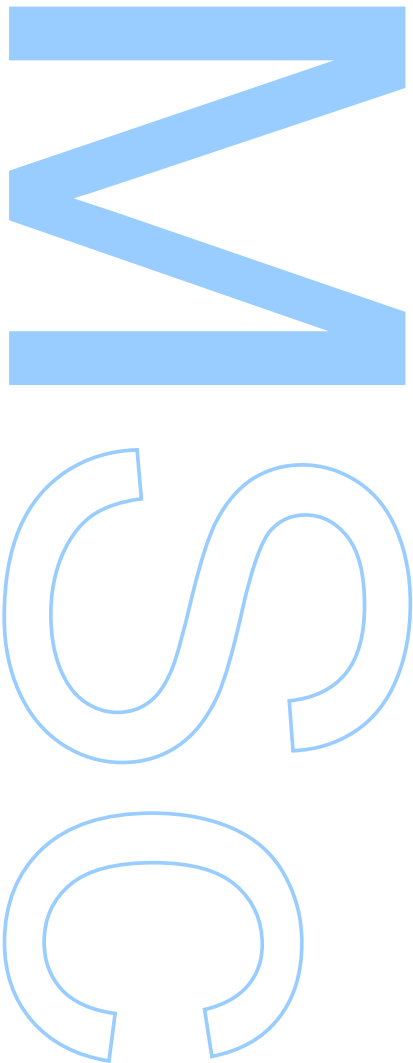
Modeling TTR-FAP Age of Onset survival curves using Mixture Density Networks, Subgroup Discovery and Sensitivity Analysis

Mariana Monteiro
Mestrado em Ciência de Dados
Departamento de Ciência de Computadores
2021

Orientador
Prof. Dr. Alípio Jorge, Faculdade de Ciências



U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____

UNIVERSIDADE DO PORTO

MASTERS THESIS

Modeling TTR-FAP Age of Onset survival curves using Mixture Density Networks, Subgroup Discovery and Sensitivity Analysis

Author:
Mariana MONTEIRO

Supervisor:
Alípio JORGE

A thesis submitted in fulfilment of the requirements
for the degree of MSc. Data Science

at the

Faculdade de Ciências da Universidade do Porto
Departamento de Ciência de Computadores

June 24, 2021

Acknowledgements

This dissertation would not have been completed without the help and support of some people. First, to Professor Alípio Jorge, for all the guidance and patience throughout the course of this year. To Maria Pedroto, for her encouragement, kindness and advice provided. Then, to Dra. Teresa Coelho, for her unquenchable drive to fight a terrible disease. To my family and to Inês and Leonor, who were always there. And last but not least, to my boyfriend, for all the the support given and for the patience to read this dissertation countless times.

I thank you all.

UNIVERSIDADE DO PORTO

Abstract

Faculdade de Ciências da Universidade do Porto
Departamento de Ciência de Computadores

MSc. Data Science

Modeling TTR-FAP Age of Onset survival curves using Mixture Density Networks, Subgroup Discovery and Sensitivity Analysis

by Mariana MONTEIRO

Transthyretin Related Familial Amyloid Polyneuropathy is a hereditary neurodegenerative disease that highly affects the quality of life of its bearers, being sometimes fatal. So far, there is no certain treatment for this disease, since the existing solutions rely on an as-early-as-possible diagnosis, and this disease proves itself challenging to diagnose. Because of this, knowing the age interval in which a patient may start developing symptoms - their interval of Age of Onset - can be helpful in beginning to establish a treatment plan. In this work, we use Mixture density Networks together with Subgroup Discovery and Sensitivity Analysis to model the survival curve of age of onset of patients with TTR-FAP, while providing an understandable characterization of the predictions made. The results of this work show that the predictive modeling capability of the Mixture Density Networks is comparable, and at times superior, to other state-of-the-art methods in survival analysis, and that Subgroup Discovery and Sensitivity Analysis are powerful tools in increasing the interpretability of an otherwise black-box model.

UNIVERSIDADE DO PORTO

Resumo

Faculdade de Ciências da Universidade do Porto
Departamento de Ciência de Computadores

Mestrado em Data Science

Modelação de curvas de sobrevivência de Idade de Onset de PAF utilizando Mixture Density Networks, Descoberta de Subgrupos e Análise de Sensibilidade

por Mariana MONTEIRO

Polineuropatia Amiloidótica Familiar, conhecida em Portugal por Doença dos Pezinhos, é uma doença hereditária e neurodegenerativa que afeta severamente a vida de quem a herda, podendo ser fatal. Até hoje, não há tratamento fidedigno para esta doença, uma vez que as soluções existentes dependem de um diagnóstico precoce, e esta doença demonstra-se bastante difícil de diagnosticar. Por consequência, ter conhecimento prévio do intervalo de idades em que um paciente pode desenvolver sintomas - o seu intervalo de idades de *onset* - pode ser útil na definição de um plano de tratamento. Neste trabalho, utilizamos *Mixture Density Networks* juntamente com Descoberta de Subgrupos e Análise de Sensibilidade de maneira a modelar a curva de sobrevivência de idade de *onset* de um paciente e, em simultâneo, facultar uma caracterização perceptível da previsão feita. Os resultados deste trabalho mostram que a capacidade de modelação preditiva das *Mixture Density Networks* é comparável, e por vezes superior, a outros métodos correntes para análise de sobrevivência, e que a Descoberta de subgrupos e Análise de Sensibilidade são ferramentas poderosas no aumento da explicabilidade de um modelo que por si só seria uma *black-box*.

