.78 | 74.33 | 72.85 |
| | 180 days | 79.83 | 84.42 | 79.04 | 71.21 | 72.68 | 69.39 | 73.75 | 79.47 | 73.82 |
| | 365 days | 82.47 | 90.76 | 81.72 | 78.03 | 80.58 | 73.35 | 71.39 | 85.15 | 75.22 |
| Functional | | | | | | | | | | |
| p1 | 90 days | 82.42 | 83.05 | 81.40 | 75.27 | 71.50 | 71.14 | 74.79 | 77.75 | 75.65 |
| | 180 days | 82.48 | 86.55 | 82.60 | 73.54 | 74.74 | 74.23 | 77.12 | 81.96 | 76.58 |
| | 365 days | 84.13 | 91.53 | 83.87 | 76.87 | 80.24 | 75.86 | 75.00 | 86.79 | 77.53 |
| p2 | 90 days | 71.55 | 73.60 | 70.22 | 58.84 | 51.73 | 59.02 | 71.83 | 78.81 | 69.62 |
| | 180 days | 75.93 | 80.41 | 72.57 | 62.51 | 64.58 | 61.12 | 77.34 | 78.81 | 72.48 |
| | 365 days | 77.15 | 83.95 | 74.43 | 74.54 | 80.29 | 74.13 | 65.26 | 72.02 | 60.19 |

These models show higher improvements in performance when compared with the baseline results presented in Table 5.3. Some models reach AUC's higher than 90% and two have performances over 96% (both for the RF classifier using the 365 days time window). These are very promising results for this outcome. The Wilcoxon Signed Rank Test for paired samples

5. PATIENT PROFILES

comparing this results to the models without imputation yielded a p-value of approximately 0, meaning that these models are statistically better, thus, we should use MVI.

Overall, this stratification approach seems to be very useful to predict NIV, presenting better results for the specialized models for the General and Prognostic profiles, comparing to the baseline models. In future work it would be interesting to do a characterization of the patients in each patient profile. This would possibly allow us to find patterns that can be meaningful to the clinicians regarding the use of NIV or other outcomes.

Finally, Table 5.6 presents a summary of the best results achieved for this analysis.

When a patient arrives to a medical appointment, the clinician can decide which clinical profile, or set of clinical profiles, are more adequate to predict the desired clinical outcome. For each clinical profile selected, the patient's data from the evaluation will be compared to each one of the clusters created (patient profiles). The patient will then be assigned to the patient profile more similar to his/her data. Then, his/her data is used as input to the specialized model for that profile, in order to predict the need for NIV, or any other outcome.

5.2 Single Snapshot Prediction

Table 5.6: Best Results obtained for each Clinical Profile and respective set of patients profiles.

| | | | Sensitivity | Specificity | AUC |
|--------------------|------|----|-------------|-------------|-------|
| General | 90d | p1 | 65.24 | 79.41 | 80.86 |
| | | p2 | 83.97 | 82.65 | 82.5 |
| | 180d | p1 | 70.25 | 83.37 | 85.93 |
| | | p2 | 88.89 | 87.76 | 91.69 |
| | 365d | p1 | 81.48 | 84.38 | 90.75 |
| | | p2 | 92.14 | 91.1 | 96.46 |
| Prognostic | 90d | p1 | 65.36 | 79.97 | 80.56 |
| | | p2 | 88 | 81.1 | 87.47 |
| | 180d | p1 | 69.79 | 84.52 | 86.35 |
| | | p2 | 85.26 | 87.09 | 92.95 |
| | 365d | p1 | 81.64 | 84.18 | 90.62 |
| | | p2 | 90.71 | 85.15 | 96.32 |
| Respiratory | 90d | p1 | 66.67 | 68.59 | 74.11 |
| | | p2 | 66.08 | 74.33 | 78.27 |
| | 180d | p1 | 69.78 | 65.68 | 73.59 |
| | | p2 | 72.68 | 79.47 | 84.42 |
| | 365d | p1 | 71.22 | 73.42 | 79.8 |
| | | p2 | 80.58 | 85.15 | 90.76 |
| Functional | 90d | p1 | 71.5 | 77.75 | 83.05 |
| | | p2 | 51.73 | 78.81 | 73.6 |
| | 180d | p1 | 74.74 | 81.96 | 86.55 |
| | | p2 | 64.58 | 78.81 | 80.41 |
| | 365d | p1 | 80.24 | 86.79 | 91.53 |
| | | p2 | 80.29 | 72.02 | 83.95 |

Chapter 6

Conclusions and Future Work

6.1 Conclusions

We use data from a cohort of 1220 ALS patients that we preprocess into patient snapshots (a vector of features that describes the patient's current condition) and then into learning instances. Our first approach was to build a set of time independent prognostic models to predict the need within NIV 90, 180, and 365 days from the last evaluation. To make that prediction the models use the information of a patient's condition at the current appointment. These first models reached AUCs of 80.74%, 85.27%, and 88.95%. We also built similar models using the patient's clinical history instead of its current condition. However, using more than one snapshot did not show improvement when compared to the models using the last evaluation.

To deal with the heterogeneity of data we proposed two approaches to patient stratification. The first consists in assigning a new patient to a progression group according to his/her progression rate and use specialized models for each group to predict the target outcome. Although the models for each group do not show overall improvements compared to the baseline models, we proved that, in fact, for the two more dissimilar groups (Slow and Fast Progressors) the models only classifies correctly one of the classes. In this scenario, we showed the need for patient stratification to create specialized models, as generic models tend generalize around a group of patients that are more representative of the overall population. The disease progression groups approach resulted in a paper that was submitted and accepted in The Sixth Workshop on Data Mining in Biomedical Informatics and Healthcare 2018. It will be presented on November 17th in Singapore, held in conjunction with the IEEE International Conference on Data Mining (ICDM'18).

The second approach consists in grouping similar patient evaluation in order to create patient profiles. Four sets of patient profiles were created (one for each clinical profile): General, Prog-

6. CONCLUSIONS AND FUTURE WORK

nostic, Respiratory, and Functional. For each clinical profile we then compute patient profiles by clustering the patient evaluation into similar groups. The models trained using the patient profiles from the General and Functional set of clinical profiles showed better results than the others and even better results than those obtained by the baseline models. These profiles reached AUC's of over 96% for the time window of 365 days when using LOCF imputation, being the best results achieved for the NIV prediction. Once again, we showed that patient stratification is a useful tool and that specialized prognostic models can better predict the need for NIV for ALS patients.

6.2 Future Work

As future work, we propose creating an Ensemble using either all or a subset of the best models for each approach (Baseline, Disease Progression Groups and Patient Profiles). The Ensemble allows patients data to be used by several models and the final prediction will be based on the majority voting of all models. Hopefully, by gathering the advantages of each model into one place, we will be able to enhance our predictions.

The approaches proposed in this thesis are not exclusive to the NIV prediction. Therefore, replacing the clinical outcome would be interesting. One alternative could be to create models to predict other functional outcomes, such as knowing when the patient will need to use a wheelchair, knowing when the patient will lose the ability to speak, and other functional outcomes.

Temporal Data Mining is a sub-area data mining aimed to find temporal patterns and models to explain the data being analyzed. By using the aforementioned methods to find temporal patterns to be used by the prognostic models would be an interesting proposal to try to improve our models.

Creating models using classifiers that are sensitive to time constraints is also important. Thus, future work should tackle the creation of Time-Dependent Models for prognostic prediction in ALS.

References

- AHA, D.W., KIBLER, D. & ALBERT, M.K. (1991). Instance-Based Learning Algorithms. *Machine Learning*, **6**, 37–66. xvii, 12
- ALSABTI, K., RANKA, S., SINGH, V. & AMERICA, H. (1997). An efficient k-means clustering algorithm. *Electrical Engineering and Computer Science. Paper*, **43**. 14
- ANDERSEN, P.M., ABRAHAMS, S., BORASIO, G.D., DE CARVALHO, M., CHIO, A., VAN DAMME, P., HARDIMAN, O., KOLLEWE, K., MORRISON, K.E., PETRI, S., PRADAT, P.F., SILANI, V., TOMIK, B., WASNER, M. & WEBER, M. (2012). EFNS guidelines on the Clinical Management of Amyotrophic Lateral Sclerosis (MALS) - revised report of an EFNS task force. *European Journal of Neurology*, **19**, 360–375. 5, 28
- ARLOT, S. & CELISSE, A. (2009). A survey of cross-validation procedures for model selection. **4**, 40–79. 15, 16
- BOLÓN-CANEDO, V., SÁNCHEZ-MAROÑO, N., ALONSO-BETANZOS, A., BENÍTEZ, J.M. & HERRERA, F. (2014). A review of microarray datasets and applied feature selection methods. *Information Sciences*, **282**, 111–135. 7
- BREIMAN, L. (2001). Random forests. *Machine Learning*, **45**, 5–32. 11
- BROWN, R.H. & AL-CHALABI, A. (2017). Amyotrophic Lateral Sclerosis. *New England Journal of Medicine*, **377**, 162–172. 1, 5, 6, 7
- CARREIRO, A. (2016). *An integrative approach for prognostic prediction in neurodegenerative diseases..* Ph.D. thesis, Instituto Superior Técnico, Universidade de Lisboa. xvii, 2, 19, 20, 21
- CARREIRO, A.V., AMARAL, P.M., PINTO, S., TOMÁS, P., DE CARVALHO, M. & MADEIRA, S.C. (2015). Prognostic models based on patient snapshots and time windows: Predicting disease progression to assisted ventilation in Amyotrophic Lateral Sclerosis. *Journal of Biomedical Informatics*, **58**, 133–144. 27, 28, 33