Contents

Acknowledgements iii

Abstract v

Resumo vii

Contents ix

List of Figures xi

List of Tables xiii

Glossary xv

1 Introduction 1

2 Background 5
2.1 Familial Amyloid Polyneuropathy . . . 5
2.2 Electronic Health Record Data . . . 7
2.3 Subgroup Discovery . . . 8
2.3.1 Elements of a subgroup discovery methodology . . . 8
2.3.2 Distribution Rules . . . 10
2.4 Sensitivity Analysis . . . 10
2.4.1 Global sensitivity analysis . . . 10
2.4.2 Individual sensitivity analysis . . . 11
2.5 Survival Analysis . . . 11
2.5.1 Censoring . . . 12
2.5.2 Survival Curves . . . 12
2.6 Mixture Density Networks . . . 14
2.6.1 Artificial Neural Networks . . . 15
2.6.2 Gaussian Mixture Models . . . 16
2.6.3 Mixture Density Networks . . . 17

3 State of the Art 19
3.1 Subgroup discovery . . . 19
3.2 Survival Analysis . . . 22
3.2.1 Bayesian Methods . . . 23

3.2.2 Neural Networks . . . 23
3.2.3 Random Survival Forests . . . 24
3.2.4 Support Vector Machines . . . 24
3.3 Prediction of age of onset . . . 24

4 Modeling the Survival Curves **27**
4.1 Architecture . . . 27
4.2 The NaN problem . . . 28
4.3 Survival Curves . . . 30
4.4 Performance Evaluation . . . 32
4.5 Results . . . 35
4.5.1 Pointwise Prediction . . . 35
4.5.2 Survival curve prediction . . . 36

5 Beyond Predictions **39**
5.1 Overview . . . 39
5.2 Subgroup Discovery . . . 40
5.2.1 Subgroup Visualization . . . 41
5.2.2 Results . . . 43
5.3 Sensitivity Analysis . . . 44
5.3.1 Global and Subgroup sensitivity analysis . . . 45
5.3.2 Individual sensitivity analysis . . . 46
5.3.3 Subgroups . . . 47
5.3.4 Results . . . 48

6 Discussion **51**

7 Conclusion **55**

Bibliography **57**

List of Figures

2.1 Symptoms of TTR-FAP (adapted from [14]). . . . 6
2.2 From subjects to survival curves. . . . 13
2.3 A representation of a survival curve. . . . 13
2.4 Architecture of a Mixture Density Network (adapted from [53]). . . . 14
2.5 Representation of an Artificial Neural Network and its basic unit. . . . 16
2.6 Interval of values obtained by the Negative Log-Likelihood loss function. . . . 18

3.1 Example of a confounding (A) and not confounding (B and C) scenarios (adapted from [80]). . . . 21
3.2 Types of Survival Analysis methods. . . . 22

4.1 Architecture of the Mixture Density Network. . . . 28
4.2 Curves obtained for a patient. . . . 30
4.3 Survival curves of a selected group of patients. . . . 31
4.4 Calibration measurement approach. . . . 34
4.5 Calibration measure for each dataset. . . . 37

5.1 Exemplification of the proposed model. . . . 40
5.2 Process of obtaining the subgroup of a patient. . . . 41
5.3 Probability density functions and survival curves of Age of Onset for a selected group of patients. . . . 42
5.4 KS metric for all the subgroups obtained, per dataset. . . . 44
5.5 Approach for sensitivity analysis. . . . 45
5.6 Global and subgroup sensitivity analysis. . . . 45
5.7 Individual sensitivity analysis. . . . 46
5.8 KS values obtained for TTR-FAP subgroup-patient pairs using Subgroup Discovery and Sensitivity Analysis. . . . 47
5.9 Global, subgroup (1) and patient (206) sensitivity analysis. . . . 48
5.10 Feature importance obtained with Sensitivity Analysis and with Random Survival Forests. . . . 49

List of Tables

2.1 Types of target variables and consequent strategies (adapted from [31]) . . . 9

4.1 Dataset information . . . 33
4.2 Model hyperparameters . . . 33
4.3 Pointwise prediction evaluation of the tested Machine Learning regression models . . . 35
4.4 C-index of the tested survival analysis models . . . 37

5.1 Kendall rank coefficient between the MDN and the Survival Random Forests 49

Glossary

ALS	Amyotrophic Lateral Sclerosis
ANN	Artificial Neural Network
AOO	Age of Onset
ARDP	Adaptive Refinement by Directed Peeling
BN	Bayesian Networks
CART	Classification and Regression Trees
DNN	Deep Neural Network
EHR	Electronic Health Record
HIV	Human Immunodeficiency Virus
IT	Interaction Trees
K-M	Kaplan-Meier
MDN	Mixture Density Network
MOB	Model-based Recursive Partitioning
ML	Machine Learning
NB	Naïve Bayes
RECPAM	Recursive Partition and Amalgamation
RMSE	Root Mean Squared Error
RSF	Random Survival Forests
SA	Sensitivity Analysis
SIDES	Subgroup Identification Based on Differential Effect Search
STIMA	Simultaneous Threshold Interaction Modeling Algorithm

SD	Subgroup Discovery
SUBIC	Supervised Bi-Clustering
TTR	Transthyretin
TTR-FAP	Transthyretin-Related Familial Amyloid Polyneuropathy

This thesis is dedicated to Professor Francisco Pereira, who taught me that a Biochemist can chose any path she wants to.

Chapter 1

Introduction

Transthyretin-Related Familial Amyloid Polyneuropathy (TTR-FAP) is a neurodegenerative disease caused by mutations of the Transthyretin (TTR) gene Cakar et al. [1]. Endemic to Portugal, Sweden and Japan, it is mostly hereditary, with some sporadic mutations having already been identified [2, 3]. Amyloid-based diseases such as this one cause increasing damage to the patients' nervous system, often having a fatal outcome [4]. Even when it is not fatal, TTR-FAP severely affects the patients' nerves, to the point of highly deteriorating their quality of life.

In Medical Sciences, Age of Onset (AOO) is a core concept defined as the age at which a patient first experiences or acquires the symptoms of a given disease [5]. AOO is especially important when dealing with diseases such as TTR-FAP, which do not have a possible treatment and, as such, rely on as-early-as-possible diagnosis in order to provide a better quality of life to the patients [6].

Even though an early diagnosis of symptom appearance is the first step in fighting TTR-FAP, three factors make diagnosing this disease a demanding task. First, a lot of symptoms characteristic to TTR-FAP are also common to an array of different diseases, making it frequent for physicians to view the symptoms as if they were from a more trivial disease. Then, the used diagnosis methods, besides costly and sometimes invasive, might still not point right away to TTR-FAP. Because of this uncertainty, and constituting the third factor, a multitude of tests is required for medical professionals to be certain of the presence of this disease [1]. These three factors combined have proven to be a hardship in diagnosing patients, even after symptom appearance [7]. The already difficult diagnosis is even more aggravated as studies conclude that, especially in Portugal, a late-onset shift is being observed, requiring patients to be followed for a longer period of time to be

properly diagnosed [8]. Because of this, being able to know the age interval in which a patient may start developing symptoms can be helpful in beginning to establish their treatment plan.

Aiming at a personalised prediction for the age of onset, we propose a Machine Learning (ML)-based approach to model the survival curve of a patients' AOO, using Mixture Density Networks (MDN). In order to understand better the clinical and genealogical characteristics of patients related to the age of onset, we also integrate a Subgroup Discovery (SD) and Sensitivity Analysis (SA) tool in our methodology. By joining these tasks, this approach is able to provide not only a model of the (AOO) but also an understandable characterization, as patients are computationally grouped according to genealogical data. In practice, we aim at answering the following four research questions:

- **RQ-1:** How to obtain a human readable characterization of the patient subgroup, given a prediction?
- **RQ-2:** On their own, can Subgroup Discovery or Sensitivity Analysis provide enough information to obtain a robust model of Age of Onset?
- **RQ-3:** How to produce a Machine Learning model able to accurately predict personalized survival curves for the Age of Onset and similar problems?
- **RQ-4:** Is it possible to use Mixture Density Networks together with Subgroup Discovery and Sensitivity Analysis to estimate survival curves that are human readable?

Our approach was validated on patient data related to four different diseases, namely: TTR-FAP, Parkinson's, Amyotrophic Lateral Sclerosis and Cardiovascular disease. Its performance was compared to seven ML regression models and to two survival analysis models. Results show that this approach is not only able to model patients' AOO with the same performance as the tested models, but it is also able to outperform them in several situations. Additionally, SD and SA provided detailed information on the genealogical characteristics that translated into a specific AOO survival curve (i.e., the genealogical risk factors) both for specific patients and for subgroups.

The main contributions of this dissertation are the following:

- The definition of a grey-box approach to obtain human-readable survival curves of Age of Onset.

- The creation of a public GitHub repository with the practical methodology to develop human-readable survival curves, readily available in [9].
- An oral presentation in Encontro de Investigação Jovem da Universidade do Porto (IJUP).

This work is organized as follows. Chapter 2 provides some fundamental background on concepts regarding TTR-FAP, Subgroup Discovery, Sensitivity Analysis, Survival Analysis and Mixture Density Networks. Chapter 3 presents an analysis of the state-of-the-art on Subgroup Discovery, Survival Analysis and prediction of Age of Onset. Chapter 4 showcases the methodology and consequent results of the modeling of survival curves of age of onset, and Chapter 5 presents the approach used to provide a characterization of the predictions made. In, Chapter 6 we discuss the main findings while answering the proposed research questions, and Chapter 7 concludes this work.

Chapter 2

Background

In this chapter, we provide background on essential concepts regarding the main topics of this dissertation. First, in Section 2.1, we introduce the disease at focus, Transthyretin-Related Familial Amyloid Polyneuropathy (TTR-FAP), and the challenges in its diagnosis. Then, in Section 2.2, we showcase the difficulties in working with medical data, especially with Electronic Health Record (EHR) data. In Sections 2.3 and 2.4 we explain the goal of Subgroup Discovery (SD) and Sensitivity Analysis (SA) and their advantages in the context of disease diagnosis. Finally, in Sections 2.5 and 2.6, we focus on the core concepts of the approach used, namely SA and Mixture Density Networks (MDN).