REFERENCES

- CASTRILLO-VIGUERA, C., GRASSO, D.L., SIMPSON, E., SHEFNER, J. & CUDKOWICZ, M.E. (2010). Clinical significance in the change of decline in ALSFRS-R. *Amyotrophic Lateral Sclerosis*, **11**, 178–180. 23
- CEDARBAUM, J.M., STAMBLER, N., MALTA, E., FULLER, C., HILT, D., THURMOND, B. & NAKANISHI, A. (1999). The ALSFRS-R: A revised ALS functional rating scale that incorporates assessments of respiratory function. *Journal of the Neurological Sciences*, **169**, 13–21. 22
- CHANDRASHEKAR, G. & SAHIN, F. (2014). A survey on feature selection methods. *Computers and Electrical Engineering*, **40**, 16–28. 8
- CHAWLA, N.V., BOWYER, K.W., HALL, L.O. & KEGELMEYER, W.P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, **16**, 321–357. 9
- CHEEMA, J.R. (2014). A Review of Missing Data Handling Methods in Education Research. *Review of Educational Research*, **84**, 487–508. 8
- CHIÒ, A., CALVO, A., MOGLIA, C., MAZZINI, L., MORA, G., MUTANI, R., BALMA, M., CAMMAROSANO, S., CANOSA, A., GALLO, S., ILARDI, A., DURELLI, L., FERRERO, B., DE MERCANTI, S., MAURO, A., LEONE, M., MONACO, F., NASUELLI, N., SOSSO, L., GIONCO, M., MARCHET, A., BUFFA, C., CAVALLO, R., ODDENINO, E., GEDA, C., DORIGUZZI BOZZO, C., MAGLIOLA, U., PAPURELLO, D., SANTIMARIA, P., MASSAZZA, U., VILLANI, A., CONTI, R., PISANO, F., PALERMO, M., VERGNANO, F., PENZA, M.T., DI VITO, N., AGUGLIA, M., PASTORE, I., MEINER, P., GHIGLIONE, P., SELIAK, D., CAVESTRO, C., ASTEGIANO, G., CORSO, G. & BOTTACCHI, E. (2011). Phenotypic heterogeneity of amyotrophic lateral sclerosis: A population based study. *Journal of Neurology, Neurosurgery and Psychiatry*, **82**, 740–746. 1
- COULET, A., COHEN, K. & ALTMAN, R. (2012). The state of the art in text mining and natural language processing for pharmacogenomics. *Journal of Biomedical Informatics*.
- CREEMERS, H., GRUPSTRA, H., NOLLET, F., VAN DEN BERG, L.H. & BEELEN, A. (2015). Prognostic factors for the course of functional status of patients with ALS: a systematic review. *Journal of Neurology*, **262**, 1407–1423. 1

- DONDERS, A.R.T., VAN DER HEIJDEN, G.J., STIJNEN, T. & MOONS, K.G. (2006). Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, **59**, 1087–1091. 8
- FAYYAD, U., PIATETSKY-SHAPIRO, G. & SMYTH, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, **17**, 37. 7
- FUREY, T.S., CRISTIANINI, N., DUFFY, N., BEDNARSKI, D.W., SCHUMMER, M. & HAUSLER, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **16**, 906–914. 12
- GANESALINGAM, J., STAHL, D., WIJESSEKERA, L., GALTREY, C., SHAW, C.E., LEIGH, P.N. & AL-CHALABI, A. (2009). Latent cluster analysis of ALS phenotypes identifies prognostically differing groups. *PLoS ONE*, **4**, 1, 22
- GEORGES, M., MORÉLOT-PANZINI, C., SIMIŁOWSKI, T. & GONZALEZ-BERMEJO, J. (2014). Noninvasive ventilation reduces energy expenditure in amyotrophic lateral sclerosis. 1–8. 2, 7
- GEORGOULOPOULOU, E., FINI, N., VINCETI, M., MONELLI, M., VACONDIO, P., BIANCONI, G., SOLA, P., NICHELLI, P. & MANDRIOLI, J. (2013). The impact of clinical factors, riluzole and therapeutic interventions on ALS survival: A population based study in Modena, Italy. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, **14**, 338–345. 1
- GUPTA, P.K., PRABHAKAR, S., SHARMA, S. & ANAND, A. (2012). A predictive model for amyotrophic lateral sclerosis (ALS) diagnosis. *Journal of the Neurological Sciences*, **312**, 68–72. 1
- GUYON, I. & ELISSEEFF, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research (JMLR)*, **3**, 1157–1182. 8
- HALL, M.A., FRANK, E., HOLMES, G., PFAHRINGER, B., REUTEMANN, P. & WITTEN, I.H. (2009). The WEKA data mining software: an update. *SIGKDD Explorations*, **11**, 10–18. 29
- HAN, J., KAMBER, M. & PEI, J. (2012). *Data mining: concepts and techniques*. Morgan Kaufmann. xvii, 7, 10, 11, 13
- HARDIMAN, O., VAN DEN BERG, L.H. & KIERNAN, M.C. (2011). Clinical diagnosis and management of amyotrophic lateral sclerosis. *Nature Reviews Neurology*, **7**, 639–649. 6
- HE, H. & GARCIA, E.A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, **21**, 1263–1284. 9

REFERENCES

- HOSMER, D.W. & LEMESHOW, S. (2000). *Applied Logistic Regression*. 1. 13
- HSU, C.W., CHANG, C.C. & LIN, C.J. (2008). A Practical Guide to Support Vector Classification. *BJU international*, **101**, 1396–400. 12
- HU, Y.J., KU, T.H., JAN, R.H., WANG, K., TSENG, Y.C. & YANG, S.F. (2012). Decision tree-based learning to predict patient controlled analgesia consumption and readjustment. *BMC Medical Informatics and Decision Making*, **12**, 131. 10
- JOSSE, J. & HUSSON, F. (2012). Handling missing values in exploratory multivariate data analysis methods. *Journal de la Société Française de . . .*, **153**, 79–99. 8
- KIMURA, F., FUJIMURA, C., ISHIDA, S., NAKAJIMA, H., FURUTAMA, D., UEHARA, H., SHINODA, K., SUGINO, M. & HANAFUSA, T. (2006). Progression rate of ALSFRS-R at time of diagnosis predicts survival time in ALS. *Neurology*, **66**, 265 LP – 267. 23
- KRAWCZYK, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, **5**, 221–232. 9
- KRISHNA, K. & MURTY, M.N. (1999). Genetic K-means algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, **29**, 433–439. 14
- KÜFFNER, R., ZACH, N., NOREL, R., HAWE, J., SCHOENFELD, D., WANG, L., LI, G., FANG, L., MACKEY, L., HARDIMAN, O., CUDKOWICZ, M., SHERMAN, A., ERTAYLAN, G., GROSSE-WENTRUP, M., HOTHORN, T., VAN LIGTENBERG, J., MACKE, J.H., MEYER, T., SCHÖLKOPF, B., TRAN, L., VAUGHAN, R., STOLOVITZKY, G. & LEITNER, M.L. (2015). Crowdsourced analysis of clinical trial data to predict amyotrophic lateral sclerosis progression. *Nature Biotechnology*, **33**, 51–57. 19
- LARRY JAMESON, J. & LONGO, D.L. (2015). Precision Medicine—Personalized, Problematic, and Promising. *Obstetrical & Gynecological Survey*, **70**, 612–614. 1
- LAXMAN, S. & SASTRY, P.S. (2006). A survey of temporal data mining. *Sadhana*, **31**, 173–198. 7
- LI, J., CHENG, K., WANG, S., MORSTATTER, F., TREVINO, R., TANG, J. & LIU, H. (2017). Feature selection: A data perspective. *ACM Computing Surveys*, **50**. 8
- LIU, T., MOORE, A.W., GRAY, A. & YANG, K. (2004). An investigation of practical approximate nearest neighbor algorithms. *Advances in Neural Information Processing Systemsnformation*, 8. 11

- MARTIN, S., AL KHLEIFAT, A. & AL-CHALABI, A. (2017). What causes amyotrophic lateral sclerosis? *F1000Research*, **6**, 371. 5, 22
- MINE, P. & SEQUENCING, A.L.S. (2018). Project MinE: study design and pilot analyses of a large-scale whole-genome sequencing study in amyotrophic lateral sclerosis. *European Journal of Human Genetics*. 6
- NEWGARD, C.D. & LEWIS, R.J. (2015). Missing Data. Howto Best Account for What Is Not Known. *Jama*, **314**, 940. 8, 31
- PASTULA, D.M., COFFMAN, C.J., ALLEN, K.D., ODDONE, E.Z., KASARSKIS, E.J., LINDQUIST, J.H., MORGENLANDER, J.C., NORMAN, B.B., ROZEAR, M.P., SAMS, L.A., SABET, A. & BEDLACK, R.S. (2009). Factors associated with survival in the National Registry of Veterans with ALS. *Amyotrophic Lateral Sclerosis*, **10**, 332–338. 1
- PEREIRA, T., FERREIRA, F. & SANDRA, C. (2018). Neuropsychological predictors of conversion from mild cognitive impairment to alzheimer’s disease: A feature selection ensemble combining stability and predictability. 29, 49
- POLKEY, M.I., LYALL, R.A., YANG, K., JOHNSON, E., NIGEL LEIGH, P. & MOXHAM, J. (2017). Respiratory muscle strength as a predictive biomarker for survival in amyotrophic lateral sclerosis. *American Journal of Respiratory and Critical Care Medicine*, **195**, 86–95. 19
- PROUDFOOT, M., JONES, A., TALBOT, K., AL-CHALABI, A. & TURNER, M.R. (2016). The ALSFRS as an outcome measure in therapeutic trials and its relationship to symptom onset. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, **17**, 414–425. 22
- RAHMAN, M.M. & DAVIS, D.N. (2013). Addressing the Class Imbalance Problem in Medical Datasets. *International Journal of Machine Learning and Computing*, **3**, 224–228. 9
- REY, D. & NEUHÄUSER, M. (2011). *Wilcoxon-Signed-Rank Test*, 1658–1659. Springer Berlin Heidelberg, Berlin, Heidelberg. 32
- RISH, I. (2001). An empirical study of the naive Bayes classifier. *Empirical methods in artificial intelligence workshop, IJCAI*, **22230**, 41–46. 12
- SATO, Y., NAKATANI, E., WATANABE, Y., FUKUSHIMA, M., NAKASHIMA, K., KANNAGI, M., KANATANI, Y. & MIZUSHIMA, H. (2015). Prediction of prognosis of ALS: Importance of active denervation findings of the cervical-upper limb area and trunk area. *Intractable and Rare Diseases Research*, **4**, 181–189. 19

REFERENCES

- SHARMA, J., SHARMA, S., PANDEY, R. & SCHOLAR, M.T. (2018). A complete review of concept of data mining. *5*, 3143–3146. 7
- SIMON, N.G., TURNER, M.R., VUCIC, S., AL-CHALABI, A., SHEFNER, J., LOMEN-HOERTH, C. & KIERNAN, M.C. (2014). Quantifying disease progression in amyotrophic lateral sclerosis. *Annals of Neurology*, **76**, 643–657. 23
- SOMORJAI, R.L., DOLENKO, B. & BAUMGARTNER, R. (2003). Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: Curses, caveats, cautions. *Bioinformatics*, **19**, 1484–1491. 7
- SONG, Y.Y. & LU, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, **27**, 130–5. 10
- TANG, J., ALELYANI, S. & LIU, H. (2014). Feature Selection for Classification: A Review. *Data Classification: Algorithms and Applications*, 37–64. 7
- TRAYNOR, B.J., ALEXANDER, M., CORR, B., FROST, E. & HARDIMAN, O. (2003). Effect of a multidisciplinary amyotrophic lateral sclerosis (ALS) clinic on ALS survival: A population based study, 1996-2000. *Journal of Neurology, Neurosurgery and Psychiatry*, **74**, 1258–1261. 1
- VAN ES, M.A., HARDIMAN, O., CHIO, A., AL-CHALABI, A., PASTERKAMP, R.J., VELDINK, J.H. & VAN DEN BERG, L.H. (2017). Amyotrophic lateral sclerosis. *The Lancet*, **390**, 2084–2098. 1, 5, 6
- WESTENENG, H.J., DEBRAY, T.P., VISSER, A.E., VAN EIJK, R.P., ROONEY, J.P., CALVO, A., MARTIN, S., MCDERMOTT, C.J., THOMPSON, A.G., PINTO, S., KOBELEVA, X., ROSENBOHM, A., STUBENDORFF, B., SOMMER, H., MIDDELKOOP, B.M., DEKKER, A.M., VAN VUGT, J.J., VAN RHEENEN, W., VAJDA, A., HEVERIN, M., KAZOKA, M., HOLLINGER, H., GROMICHO, M., KÖRNER, S., RINGER, T.M., RÖDIGER, A., GUNKEL, A., SHAW, C.E., BREDENOORD, A.L., VAN ES, M.A., CORCIA, P., COURATIER, P., WEBER, M., GROSSKREUTZ, J., LUDOLPH, A.C., PETRI, S., DE CARVALHO, M., VAN DAMME, P., TALBOT, K., TURNER, M.R., SHAW, P.J., AL-CHALABI, A., CHIÒ, A., HARDIMAN, O., MOONS, K.G., VELDINK, J.H. & VAN DEN BERG, L.H. (2018). Prognosis for patients with amyotrophic lateral sclerosis: development and validation of a personalised prediction model. *The Lancet Neurology*, **17**, 423–433. 1, 19, 21

REFERENCES

- ZAREI, S., CARR, K., REILEY, L., DIAZ, K., GUERRA, O., ALTAMIRANO, P., PAGANI, W., LODIN, D., OROZCO, G. & CHINEA, A. (2015). A comprehensive review of amyotrophic lateral sclerosis. *Surgical Neurology International*, **6**, 171. 5, 6
- ZOU, Z.Y., LIU, C.Y., CHE, C.H. & HUANG, H.P. (2016). Toward precision medicine in amyotrophic lateral sclerosis. *Annals of translational medicine*, **4**, 27. 1, 6
page

Appendices

Appendix A

Predicting Non-Invasive Ventilation in ALS Patients using Stratified Disease Progression Groups

Sofia Pires*

LASIGE, Faculdade de Ciências, Universidade de Lisboa
Lisbon, Portugal
sofiaferropires@gmail.com

Susana Pinto

Instituto de Medicina Molecular, Instituto de Fisiologia
Faculdade de Medicina, Universidade de Lisboa
Lisbon, Portugal
susana.c.pinto@medicina.ulisboa.pt

Sara C. Madeira*

LASIGE, Faculdade de Ciências, Universidade de Lisboa
Lisbon, Portugal
sacmadeira@ciencias.ulisboa.pt

Marta Gromicho

Instituto de Medicina Molecular, Instituto de Fisiologia
Faculdade de Medicina, Universidade de Lisboa
Lisbon, Portugal
martalgms@gmail.com

Mamede de Carvalho

Instituto de Medicina Molecular, Instituto de Fisiologia
Faculdade de Medicina, Universidade de Lisboa
Lisbon, Portugal
Department of Neurosciences and Mental Health
Hospital de Santa Maria CHLN
Lisbon, Portugal
mamedealves@medicina.ulisboa.pt

Abstract—Amyotrophic Lateral Sclerosis (ALS) is a neurodegenerative disease highly known for its rapid progression, leading to death usually within a few years. Respiratory failure is the most common cause of death. Therefore, efforts must be taken to prevent respiratory insufficiency. Preventive administration of non-invasive ventilation (NIV) has proven to improve survival in ALS patients. Using disease progression groups revealed to be of great importance to ALS studies, since the heterogeneous nature of disease presentation and progression presents challenges to the learn of predictive models that work for all patients. In this context, we propose an approach to stratify patients in three progression groups (Slow, Neutral and Fast) enabling the creation of specialized learning models that predict the need of NIV within a time window of 90, 180 or 365 days of their current medical appointment. The models are built using a collection of classifiers and 5x10-fold cross validation. We also test the use of a Feature Selection Ensemble to test which features are more relevant to predict this outcome. Our specialized predictive models showed promising results, proving the utility of patient stratification when predicting NIV in ALS patients.

Index Terms—Amyotrophic Lateral Sclerosis, Patient Stratification, Prognostic Prediction, Disease Progression Groups

I. INTRODUCTION

Amyotrophic Lateral Sclerosis (ALS) is a devastating disease characterized by the progressive loss of motor neurons in the brain and spinal cord [1]. Patients with ALS generally die from respiratory failure within 3 to 5 years [2]. However, due to the heterogeneity of the disease, some patients die in less than a year from disease onset, while some can live with it for

over 10 years [3]. This aspect of the disease hinders our understanding of it, making it difficult to provide early diagnosis and develop treatments based on disease progression.

Although approximately 10% of patients has a family history of ALS, genetic factors are not the only cause. Epigenetic factors, as well as environmental and internal factors, can have a role in the causality of the disease [4]. At the moment there is no cure to ALS. Thus, efforts to maintain the quality of life and improve the prognosis are imperative.

Respiratory Muscle weakness, leading to respiratory failure, is the most common cause of death in ALS patients. Thus, clinicians should take especial attention to early signs of respiratory insufficiency [5]. Non-invasive ventilation (NIV) is generally used in ALS to improve prognosis and quality of life [6].

In this context, Carreiro et al. [7] proposed a prognostic prediction approach to predict the need for NIV in ALS patients given a predefined time window. The use of demographic and clinical data together with state of the art data mining techniques provided a tool that can help clinicians to anticipate prescription of NIV and possibly improve patients' prognosis.

Due to ALS heterogeneous nature, special attention has been given in recent years to patient stratification. The idea is that designing specialized models using groups of patients who are similar either according to their progression [8] or to sets of prognostic biomarkers [9], may eventually help to understand the underlying mechanisms of the disease, providing a new perspective on how to plan clinical trials and better manage the

*Corresponding Authors

disease. The DREAM-Phil Bowen ALS Prediction Prize4Life Challenge [10] encouraged researchers to use clinical trial data to predict disease progression in 3 to 12 months, leading to several proposals that helped validating various prognostic features described in the literature [8].

In this work, we revisit the work of Carreiro et al. [7] using the follow-up data from our ALS clinic. We also propose an for patient stratification using the patients’ progression rate. We divide patients into three separate disease progression groups (Slow, Neutral and Fast). For each group we train specialized models to predict the need for NIV, within a given time window (k days) from the date of current appointment. Given that patients usually have medical appointments every 3 months, we use three time windows: 90, 180, and 365 days.

II. METHODS

First we provide an overview of data used in this work. Our data comprises demographic, clinical, and genetic information from 1220 ALS patients observed from 1992 until March 2018. It has 27 features, most of them known prognostic biomarkers for ALS. These features can be divided into two groups: static and temporal. Static features do not change over time, such as demographic and genetic features. Temporal features are clinical tests that are usually measured at each appointment (every 3 months). Table I presents the features used in this work.

Figure 1 illustrates our baseline approach (following Carreiro et al. [7]): "Given a patient’s current condition, can we predict the need for NIV in a predefined time window (k days)?" To tackle this problem we follow the workflow presented in Figure 2. First we need to transform data into snapshots and then into learning examples, according to the chosen time window. Then we optionally run a feature selection ensemble to select the best features for classification. We then build the models from various classifiers that try to predict the need for NIV according to the previously chosen time window. After the models are trained, when a new patient goes to an appointment, his/her data is fed to the predictive model that predicts whether or not the patient will require the use of NIV within k days.

In our proposed approach we will also aim at predicting the need for NIV in k days. However, we perform patient stratification using patients’ disease progression rate to create three disease progression groups: Slow, Neutral, and Fast Progressors. Then, we build specialized models using data from each group, aiming to improve prognostic prediction. Our revised version of the prognostic prediction problem is illustrated in Figure 3.