2.1 Familial Amyloid Polyneuropathy

First documented in 1952 by Dr. Mário Corino de Andrade in Hospital Santo António in Portugal [10], TTR-FAP is an autosomal dominant amyloidosis caused by mutations of the Transthyretin (TTR) gene, leading to the deposit of amyloidogenic transthyretin, an insoluble protein-derived material, in tissues and organs [1]. It is generally paired with fatal outcomes, registering, on average, a life expectancy of 10 years since the symptoms appear [11].

TTR-FAP is a progressive genetic disease which, although endemic to Portugal, Sweden and Japan, has already registered cases all over the world [12]. While most types of this disease are hereditary, a wild-type (i.e., non-hereditary, sometimes referred to as sporadic-type) form of it was also observed with cases around the world [3]. In order to better understand the magnitude of the disease, it is necessary to introduce two concepts: incidence and prevalence. The former specifies the rate of occurrence of new cases,

while the latter measures the quantity of affected individuals at a certain point in time [13]. Having these into account, in Europe, TTR-FAP has an incidence of 0.3 new cases per year per 1 million inhabitants and a prevalence of 5.2 cases per 1 million inhabitants, classifying it as a rare disease [4].

Despite the fact that the disease is considered rare, it is estimated that its prevalence may be much higher. This estimation is made based on two main factors: first, due to the wide range of symptoms (Figure 2.1) that can be common to other diseases, and second, due to the lack of proper diagnosis in early stages [7]. This scarcity in proper diagnosis is mostly attributed to three factors. For one, the symptoms characteristic to TTR-FAP are also characteristic to a multitude of other diseases, making them non-specific and therefore hard to associate to this disease specifically [11]. Then, the used diagnostic methods (i.e., genetic testing and biopsy of tissues) do not always present clear answers on the exact disease presented by a patient. Because of this, and constituting the third factor, a panoply of assessments is required after these tests in order to be certain of the presence of this disease [1].

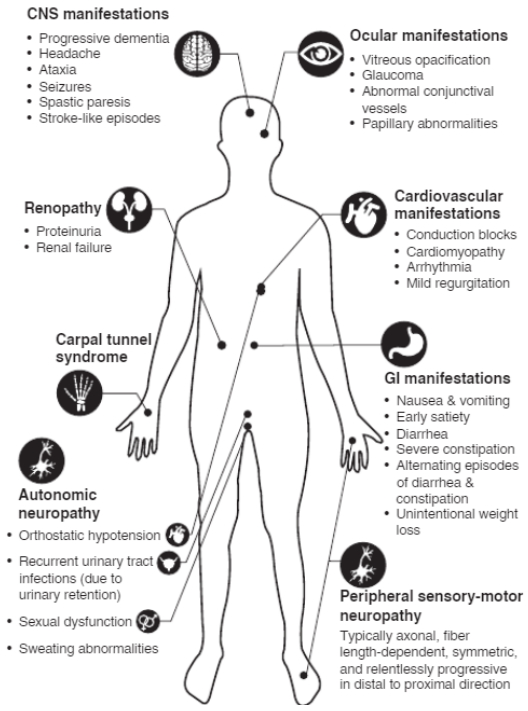


FIGURE 2.1: Symptoms of TTR-FAP (adapted from [14]).

In defiance of this reality, as the number of diagnosed patients increases, and, simultaneously, more documentation and knowledge are gathered, more tools are added to the

treatment and management of TTR-FAP. At the time of writing, two of the most groundbreaking advances in its combat should be noted. First, the liver transplant, as TTR is mostly formed in the liver [15]. This procedure has proved to be effective in slowing or even stopping the progression of this disease [16]. Then, Tafamidis, Patisiran and Inotersen, three drugs that also showed positive results in stopping the disease progression [17]. Although these constitute respectable advances in managing TTR-FAP, diagnosis keeps proving itself as an indispensable aspect in the battle against it, as most therapies are only reliable for patients with an early-stage disease [18].

In Medicine, Age of Onset (AOO) is referred to as the age at which a patient experiences for the first time the symptoms of a disease [5]. AOO constitutes a vital information for health professionals, especially when dealing with patients with neurodegenerative fatal diseases like TTR-FAP. Because of this disease having no dependable treatment and patients relying only on symptom management, it is a priority to provide therapies as soon as possible. Having in consideration the aforementioned complications in the early diagnosis of TTR-FAP, modeling the AOO appears as a solution to receive early assistance and better develop a therapy plan.

2.2 Electronic Health Record Data

sec:bEHR data consists in digitized longitudinal information about a specific patient or set of patients, allowing physicians to track their medical history. The rise in this type of patient tracking in hospitals paved the way to the use of Machine Learning (ML) and data analysis tools in this type of data. However, some challenges also surfaced with its growing use [19].

One of the main problems with EHR data is the high frequency of patient censoring (i.e., those with an incomplete medical history), as a patient can suddenly stop being followed, being due to death or dropout (e.g., in the case of clinical trials), amongst a panoply of other reasons. Even if a patient is always followed, there is always the chance of missing data, as data is added manually by physicians, and, as such, it is prone to human error. Even if these issues were not present, medical problems often require a high level of patient personalization when offering solutions. This can be a challenge when dealing with highly heterogeneous data like EHR, as finding patterns within patient data can turn into a difficult task [20].