In Figure 4, We adapted the workflow used for the baseline approach, to account for the use of disease progression groups. The progression groups are created from the whole population of patients using the information of their first symptoms and their first visit. The patients snapshots and learning instances are created in the same way as before [7]. After created they are split into separate datasets for each disease progression group. Then classifiers are trained for each progression group

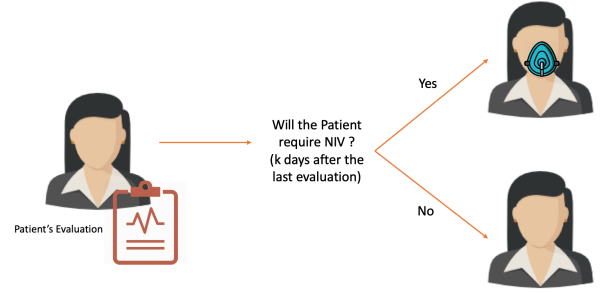


Fig. 1. Problem Formulation: Given the Patient current condition, can we predict the need for Non-Invasive Ventilation (NIV) in a predefined time window (k days)?

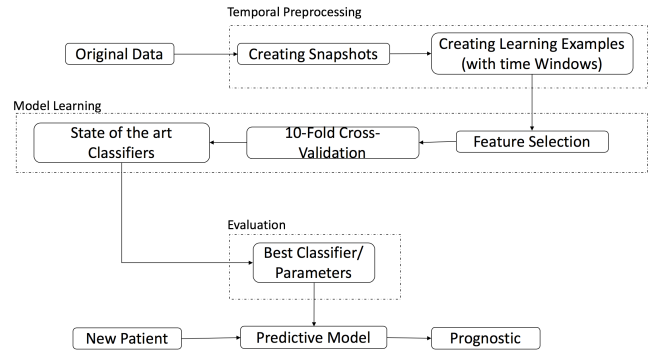


Fig. 2. Workflow of the methodology for ALS prognostic prediction using patient snapshots (following Carreiro et al.). Original data is preprocessed in order to create patient snapshots, which are then used to create the learning instances. The models are then built using a stratified 5 x 10-fold cross validation scheme. The models are then evaluated and the best parameters are chosen. After the final model is learnt, whenever a new patient is evaluated, his/her data is fed to the predictive model that outputs a prediction: need or not need for NIV.

to predict if a patient will need NIV in k days. Once the models are trained, whenever a new patient comes to an appointment, we compute his/her progression rate to identify the corresponding progression group, and his/her data is used by the specialized model to predict the desired target.

A. Data Preprocessing

In this section we further explain how we preprocess data in order to be used in the predictive models. First we need to transform the original data into patient snapshots that have all the information of a patient’s current state. Then we need to label those snapshots with an Evolution class. The class has information about that patient’s need for NIV within a time window (Yes or No). We also present our approach to create disease progression groups, using all population.

1) *Creating Snapshots and Learning Instances*: Data used in this work is a combination of static data (demographics, family history, onset evaluation and genetic information), and temporal data (set of clinical tests performed at each

TABLE I
AVAILABLE FEATURES IN THE ALS DATASET

| Static | Demographics | Gender | Body Mass Index (BMI) | | Age at onset | |
|---|---------------------------------|--|-----------------------|----------------------------------|------------------|--|
| | Medical and Family History | Family History of Motor Neurone Disease (MND) | | | | |
| | Onset Evaluation | UMN vs LMN | | Onset Form | Diagnostic Delay | |
| | Genetic | El Escorial Reviewed Criteria | | | | |
| Temporal | Expression of c9orf72 Mutations | | | | | |
| | Functional Scores | ALSFRS* | ALSFRS-R* | ALSFRSb* | R* | |
| | | ALSFRSsUL* | | ALSFRSsLL* | ALSFRSf* | |
| | Respiratory Tests | Vital Capacity (VC) | Forced VC (FVC) | Airway Occlusion Pressure (P0.1) | | |
| | | Maximal Sniff nasal Inspiratory Pressure (SNIP) | | | | |
| | | Maximum Inspiratory/Expiratory Pressures (MIP/MEP) | | | | |
| | Respiratory Status | Date of Non-invasive ventilation (NIV) start | | | | |
| | Neurophysiological Tests | Phrenic nerve response amplitude (PhrenMeanAmpl) | | | | |
| Phrenic nerve response latency (PhrenMeanLat) | | | | | | |
| Other physical values | Cervical Extension | | Cervical Flexion | | | |

* Scores and Sub-scores of the ALS Functional Rating Scale

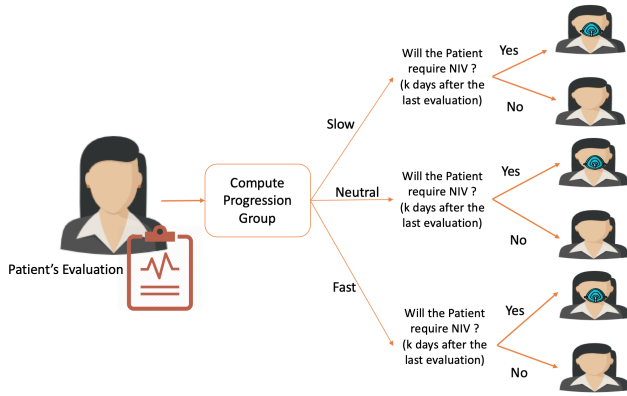


Fig. 3. Revised Problem Formulation using Disease Progression Groups: Knowing the patient current state, as well as his/her disease progression group can we predict the need for NIV in a given time window, using group specialized prognostic models?

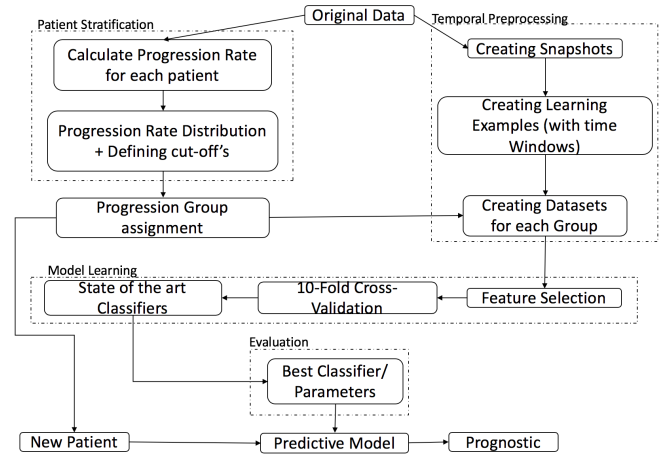


Fig. 4. Adapted Workflow of the proposed methodology for ALS prognostic prediction using patient snapshots and disease progression groups. Original data is preprocessed in order to create patient snapshots and learning instances. At the same time, original data is used to create the progression groups. The progression groups and learning instances are then used to create separate sets of data for each group. The models are built using a stratified 5 x 10-fold cross validation scheme, evaluated, and the best parameters are chosen. After the final model is trained, whenever a new patient is evaluated, his/her progression group is computed and his/her data is fed to the specialized prognostic model.

appointment). In original data, each exam or test is stored in its separate file, where temporal features are presented as a multivariate time series. After each appointment, a set of clinical tests are prescribed to the patient (to be presented at the next appointment). However, it is common for patients to not be able to perform all tests in the same day. This makes it difficult to merge all features into a single observation. To solve this problem, we follow the approach proposed by Carreiro et al. [7], that uses a bottom-up hierarchical clustering with constraints strategy to group exams and tests that are closer, creating a patient snapshot (a summary of the patient condition around that time). The constraints are: 1) two evaluations of the same test/exam cannot be in the same snapshot and 2) there cannot be snapshots where the patient uses NIV in some exams and does not in others, thus disrupting class coherence. At the end of this step we have a snapshot that is composed by static and temporal information of a patient (at the time around each appointment) and a class that tells us if the patient was or was not using NIV at the

time. This yields a dataset that can have multiple snapshots per patient, one for each appointment.

Finally, there is still one additional step to be performed: to create the learning instances from which the predictive models will learn from [7]. A single snapshot has only information about the NIV status at that specific time. However, what we want to know is the NIV status k days after the appointment time. To do this, we look to the next appointments of the same patient to create an Evolution class that takes the value 1 when the NIV status changes (from not needing NIV at the time of the appointment to needing NIV within the selected time window), and 0 when the NIV status does not change in that period of time. Snapshots where the patient already requires NIV at the current time and snapshots that have no

information about the NIV status after the time we are looking at cannot be used as learning examples and are thus discarded. As appointments are usually every three months, we use the time windows of 90, 180, and 365 days (3, 6, and 12 months respectively), as recommended by clinicians and the literature [11].

Figure 5 is an example of the preprocessing steps just described, adapted form [7].

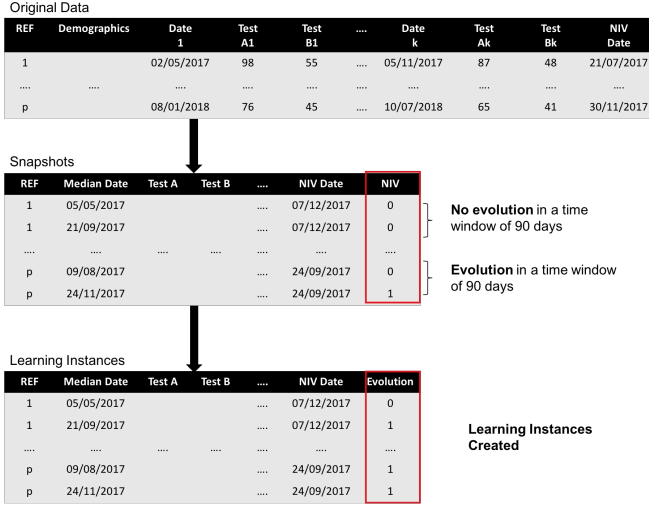


Fig. 5. Example of creating Snapshots and Learning Instances

2) *Creating Disease Progression Groups:* ALS is a highly heterogeneous disease that can present itself in various phenotypes. Some phenotypes, such as bulbar palsy, are usually associated with a worse prognosis, while flail leg or flail arm are usually associated with a better prognosis [4]. However, even with this knowledge, it is still difficult to understand disease progression, and what makes a patient progress faster or slower.

The ALS Functional Rating Scale (ALSFRS) is a standard test used by physicians to estimate the outcome of a treatment or the progression of the disease. Although very popular, this scale has only a small respiratory component. Given that respiratory failure is the most common cause of death in ALS patients, the ALS functional rating scale revised (ALSFRS-R) was proposed [12]. This new scale adds additional assessments of dyspnea, orthopnea, and need for ventilatory support, and quickly became the preferred test to quantify disease progression [13]. In this scenario, by measuring the change in ALSFRS-R over time, we can obtain an estimation of how the disease is progressing and infer about the survival of the patient [14]. In this work, we build upon this idea and use it to create three groups of patients with similar progression rate (with the time of first symptoms and the information about the patients in their first appointment). We compute the progression rate, using the following equation:

$$ProgressionRate = \frac{48 - ALSFRSR_{1stVisit}}{\Delta t_{1stSymptoms;1stVisit}}, \quad (1)$$

where 48 is the maximum score for the ALSFRS-R scale, $ALSFRSR_{1stVisit}$ is the ALSFRS-R score of a given patient in the first appointment (diagnosis) and $\Delta t_{1stSymptoms;1stVisit}$ is the time in months between the first symptoms and the first appointment.

We compute the progression rate for each patient. Then after analyzing the distribution of the progression rate from all patients in our dataset, and following consensual clinical recommendation, we divided the patients in three groups: 25% of the patients with lower or higher progression rates are grouped to create the Slow and Fast progressors groups; While the remaining 50 % of the patients with an average progression rate are grouped together and called Neutral progressors.

3) *Feature Selection:* Feature Selection (FS) methods are known for their capability to improve prediction performance [15]. These methods select a set of features that better describe data and reduce the effects of noisy and irrelevant features [16]. In clinical data, FS is especially useful for two reasons: 1) Data is usually high dimensional, with high number of missing values and usually with few observations (known challenges in machine learning, often called the curse of dimensionality and curse of sparsity [17]); 2) The set of chosen features can be of importance to clinicians to know which tests or exams are more important to the problem under study and also reduce the number and cost.

In this work, we follow the FS methodology proposed by [18]. They propose a FS Ensemble (FSE) that combines both stability and predictability to chose the best features for prognostic prediction in Alzheimer’s Disease. This ensemble combines multiple FS algorithms to return a set of features less biased by the characteristics of each FS algorithms. The FS algorithms used are: ReliefF, Information Gain, Conditional Mutual Information Maximization, Minimum Redundancy Maximum Relevance, and Chi-Squared. This methodology has two phases: first data is processed by FS algorithms to select a reduced set of features and then this set of features is optimized for stability and predictability [18].

B. Predictive Models

After creating learning instances for the main dataset (all patients) and each disease progression group (for the three time windows of 90, 180, 365 days), we use them to train the models that will allow us to predict whether or not a patient will need NIV k days after the current appointment.

Each dataset was imputed to 6 different classifiers while using stratified 5 x 10-fold cross-validation (CV) scheme [19]. The classifiers are: Decision Tree (DT), k-Nearest Neighbors (kNN), Support Vector Machines (SVM) with Polynomial (Poly) and Gaussian (G) kernels, Naïve Bayes (NB), Random Forest (RF), and Logistic Regression (LR). All classifiers used are available in Weka [20].

To deal with imbalanced data, we used a combination of Random Undersampling [21] and SMOTE [22] techniques. We used this combination due to two reasons: 1) by only undersampling the majority class to have equal size to the minority class, we would be using less data, thus impairing

classifiers (as well as their ability to generalize, due to the potential loss of variability in the data); 2) SMOTE creates synthetic learning instances for the minority class by using kNN to find similar instances to one of the learning examples. When dealing with a dataset that has a slight imbalance, this works well, since it creates synthetic instances close to the original ones, rather than simply oversampling them, which could lead to model overfitting. However, when dealing with a highly imbalanced dataset, in order to balance the data, the number of synthetic instances that are created easily surpasses the number of true instances of the minority class. As such, we randomly undersample the majority class, until we have a class imbalance of 60%/40% between the majority class and the minority class, and then use SMOTE to oversample the minority class until we achieve a balance of 50%/50%.

We performed a grid search to find the best parameters for each classifier. The parameters and corresponding ranges are detailed in Table II. The best parameters were chosen according to the best average AUC across the 5 x 10-fold CV classification results.

For model evaluation several metrics were retrieved, such as AUC, prediction accuracy, confusion matrix, sensitivity and specificity.

Regarding FS, we used the methodology presented before. We performed FS for each time window and disease progression group. We then created new datasets containing only the features selected. For the disease progression groups we also created datasets using the features selected from the main model, to understand if those features were better for prediction than the ones specific for each group. We then compared the models with and without FS using the Wilcoxon Signed-Rank Test for paired instances [23]. For the disease progression groups we also compared the results from using all features with the ones using the features selected for the baseline model (all patients).

TABLE II
PARAMETERS AND CORRESPONDING RANGES TESTED FOR EACH CLASSIFIER.

| Classifier | Parameter | Range |
|------------|-------------------|--|
| DT | Confidence factor | {0.15,0.20,0.25,0.30} |
| kNN | Nr of Neighbours | {1,3,5,7,9,11} |
| SVM P/G | Complexity | { $10^{-2}, 10^{-1}, 10^1, 10^2$ } |
| SVM P | Polynomial Degree | {1,2,3} |
| SVM G | Gamma | { $10^{-3}, 10^{-2}, 10^{-1}, 10^1, 10^2, 10^3$ } |
| NB | Kernel | {True,False} |
| RF | Nr of Trees | {5,10,15,20} |
| LR | Ridge Factor | { $10^{-9}, 10^{-8}, 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}$ } |

III. RESULTS

We first analyze the results from creating Patient Snapshots and Learning Instances for the time windows of $k = 90, 180,$ and 365 days. Then, we describe the creation of disease progression groups together with results and show the results for the FSE. Follows the results using the stratified 5x10 fold CV for the prognostic models. Finally, we compare the two approaches (Main and Progression Groups).

Our results are based on a cohort of 1220 ALS patients, described in Section II. For each patient we analyzed data as described in Table I. All patients were followed by the same clinician and all tests and exams were performed using the same procedure, to reduce clinical bias.

A. Creating Snapshots and Learning Instances

The approach used to create patient snapshots, consists in a bottom-up hierarchical clustering with constraints. From the 1220 patients only 1070 could be used to create snapshots, as the other 150 patients either had NIV at the time of the first appointment, or the information about their NIV status was unavailable.

Using this 1070 patients, we were able to create 5553 patient snapshots, resulting in an average of 5.19 snapshots per patient (15.57 months of follow-up data for each patient on average).

From these snapshots we created the learning instances for each time window to be used by the FSE and classifiers. This consisted in creating an Evolution class (E), where $E=1$ if the patient evolves to needing NIV in k days, and $E=0$ if the patient does not evolve to need NIV in that time period. Table III shows the results for each value of k .

TABLE III
STATISTICS AND CLASS DISTRIBUTION FOR TIME WINDOWS OF $k=90, 180$ AND 365 DAYS

| k | 90 | 180 | 365 |
|----------------------|---------------|---------------|---------------|
| Nr of Snapshots | 3178 | 3018 | 2762 |
| Nr of Patients | 861 | 823 | 775 |
| Snapshots p/ Patient | 3.69 | 3.67 | 3.56 |
| Evolution (E=1) | 559 (17.59%) | 906 (30.02%) | 1342 (48.59%) |
| No Evolution (E= 0) | 2619 (82.41%) | 2112 (69.98%) | 1420 (51.41%) |

Table III shows that the number of snapshots decreases with the increasing of k . This is due to the fact that for some snapshots we no longer have information about their NIV status within these time window. The same goes for the number of patients, since some patients require NIV soon after their first appointment. The class distribution shows that in the first time window of 90 days, less than 20% evolve to NIV. However, when in the window of 365 days, the number of NIV evolutions rises to 50%. This is expected since the probability of a patient requiring NIV increases when we consider longer periods of time, meaning that if we extended k , the number of instances where $E=1$ would continue to increase while instances with $E=0$ would keep decreasing. By extending k to a period long enough we believe a class distribution of 100%($(E=1)/0\%(E=0)$) would be reached.

B. Creating Disease Progression Groups

To create the disease progression groups we start by computing the progression rate for each patient. Using (1) we calculate the change in the ALSFRS-R scale over a period of time. We chose to use the change between the first symptoms and the first appointment (following clinical advice), since after the 1st appointment the progression rate usually remains constant. Out of the 1220 patients, only 1093 were used to create the progression groups, due to the fact that some

patients did not have information on either the ALSFRS-R scale or the interval between the first symptoms to the first visit (Diagnostic Delay).

Figure 6 shows the progression rate distribution for all patients. Similar results for the disease progression rate distribution were obtained by [24], although using the change in ALSFRS rather than ALSFRS-R, suggesting that the distribution obtained is not specific to our ALS population. We then divided the patients as aforementioned, by creating three disease progression groups: Slow, Neutral and Fast progressors. This resulted in 271, 552, and 270 patients in each group respectively. By using the whole population to create the groups, we allow their use for other problems, such as predicting other outcomes.