Above all of these setbacks, the biggest challenge with this type of data is the reluctance that still exists in medical facilities to digitize all patient information. This means that the data provided can be incomplete without the user's knowledge. Because of everything here mentioned, this data deserves special attention when being modeled. Not only is the cleaning and preparation of data important but also the choice of the correct ML model to be able to overcome the challenges it poses [21].

2.3 Subgroup Discovery

Introduced by Kloesgen [22] and Wrobel [23], SD is a widely used data mining technique with the purpose of extracting statistically interesting relationships between variables with respect to a specific target [24]. Data mining methodologies can be categorized in two groups, depending on their goals: predictive and descriptive induction.

- **Predictive Induction** aims to discover knowledge by predicting or classifying an unknown object. In this type of induction we find most of ML methods, namely classification and regression methods. Another method in this category is time series [25].
- **Descriptive Induction** tries to extract interesting knowledge from unlabelled data (i.e., data without the target information) [26].

In the literature, one can observe that SD does not fit entirely in either of these categories, as it has the goal of extracting interesting knowledge from labelled data (i.e., according to a property of interest) [27]. This characteristic is what allows for the explicability of SD models (i.e., ability of a model to be human-readable), allowing SD to be applicable in a vast set of areas, such as finance [28], medicine [29] or simply day-to-day problems [30].

2.3.1 Elements of a subgroup discovery methodology

The usage of an SD approach requires special attention to some key aspects:

- *Target Variable:* One thing to consider right away when performing SD is the type of target one is working with, as it can influence the overall strategy to be applied. These can be observed in Table 2.1.

TABLE 2.1: Types of target variables and consequent strategies (adapted from [31])

Target type	Strategy
Binary	Finding subgroups for each of the possible values
Nominal	Similar to the binary approach, but subgroups are found for the different categories
Numeric	Target is discretized in intervals and subgroups are provided according to them

- *Search Strategy:* The strategy to generate the candidate subgroups. Even though a lot of strategies have been deployed throughout time, three of them stand out as the most widely used.

  - **Exhaustive Search:** Generates all possible candidates taking into account previously defined constraints. This type of search can be computationally costly, as the cost is proportional to the amount of candidates [32].
  - **Beam Search:** Only a selection (referred to as beam width) of the best partial candidates is taken into consideration. By exploring only a part of the search space, beam search does not guarantee a solution, unlike exhaustive search [33].
  - **Genetic Algorithms:** An evolutionary methodology in which, just like in natural evolution, the solutions with the best fitness measure can evolve. This method proved to have particular advantages when compared to the previous two [34].

- *Pruning:* After candidate generation, it is vital that only the significant candidates are kept. The most used types of pruning are coverage pruning, optimistic pruning and constraint pruning. Helal et. al [25] presented a thorough survey with the existing methodologies for SD and the search strategy and pruning method each of them employ.

- *Quality measures:* Used as the final step in an SD algorithm, the usage of a quality measure to rank the obtained subgroups allows for a post-processing of the best candidates by assigning a numerical value to a subgroup. Depending on the objective of the algorithms applied, different quality measures can be used. Two popular quality measures are the Piatetsky-Shapiro [22] and the unusualness [35].

2.3.2 Distribution Rules

Subgroups are represented under the form of rules. For this specific problem, we use distribution rules, a type of association rules that results in a distribution [36]. Distribution rules are defined as follows [26]:

$$A \rightarrow y = D_{\mathrm{y—A}}$$

with A being a set of items, y the property of interest (i.e., the target) and $\mathrm{D_{y—A}}$ the empirical distribution of y when A is observed. It is represented as a pair of y_j/*freq*(y_j), with $\mathrm{y_j}$ being a value of y and freq($\mathrm{y_j}$) the frequency of $\mathrm{y_j}$ when A is observed.

In an empirical clinical example comprising characteristics of patients and their heart rate, we could have the following distribution rule:

$$female \wedge smoker \rightarrow \{80/5, 90/3, 100/2, 110/4\} \tag{2.1}$$

This rule represents the conjunction of female smoker in all the data, and tells us that 5 have heart rates of 80bpm, 3 of 90bpm, 2 of 100 bpm and 4 of 110bpm, which could be represented under the form of a distribution.

2.4 Sensitivity Analysis

In ML, the sensitivity of a model is related to how the output is affected by changes in the input [37]. Therefore, SA consists in varying the input of the model slightly, and registering the change that occurs in the output. This is particularly interesting in neuronal methodologies, as they're usually considered to be black boxes, and therefore explaining their predictions can turn into an arduous task. With this analysis, it is possible to obtain knowledge on which features have a higher impact on a given prediction. Sensitivity can be analysed both from a global and local standpoint.

2.4.1 Global sensitivity analysis

Globally, SA can be used almost as a feature importance algorithm, as it allows for the ranking of the features according to their effect on the target variable. To this effect, two types of analysis are commonly more used, the input perturbation algorithm and the

partial derivative algorithm. Both of these have the same basal concept: a feature is more important when it causes a greater change in the model's output.

The partial derivative algorithm makes use of the Jacobian matrix, which computes the first order derivatives of the outputs regarding the inputs [38]. As a result, this method is particularly useful when in use of neural networks that apply first-derivative activation functions, such as back-propagation neural networks [39]. The input perturbation algorithm, contrary to the latter, has a more general approach. Here, a perturbation (Δx) is applied to the input of the network (x), and the effects on the output are registered, represented in Equation 2.2.

$$x = x + \Delta x \tag{2.2}$$

This effect is observed in the error of the prediction, with the most important feature being the one that was able to cause a higher shift in the error [40].