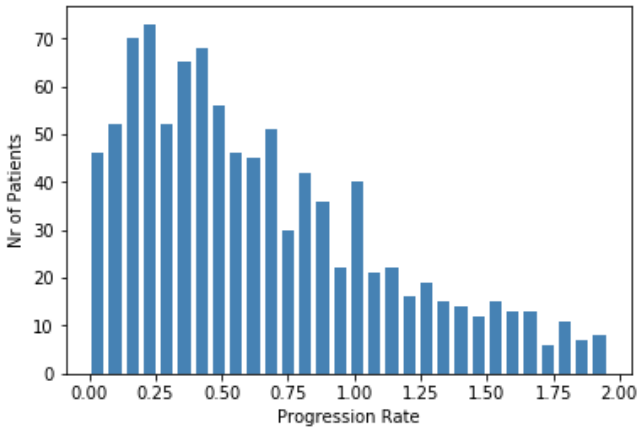


Fig. 6. Progression Rate Distribution among all patients.

After the progression groups were created, we merged the information each patient's group with the learning instances created before, to create datasets for each group and time window. The class distribution and statistics for each group are presented in Table IV, where we see that Slow progressors have more Snapshots per Patient than the other groups, and that Fast progressors have less snapshots than the other groups. This is in accordance with our expectations, as Slow progressors have a lower progression rate, meaning they survive longer, thus having more follow-up appointments. The number of patients for each time window is also higher in Slow progressors than in Fast progressors, despite the fact that they had almost the same number of patients when the groups were created. This is because Fast progressors evolve faster before k and some of them arrive to the clinic already using NIV, thus not being used to create learning instances.

Finally, by looking at class distribution we observe that, as before, the instances with $E=1$ tend to increase with the value of k . However, class distribution for Fast and Slow progressors differs reasonably from class distribution for the datasets with all patients. Neutral progressors have a class distribution closer to the overall population. Slow progressors have less instances with $E=1$, having in the longer k of 365 days less than 25% of evolutions (approximately half of what we have using all

instances). Regarding Fast progressors we observe the opposite from Slow progressors in terms of class distribution. Here the number of $E=1$ instances is in total $\sim 35\%$ for the smaller k of 90 days and goes up to $\sim 77\%$ for the longer k of 365 days.

TABLE IV
STATISTICS AND CLASS DISTRIBUTION FOR TIME WINDOWS OF
 $k=90, 180, 365$, FOR EACH DISEASE PROGRESSION GROUP

| k | 90 | 180 | 365 |
|------------------------|---------------|---------------|--------------|
| Slow Progressors | | | |
| Nr of Snapshots | 1242 | 1191 | 1090 |
| Nr of Patients | 215 | 210 | 200 |
| Snapshots p/ Patient | 5.78 | 5.67 | 5.45 |
| Evolution ($E=1$) | 88 (7.09%) | 163 (13.69%) | 269 (24.68%) |
| No Evolution ($E=0$) | 1154 (92.91%) | 1028 (86.31%) | 821 (75.32%) |
| Neutral Progressors | | | |
| Nr of Snapshots | 1459 | 1390 | 1278 |
| Nr of Patients | 441 | 425 | 399 |
| Snapshots p/ Patient | 3.31 | 3.27 | 3.20 |
| Evolution ($E=1$) | 328 (22.48%) | 527 (37.91%) | 801 (62.68%) |
| No Evolution ($E=0$) | 1131 (77.52%) | 863 (62.09%) | 477 (37.32%) |
| Fast Progressors | | | |
| Nr of Snapshots | 384 | 348 | 311 |
| Nr of Patients | 171 | 158 | 148 |
| Snapshots p/ Patient | 2.24 | 2.20 | 2.10 |
| Evolution ($E=1$) | 131 (34.11%) | 193 (55.46%) | 238 (76.53%) |
| No Evolution ($E=0$) | 253 (65.89%) | 155 (44.54%) | 73 (23.47%) |

C. Feature Selection

Table V shows the features selected for each dataset by the FSE. Most of the selected features across all datasets are recognized in the literature as prognostic indicators in ALS patients. Since we are predicting a respiratory target, we would expect that respiratory features would be more important than other features regarding other aspects of the disease. Those expectations were somewhat confirmed. However, demographic features are also relevant for the prognostic models.

There are differences between the features selected for each group. Slow progressors tend to need more features to build good prognostic models, while Fast progressors seem to rely on fewer features. Moreover, while in Slow progressors, for longer time windows the number of selected features diminishes, in Fast progressors it is exactly the opposite. In Neutral progressors, the selected features are actually the same for all values of k . The differences in the selected features between the progression groups is also important to the clinicians, since knowing which tests and exams are more important for each patient, can save time and resources, which in turn can lead to a better prognosis.

Features such as family history and region of disease-onset are described in literature as independent prognostic predictors for ALS. However, they were not selected by the FSE. This does not mean that those features are not important in the global view of ALS, but rather that they are not the best predictors for the problem presented in this work. Prognostic Features in literature are mainly from studies predicting survival and functional declines. However, they may not be the most useful for other approaches, datasets or outcome predictions.

TABLE V
SELECTED FEATURES

| Features | All Patients | | | Slow Progressors | | | Neutral Progressors | | | Fast Progressors | | |
|-------------------------------|--------------|-------|-------|------------------|-------|-------|---------------------|-------|-------|------------------|-------|-------|
| | 90 d | 180 d | 365 d | 90 d | 180 d | 365 d | 90 d | 180 d | 365 d | 90 d | 180 d | 365 d |
| Gender | | | | | | | | | | | | |
| Age at Onset | X | X | X | X | X | X | X | X | X | X | | X |
| BMI | X | X | X | X | X | X | X | X | X | X | X | X |
| Family History MND | | | | | | | | | | | | |
| Diagnostic Delay | X | X | X | X | X | X | X | X | X | | | |
| El Escorial Reviewed Criteria | | | | | | | | | | | | |
| UMN vs LMN | | | | | | | | | | | | |
| Onset form | | | | X | | | | | | | | |
| c9orf72 | | | | X | X | | | | | | | |
| ALS-FRS | X | X | X | X | X | X | X | X | X | | | |
| ALS-FRS-R | X | | X | X | X | X | X | X | X | | | |
| ALS-FRSb | X | | | X | X | | X | X | X | | | |
| ALS-FRSsUL | | | | X | X | | X | X | X | | | |
| ALS-FRSsLL | | | | X | X | | X | X | X | | | |
| ALS-FRSr | | | | X | X | | | | | | | |
| R | | | | X | X | | X | X | X | | | |
| VC | X | X | X | X | X | X | X | X | X | X | X | X |
| FVC | X | X | X | X | X | X | X | X | X | X | X | X |
| MIP | X | X | X | X | X | X | X | X | X | | | X |
| MEP | X | X | X | X | X | X | X | X | X | | | X |
| P0.1 | X | X | | X | X | | X | X | X | | | |
| SNIP | | | | | | | | | | | | |
| PhrenMeanLat | X | | | X | X | | X | X | X | | | |
| PhrenMeanAmpl | X | | | X | X | | X | X | X | | | |
| Cervical Flexion | | | | | | | | | | | | |
| Cervical Extension | | | | | | | | | | | | |

Although there are some differences in the selected features, there is a subset of features that is selected in almost all datasets. Age at onset, Body Mass Index (BMI), Disease Duration, ALSFRS, FVC, and Vital Capacity (VC) are some of the most frequent features, being selected at least 75% of the time. Three features are selected 100% of the time (BMI, FVC, and VC) making them the most important features to predict whether a patient will require NIV or not k days from his current appointment. It is no surprise that FVC is one of the most important features, since it is one of the most common tests used to evaluate respiratory declines in ALS [28].

D. Learning Predictive Models

We now look at the results of the predictive models. For each model, we followed a stratified 5 x 10-fold CV scheme, followed by a grid search to check for the best parameters for each classifier. The classifiers used were: Decision Trees (DT), k-Nearest Neighbors, Support Vector Machine (SVM) with Polynomial (P) and Gaussian (G) kernels, Naïve Bayes (NB), Random Forest (RF), and Linear Regression (LR). To evaluate the models, three metrics were chosen: AUC, Sensitivity, and Specificity. AUC is used to perform model comparisons, since it combines the results of the other two metrics.

We first run all the classifiers for the main models (with all snapshots for all patients) to build our baseline classifiers. Results are presented in Table VI.

These baseline results show that the performance of all classifiers improves for longer values of k. This can be due to the lower class imbalance for the longer time windows. Moreover, three of the classifiers (NB, RF, and LR) show better performances than the rest. These results can also be

TABLE VI
BASELINE RESULTS: AUC, SENSITIVITY, AND SPECIFICITY RESULTS FOR THE PROGNOSTIC MODELS FOR 90, 180, AND 365 DAYS. THE CLASSIFIERS ARE: DECISION TREES (DT), K-NEAREST NEIGHBORS, SUPPORT VECTOR MACHINE (SVM) WITH POLYNOMIAL (P) AND GAUSSIAN (G) KERNELS, NAÏVE BAYES (NB), RANDOM FOREST (RF) AND LINEAR REGRESSION (LR).

| | DT | kNN | SVM P | SVM G | NB | RF | LR |
|-------------|-------|-------|-------|-------|-------|-------|-------|
| UC | | | | | | | |
| 90d | 72.21 | 68.56 | 62.63 | 58.92 | 79.84 | 80.74 | 79.15 |
| 180d | 74.79 | 69.81 | 63.30 | 58.64 | 80.57 | 85.27 | 79.68 |
| 365d | 71.27 | 71.51 | 70.56 | 61.54 | 82.25 | 89.20 | 80.21 |
| Sensitivity | | | | | | | |
| 90d | 60.47 | 58.71 | 66.98 | 69.30 | 69.87 | 68.55 | 69.41 |
| 180d | 63.62 | 52.65 | 69.60 | 67.99 | 73.22 | 70.73 | 72.30 |
| 365d | 75.10 | 67.96 | 70.61 | 71.49 | 78.41 | 82.07 | 73.85 |
| Specificity | | | | | | | |
| 90d | 74.81 | 69.93 | 58.29 | 48.53 | 74.99 | 77.60 | 74.78 |
| 180d | 74.01 | 76.15 | 57.01 | 49.29 | 73.94 | 81.35 | 72.47 |
| 365d | 60.68 | 63.59 | 70.51 | 51.51 | 71.96 | 80.66 | 73.55 |

observed in the models proposed by [7]. However, the results obtained here outperform those models. This improvement can be due to the changes in the pipeline and/or by the larger amount of data used. Moreover, contrary to previous results, where there was a great imbalance in the values of sensitivity and specificity, ours are more balanced. This is a great improvement. Overall, the results obtained are promising for all time windows, with some measures above 80% and one result very close to 90%.