2.4.2 Individual sensitivity analysis

For individual sensitivity analysis, the goal is to find, for each patient, the set of feature values that leads to the same prediction as the original values. The general concept is similar to the input perturbation method, only the effect on the output is not measured in terms of error but rather in terms of the KS metric (for further detail on this change, please refer to Subsection 5.3).

2.5 Survival Analysis

Survival Analysis encompasses a wide variety of statistical methods that have as their goal the estimation of time until an event (also in the literature as failure) happens (i.e., time-to-event) [41]. Even though we can see a panoply of events being studied with SA, the ones that are more commonly used in the medical context are age of onset, relapse, recovery or death [42]. The main difference between survival analysis and regular statistic approaches is that the former accounts for censored data.

2.5.1 Censoring

Censored data comprises individuals that did not meet the event at study. As an example, imagining we are studying the time until a patient develops metastases following a successful cancer therapy, the event would be the appearance of metastases. This event may not be reached by some patients, either because of death or because they stayed healthy, among others. How survival analysis deals with incomplete data constitutes its biggest advantage, as, unlike standard statistical analysis, it is not treated as missing data, and, therefore, not discarded [43].

Censoring can appear mainly under the form of right-censoring, left-censoring or interval-censoring. Right censoring is the most common type, as it happens when the event being studied is not achieved. Left-censoring is a more complex case, corresponding to cases where the true survival time may be less than the observed survival time (e.g., when the event is the time a patient tests positive for a virus but we do not know exactly when he was exposed to it). Both right-censoring and left-censoring constitute point censoring, as they present information on either the beginning or the end of the survival time. Contrarily to this, interval censoring happens when the time-to-event is known but only within an interval [44].

2.5.2 Survival Curves

In order to deal with censoring, in 1958 Kaplan and Meier established a methodology which nowadays is the most widely used approach to deal with incomplete observations, the Kaplan-Meier (K-M) estimate [45]. This strategy considers time by dividing it into a set of smaller intervals, thus allowing for the integration or discarding of individuals throughout the length of a study [46]. For this reason, this method is defined as the probability of survival in a given length of time, while considering time in small intervals [47]. In order to better understand K-M analysis, it is important to take into account three basal concepts [48]:

- **Serial time:** The time each patient/individual remained in the study.
- **Status at the end of serial time:** Either the event being studied was achieved or not (i.e., censored).
- **Study group:** The study group/subgroup each patient belongs to.

Since this information is collected for each individual, by the end of the study there should be enough data to build a survival table. Survival tables constitute not only a methodology to store information but also a bridge to the computation of survival probabilities and the survival function [49]. Consecutively, these are used to build the survival curves, or K-M curves. This pipeline can be observed in Figure 2.2.

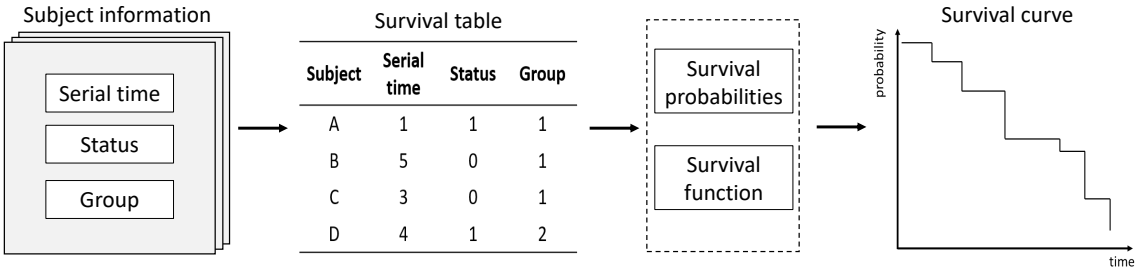


FIGURE 2.2: From subjects to survival curves.

In figure 2.3 we present an example of a survival curve. In order to interpret it, one needs to observe both the horizontal and the vertical lines. The former represents the survival probability for the interval being contemplated, while the latter (although sometimes not represented) makes it possible to scrutinize the change in the cumulative probabilities as we move further in time. In some representations a point or mark can also be observed along the horizontal lines, which represents the censored subjects [50].

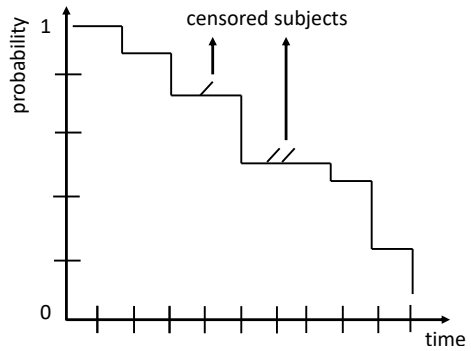


FIGURE 2.3: A representation of a survival curve.

When it comes to the interpretation of survival curves, besides the visual aspects, some notions are required [51], namely:

- **T:** Denotes the random variable for a person's survival time.
- **t:** Specifies a value of interest within T.

- **d(0,1):** A random variable which indicates whether a subject met the event (1) or was censored (0).
- **S(t):** Corresponds to the survival function and expresses the probability of T exceeding a specific t.
- **h(t):** Hazard function. Unlike the survival function, this one focuses on failure, not survival. It is presented as a rate, not a probability, so its values range from 0 to infinity. It amounts to the instantaneous potential that the event is achieved within a narrow time frame. This function can be derived from the survival function and vice versa.

2.6 Mixture Density Networks

When it comes to ML problems, where the goal is to model the conditional distribution of a random variable, we typically observe a Gaussian distribution being assumed. Although this works for simpler problems, real-world problems do not often follow this type of distribution, and so assuming a Gaussian distribution may lead to poor results [52]. Presented in 1994 by Bishop, MDN proved to be the solution for these types of problems [53]. As seen in Figure 2.4, MDN have a simple approach which consists in combining a DNN with a Mixture model. This joint task allows MDN to theoretically represent any conditional probability distribution, thus making it a powerful tool in problems not represented by a simple distribution, such as Gaussian.

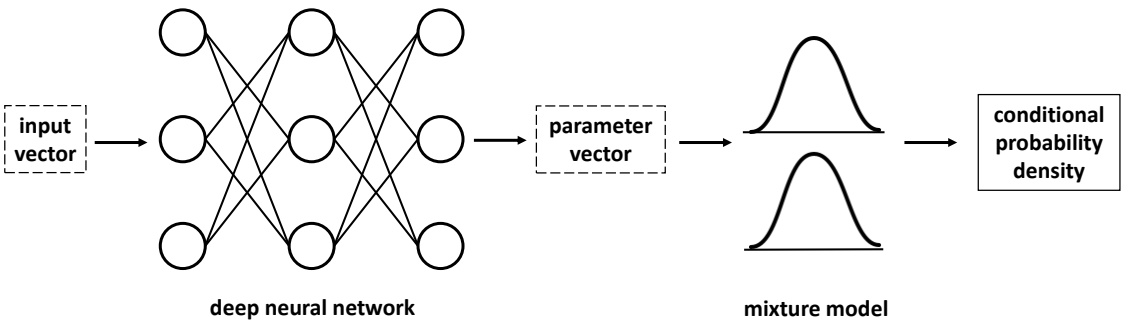


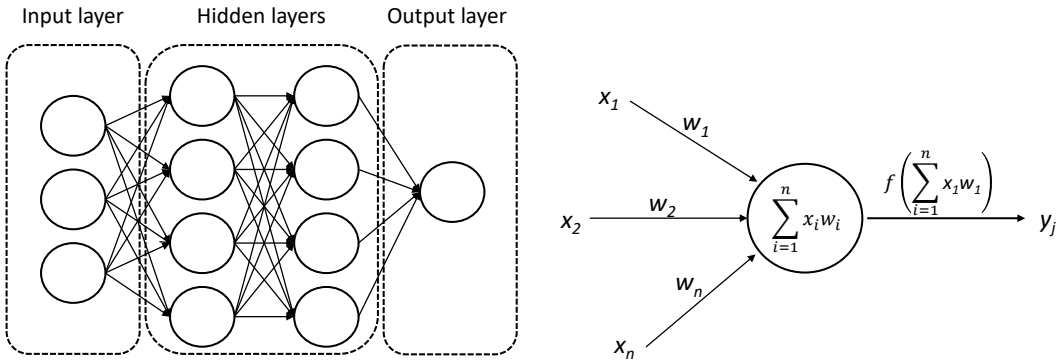
FIGURE 2.4: Architecture of a Mixture Density Network (adapted from [53]).

In the following subsections, we provide the theoretical foundations of each of the components of the Mixture Density Networks, namely Deep Neural Networks and Gaussian Mixture Models. Finally, we demonstrate how these components integrate in a single model.

2.6.1 Artificial Neural Networks

Artificial Neural Networks (ANN) are computational models made to mirror the functioning of the human brain [54]. Unlike traditional ML models, ANN make use of a layered type of learning in order to retrieve more information from the given data. The basic unit of this model is the neuron, and a set of neurons composes a layer. There are three types of layers: the input layer, which receives the data, the hidden layer, responsible for the learning, and the output layer, which outputs the result of the learning process.

The layers in a network are linked through connections between neurons, as they receive data (i.e., input) and send data (i.e., output). One neuron can have multiple inputs (i.e., ingoing connections) and outputs (i.e., outgoing connections). In order to distinguish the importance of the variables in the data, a weight is associated to each neuron input [55]. The bigger the weight, the more a variable will contribute to the output. Inside a neuron, the weighted sum of all its inputs is calculated. However, the result of this sum is only output if a threshold inherent to the neuron is activated. If the weighted sum of the inputs of a neuron is bigger than the threshold, then the neuron is activated and information can pass to the next layer. In order to create an output, the weighted sum is passed through an activation function, whose main goal is to introduce non-linearity to the network. The activation function can apply different types of operations, depending on the needed output. A general depiction of an ANN is presented in Figure 2.5.



(A) Artificial Neural Network (B) Neuron

FIGURE 2.5: Representation of an Artificial Neural Network and its basic unit.