Although all classifiers present acceptable AUC values, for clarity sake and due to space restrictions, we use the best 3 classifiers (NB, RF and LR) for the following tests.

After training the baseline classifiers for each time window

using all features, we built the classifiers for the datasets, using the features selected by the FSE. We then compared them to see if the models with less features perform at least as good as the ones using all the information available. The results for this test are presented in the Table VII.

Comparing AUC results, we see that for most models, the results are better when using all features rather than using the set of selected features. FS methods usually have better results in situations where the number of features is higher than the number of observations. In our case, the number of observations is higher and thus the models benefit from using more features.

Although the results suggest that the models without FS are better than the ones with it, we performed the Wilcoxon Signed-Ranks Test for Paired Samples to check if the differences between them were statistically significant. The test yielded a p-value of 0.0284, meaning that the differences are indeed significant. Therefore, the baseline models used to predict the use of NIV should be the ones using all features.

We follow by using the same FS approach for the progression groups. We also tested how the models performed when using the features selected for the baseline results for each group to see how they compared to the specific feature set in terms of model performance. Table VIII presents these results.

By comparing with baseline results, we can see that in terms of AUC, baseline results are apparently better for the Fast and Neutral progressors and worse for Slow progressors. However, we highlight that this does not mean that the baseline models are better at predicting those groups than the actual specialized models. The results also seem to be better for the larger time windows, with an exception for the Fast progressors, where the window of 180 days shows the best results. This could be due to the decreasing number of learning instances that can be deteriorating the model's performance. This is also the group with the lower results, probably due to the difference in number of learning instances.

Similarly to the baseline models, the results using the FSE are overall lower than when using all features. Furthermore, the results with the selected features for the baseline models are also lower. As before, we used the Wilcoxon Signed-Ranks Test for Paired Samples to check if the differences were high enough to be statistically significant. We obtained p-values of 0.8288 and 0.0014 (not statistically significant and statistically significant) when comparing the all feature models to the results using the specific FS features and using the baseline FS features, respectively. Thus, we can conclude that using the baseline set of features is detrimental to the models performance in the progression groups. However, the differences between the no FS and the FS models for each group is not high enough to be statistically significant. As such, using the models with the FSE features should be preferred, since a lower number of features leads to simpler models, which generalize better.

E. Comparing Baseline and Disease Progression Groups

To better compare baseline and disease progression groups results, we trained the classifiers using all instances labeled with the progression group (not used in the classifiers). Then we retrieved the predictions for each instance and computed the confusion matrix and AUC to retrieve the necessary information to better compare the two approaches. The goal was to evaluate how well was the baseline model classifying patients from each group individually. Table IX shows the results for the NB classifier.

The AUC results for each group in Table VIII and Table IX are similar. However, when looking at sensitivity and specificity, we observe that the results are very different and groups are indeed relevant.

Regarding Slow progressors, sensitivity and specificity measures are highly imbalanced (very high specificity and very low sensitivity), meaning that the baseline model is correctly classifying negative instances and poorly classifying positive instances. The opposite can be seen with Fast progressors, where the baseline model correctly predicts almost all positive instances but incorrectly classifies the majority of negative instances. Moreover, results for the Fast progressors show that the AUC difference between approaches is higher than that of other groups. Once again this can be due to the fewer number of instances hindering the classifiers performance.

Neutral progressors show closer results between approaches. This is probably due to the fact that this group gives a higher contribution in terms of patients and learning instances to the baseline model, as well as being more representative of the average of the population. This means that the results for the baseline model generalize around the Neutral progressors, thus it predicts better the instances from that group (or the ones from the other groups that are closer to them). For the instances that are dissimilar, the model tries to predict the outcome, but generally fails. Using specialized models for the groups ensures that each model learns with a subset of patients that are similar to each other, generalizing around a less heterogeneous set of patients.

The results show that there are benefits in using the specialized models for the disease progression groups rather than the baseline ones, provided that we are able to compute the disease progression group of the patient. In this scenario, patient stratification thus proves useful when predicting NIV.

IV. CONCLUSIONS AND FUTURE WORK

We propose an approach using patient stratification, and prognostic prediction to tackle a known problem in ALS literature: prognostic prediction of NIV. We created three disease progression groups (Slow, Neutral and Fast progressors) to learn specialized prognostic models that predict the need of NIV within predefined time windows of 90, 180, and 365 days. The results are promising, achieving up to 91% in AUC for Slow progressors in 375 days. We also tested a Feature Selection Ensemble outputting set of features that are more important for the prognostic models. These features can help clinicians understand what are the best tests and medical

TABLE VII

BASELINE RESULTS: AUC, SENSITIVITY AND SPECIFICITY RESULTS FOR THE PROGNOSTIC MODELS FOR THE 90, 180 AND 365 DAYS TIME WINDOWS. ORIG IS THE ORIGINAL DATASET, FS IS THE DATASET WITH FEATURES SELECTED BY FEATURE SELECTION ENSEMBLE (FSE). THE CLASSIFIERS ARE: NAÏVE BAYES (NB), RANDOM FOREST (RF) AND LINEAR REGRESSION (LR).

| | | AUC | | | Sensitivity | | | Specificity | | |
|------|------|-------|-------|-------|-------------|-------|-------|-------------|-------|-------|
| | | NB | RF | LR | NB | RF | LR | NB | RF | LR |
| 90d | Orig | 79.84 | 80.74 | 79.15 | 69.87 | 68.55 | 69.41 | 74.99 | 77.6 | 74.78 |
| | FS | 79.05 | 77.48 | 79.29 | 76.06 | 71.34 | 74.78 | 67.25 | 68.23 | 69.98 |
| 180d | Orig | 80.57 | 85.27 | 79.68 | 73.22 | 70.73 | 72.3 | 73.94 | 81.35 | 72.47 |
| | FS | 77.51 | 82.02 | 78.21 | 74.61 | 74.5 | 73.77 | 66.36 | 73.45 | 68.5 |
| 365d | Orig | 82.25 | 89.20 | 80.21 | 78.41 | 82.07 | 73.85 | 71.96 | 80.66 | 73.55 |
| | FS | 79.91 | 88.95 | 80.46 | 74.63 | 82.07 | 75.02 | 70.77 | 80.9 | 72.21 |

TABLE VIII

RESULTS WITH DISEASE PROGRESSION GROUPS: AUC, SENSITIVITY AND SPECIFICITY RESULTS FOR THE PROGNOSTIC MODELS FOR 90, 180 AND 365 DAYS, WHEN USING DISEASE PROGRESSION GROUP. ORIG IS THE ORIGINAL DATASET, FS IS THE DATASET WITH FEATURES SELECTED FOR EACH PROGRESSION GROUP BY THE FEATURE SELECTION ENSEMBLE (FSE), AND ORIG IS THE DATASET USING THE FEATURES SELECTED FOR THE MAIN MODELS. THE CLASSIFIERS ARE: NAÏVE BAYES (NB), RANDOM FOREST (RF) AND LINEAR REGRESSION (LR).

| | | AUC | | | Sensitivity | | | Specificity | | | |
|----------------|------|---------|-------|-------|-------------|-------|-------|-------------|-------|-------|-------|
| | | NB | RF | LR | NB | RF | LR | NB | RF | LR | |
| Slow | 90d | Orig | 81.11 | 81.11 | 74.15 | 70.91 | 69.55 | 64.09 | 79.41 | 75.58 | 73.99 |
| | | FS | 82.24 | 79.77 | 79.17 | 72.27 | 68.86 | 70.45 | 76.79 | 73.97 | 75.06 |
| | | Orig FS | 80.60 | 76.48 | 77.91 | 77.05 | 70.91 | 71.59 | 73.29 | 69.84 | 72.69 |
| | 180d | Orig | 85.21 | 86.51 | 80.94 | 77.42 | 72.27 | 72.88 | 80.88 | 83.29 | 77.67 |
| | | FS | 84.78 | 85.72 | 82.50 | 75.58 | 74.11 | 73.50 | 79.73 | 80.64 | 78.79 |
| | | Orig FS | 81.07 | 82.11 | 80.64 | 79.14 | 72.76 | 75.21 | 70.04 | 75.86 | 73.99 |
| | 365d | Orig | 84.40 | 90.58 | 80.22 | 73.16 | 78.29 | 69.52 | 80.56 | 85.87 | 77.47 |
| | | FS | 82.27 | 88.92 | 82.71 | 75.54 | 80.00 | 73.01 | 75.44 | 82.39 | 75.98 |
| | | Orig FS | 82.24 | 88.66 | 82.77 | 75.61 | 79.33 | 73.38 | 75.42 | 81.78 | 76.22 |
| Neutral | 90d | Orig | 77.29 | 76.20 | 74.88 | 67.74 | 58.17 | 64.88 | 74.06 | 77.84 | 71.71 |
| | | FS | 76.34 | 74.94 | 77.43 | 68.66 | 62.87 | 69.57 | 71.18 | 72.25 | 72.36 |
| | | Orig FS | 75.75 | 72.74 | 76.06 | 70.30 | 60.24 | 68.84 | 67.06 | 70.47 | 69.05 |
| | 180d | Orig | 75.32 | 81.61 | 74.68 | 60.91 | 62.01 | 63.76 | 77.27 | 82.92 | 72.84 |
| | | FS | 75.78 | 80.24 | 76.62 | 66.72 | 65.24 | 66.22 | 73.02 | 78.61 | 73.42 |
| | | Orig FS | 72.42 | 77.78 | 71.89 | 67.13 | 60.46 | 65.88 | 64.26 | 77.71 | 66.56 |
| | 365d | Orig | 74.40 | 85.71 | 73.54 | 70.64 | 78.80 | 68.76 | 67.04 | 77.06 | 66.50 |
| | | FS | 75.27 | 83.93 | 75.43 | 59.48 | 68.39 | 68.29 | 79.45 | 82.01 | 71.78 |
| | | Orig FS | 71.74 | 84.10 | 72.22 | 60.85 | 66.77 | 65.74 | 73.04 | 82.39 | 68.89 |
| Fast | 90d | Orig | 72.69 | 71.82 | 74.94 | 63.66 | 51.30 | 61.37 | 74.39 | 76.68 | 77.39 |
| | | FS | 72.62 | 72.43 | 70.45 | 75.88 | 62.44 | 64.12 | 58.58 | 70.12 | 65.06 |
| | | Orig FS | 72.97 | 70.37 | 75.04 | 68.40 | 64.12 | 68.09 | 66.64 | 65.30 | 69.96 |
| | 180d | Orig | 71.48 | 81.23 | 70.56 | 65.28 | 70.57 | 68.81 | 68.00 | 75.61 | 61.55 |
| | | FS | 70.62 | 81.26 | 69.14 | 61.14 | 66.22 | 60.10 | 67.35 | 76.26 | 63.23 |
| | | Orig FS | 70.47 | 79.73 | 70.02 | 59.59 | 67.56 | 62.80 | 71.61 | 74.58 | 65.29 |
| | 365d | Orig | 65.60 | 79.41 | 65.02 | 64.96 | 74.37 | 68.15 | 57.26 | 71.23 | 51.23 |
| | | FS | 69.57 | 77.52 | 63.00 | 56.13 | 65.71 | 59.08 | 72.60 | 72.33 | 55.07 |
| | | Orig FS | 66.97 | 75.54 | 60.51 | 54.54 | 64.03 | 57.06 | 67.40 | 70.14 | 55.34 |

TABLE IX

DETAILS OF BASELINE RESULTS FOR EACH PROGRESSION GROUP: AUC, SENSITIVITY AND SPECIFICITY RESULTS FOR THE PROGNOSTIC MODELS FOR 90, 180 AND 365 DAYS RELATIVE TO EACH DISEASE PROGRESSION GROUP. THE CLASSIFIER IS NAÏVE BAYES (NB).

| | | Sensitivity | Specificity | AUC |
|------|---------|-------------|-------------|-------|
| 90d | Slow | 33.41 | 94.52 | 81.88 |
| | Neutral | 68.54 | 71.03 | 76.48 |
| | Fast | 85.04 | 34.7 | 68.36 |
| 180d | Slow | 40.86 | 95.91 | 85.31 |
| | Neutral | 69.15 | 68.76 | 75.22 |
| | Fast | 84.87 | 31.87 | 66.11 |
| 365d | Slow | 51.45 | 92.98 | 84.3 |
| | Neutral | 78.72 | 55.81 | 74.44 |
| | Fast | 90.84 | 31.87 | 66.37 |

exams to predict the need for NIV. The models created can be a useful tool for clinicians, either to reinforce the decision of prescribing or not NIV, or to help them decide when in doubt.

The proposed approach based on disease progression groups is not restricted to predict NIV. In future work we plan to apply it to prognostic prediction of other clinical outcomes, such as functional declines.

We reinforce the need for patient stratification when studying heterogeneous diseases such as ALS. We showed that when we learn with all patients (disregarding how different they are) in the same model, we risk to obtain a model that only performs well for a subgroup of patients and not for the whole population. Furthermore, by having specialized models to address different groups of patients, we are able to provide a more personalized care to each patient needs, thus improving prognosis and quality of life.

ACKNOWLEDGMENT

This work was partially supported by FCT, through funding of Neuroclinomics2 project (PTDC/EEI-SII/1937/2014) and the LASIGE Research Unit, (ref. UID/CEC/00408/2013). We are grateful to Genomed for performing the genetic tests in a subset of the included patients.

REFERENCES

- [1] M. A. van Es et al., Amyotrophic lateral sclerosis, *Lancet*, vol. 390, no. 10107, pp. 20842098, 2017.
- [2] R. H. Brown and A. Al-Chalabi, Amyotrophic Lateral Sclerosis, *N. Engl. J. Med.*, vol. 377, no. 2, pp. 162172, 2017.
- [3] E. Beghi et al., The epidemiology and treatment of ALS: Focus on the heterogeneity of the disease and critical appraisal of therapeutic trials, *Amyotroph Lateral Scler.*, vol. 12, no. 1, pp. 110, 2012.
- [4] S. Martin, A. Al Khleifat, and A. Al-Chalabi, What causes amyotrophic lateral sclerosis?, *F1000Research*, vol. 6, no. 0, p. 371, 2017.
- [5] O. Hardiman, L. H. Van Den Berg, and M. C. Kiernan, Clinical diagnosis and management of amyotrophic lateral sclerosis, *Nat. Rev. Neurol.*, vol. 7, no. 11, pp. 639649, 2011.
- [6] M. Georges, C. Morlot-panzini, T. Similowski, and J. Gonzalez-bermejo, Noninvasive ventilation reduces energy expenditure in amyotrophic lateral sclerosis, no. Vc, pp. 18, 2014.
- [7] A. V. Carreiro, P. M. T. Amaral, S. Pinto, P. Toms, M. de Carvalho, and S. C. Madeira, Prognostic models based on patient snapshots and time windows: Predicting disease progression to assisted ventilation in Amyotrophic Lateral Sclerosis, *J. Biomed. Inform.*, vol. 58, pp. 133144, 2015.
- [8] H. J. Westenberg et al., Prognosis for patients with amyotrophic lateral sclerosis: development and validation of a personalised prediction model, *Lancet Neurol.*, vol. 17, no. 5, pp. 423433, 2018.
- [9] J. Ganesalingam et al., Latent cluster analysis of ALS phenotypes identifies prognostically differing groups, *PLoS One*, vol. 4, no. 9, 2009.
- [10] R. Kffner et al., Crowdsourced analysis of clinical trial data to predict amyotrophic lateral sclerosis progression, *Nat. Biotechnol.*, vol. 33, no. 1, pp. 5157, 2015.
- [11] P. M. Andersen et al., EFNS guidelines on the Clinical Management of Amyotrophic Lateral Sclerosis (MALS) - revised report of an EFNS task force, *Eur. J. Neurol.*, vol. 19, no. 3, pp. 360375, 2012.
- [12] J. M. Cedarbaum et al., The ALSFRS-R: A revised ALS functional rating scale that incorporates assessments of respiratory function, *J. Neurol. Sci.*, vol. 169, no. 12, pp. 1321, 1999.
- [13] N. G. Simon et al., Quantifying disease progression in amyotrophic lateral sclerosis, *Ann. Neurol.*, vol. 76, no. 5, pp. 643657, 2014.
- [14] F. Kimura et al., Progression rate of ALSFRS-R at time of diagnosis predicts survival time in ALS, *Neurology*, vol. 66, no. 2, p. 265 LP-267, Jan. 2006.
- [15] I. Guyon and A. Elisseeff, An Introduction to Variable and Feature Selection, *J. Mach. Learn. Res.*, vol. 3, no. 3, pp. 11571182, 2003.
- [16] G. Chandrashekar and F. Sahin, A survey on feature selection methods, *Comput. Electr. Eng.*, vol. 40, no. 1, pp. 1628, 2014.
- [17] R. L. Somorjai, B. Dolenko, and R. Baumgartner, Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: Curses, caveats, cautions, *Bioinformatics*, vol. 19, no. 12, pp. 14841491, 2003.
- [18] T. Pereira, F. L. Ferreira et al., Neuropsychological predictors of conversion from Mild Cognitive Impairment to Alzheimers Disease: A feature selection ensemble combining stability and predictability, Under Submission, 2018
- [19] R. Kohavi, A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection, *Appear. Int. Jt. Conf. Artificial Intell.*, vol. 5, pp. 17, 1995.
- [20] M. A. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, The WEKA data mining software: an update, *SIGKDD Explor.*, vol. 11, no. 1, pp. 1018, 2009.
- [21] H. He and E. A. Garcia, Learning from imbalanced data, *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 12631284, 2009.
- [22] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, *J. Artif. Intell. Res.*, vol. 16, pp. 321357, 2002.
- [23] D. Rey and M. Neuhuser, Wilcoxon-Signed-Rank Test, in *International Encyclopedia of Statistical Science*, M. Lovric, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 16581659.
- [24] M. Proudfoot, A. Jones, K. Talbot, A. Al-Chalabi, and M. R. Turner, The ALSFRS as an outcome measure in therapeutic trials and its relationship to symptom onset, *Amyotroph. Lateral Scler. Front. Degener.*, vol. 17, no. 56, pp. 414425, 2016.
- [25] N. G. Simon, W. Huynh, S. Vucic, K. Talbot, and M. C. Kiernan, Motor neuron disease: Current management and future prospects, *Intern. Med. J.*, vol. 45, no. 10, pp. 10051013, 2015.
- [26] H. Creemers, H. Grupstra, F. Nollet, L. H. van den Berg, and A. Beelen, Prognostic factors for the course of functional status of patients with ALS: a systematic review, *J. Neurol.*, vol. 262, no. 6, pp. 14071423, 2015.
- [27] M. I. Polkey, R. A. Lyall, K. Yang, E. Johnson, P. Nigel Leigh, and J. Moxham, Respiratory muscle strength as a predictive biomarker for survival in amyotrophic lateral sclerosis, *Am. J. Respir. Crit. Care Med.*, vol. 195, no. 1, pp. 8695, 2017.
- [28] N. Lechtzin et al., Respiratory measures in amyotrophic lateral sclerosis, *Amyotroph. Lateral Scler. Front. Degener.*, vol. 19, no. 56, pp. 110, 2018.