In order to learn, the neural network continuously adjusts the weight values until an optimum set of values is found. Since it is very difficult to achieve a perfect solution (i.e., the perfect weight values), the learning process of a network consists in an optimization process. In a supervised learning context, this process happens as follows:

1. The instances of the training data are fed to the network one by one.
2. In each instance, the network observes how much the output (i.e., the predicted value) differs from the expected value (i.e., the true value). This difference is given by the Loss function.
3. In order to move in the direction of a local optimum solution, the Loss function is optimized (i.e., we try to minimize the loss value).

The optimization of the Loss function is usually done using the Gradient Descent algorithm [56] paired with the Backpropagation algorithm [57]. The first step in the optimization process consists in computing the gradient of the neural network, using backpropagation. Mathematically, the gradient is a partial derivative of the loss function with respect to the weights. In practice, the gradient shows how much the weights need to change to minimize the loss function. Then, the Gradient Descent algorithm will use this gradient as the goal of the optimization. This means that the values of the weights will be updated in order to follow the direction of the gradient.

2.6.2 Gaussian Mixture Models

Gaussian Mixture Models are a parametric model used to model complex probability distributions using a set of simpler distributions (i.e., Gaussian distributions) [58]. These

are defined as the weighted sum of M Gaussian distributions, called the components (Equation 2.3).

$$p(x) = \sum_{i=1}^{M} w_i \, g\left(x|\mu_i, \sum_i\right) \tag{2.3}$$

Where x is a continuous data vector, w_i with $i = 1, ..., M$ are the mixture coefficients and $g\left(x|\mu_i, \sum_i\right)$ with $i = 1, ..., M$ are the component Gaussian distributions. Even though the form of the distributions is known (i.e., Gaussian in this case), the parameters for the distributions and the respective weights are unknown. Usually, in order to calculate them, Mixture Models employ the Expectation Maximization algorithm. Since we are in the context of Mixture Density Networks, the parameters are obtained using an ANN.

2.6.3 Mixture Density Networks

MDN are generative models that rely on the straightforward concept of combining an ANN with a Gaussian Mixture Model to predict a conditional distribution $p(t|x)$. Each distribution is constituted by K components, which are Gaussian in the case of continuous target variables, as it is the case of this work. In Equation 2.4 we can observe how the conditional distribution $p(t|x)$ is obtained.

$$p\left(t|x\right) = \sum_{k=1}^{K} \alpha_k\left(x\right) N\left(t|\mu_k\left(x\right), \sigma_k^2\left(x\right)\right) \tag{2.4}$$

In each component, α corresponds to the mixing coefficient, μ the mean and σ the variance. For all of the K components, these parameters are dictated on the output layer of the neural network, which is composed of three different nodes, each predicting one parameter. Each of these nodes needs to respect a set of conditions:

- The mixing coefficients (α) must satisfy $0 \leq \alpha_k(x) \leq 1$ and $\sum_{k=1}^{K} \alpha_k(x) = 1$. Therefore, the node used to predict this parameter uses a softmax activation function.
- The means μ are directly computed using linear activations.
- The variances σ need to satisfy the condition $\sigma_k^2(x) \geq 0$, and thus this node uses exponentials of activations.

The weights of the neural network are obtained by minimizing the error function, the negative logarithm of the likelihood, represented in Equation 2.5.

$$E(w) = -\sum_{n=1}^{N} ln\left\{\sum_{k=1}^{K} \alpha_k(x_n, w)\, N\left(t|\mu_k(x_n, w), \sigma_k^2(x_n, w)\right)\right\} \tag{2.5}$$

This simple loss function allows the problem to be framed as a minimization, and therefore meaning that the best solution is the one with a Negative Log-Likelihood value closest to zero, as can be seen in Figure 2.6 [59].

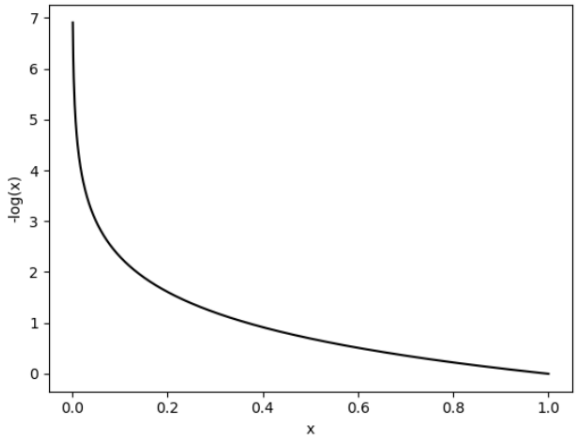


FIGURE 2.6: Interval of values obtained by the Negative Log-Likelihood loss function.

The performance of an MDN model can be evaluated in a distribution prediction task or a pointwise prediction task. If we want to compare it to approaches that predict a distribution, the Negative Log-Likelihood is used. However, when the goal is to compare it to other poitwise prediction approaches, the root mean squared error (RMSE) is used, defined in Equation 2.6.

$$RMSE = \sqrt{\frac{\sum_{t=1}^{T}(x_{1,t} - x_{2,t})^2}{T}} \tag{2.6}$$

Where $x_{1,t}$ and $x_{2,t}$ represent the ground-truth and the predicted ages of onset, respectively, for a given patient t.

In the next chapter, we discuss the most relevant methodologies used in the context of some of the concepts here presented, namely Survival Analysis, Subgroup Discovery, Sensitivity Analysis and the prediction of Age of Onset.

Chapter 3

State of the Art

In this chapter, we analyze the state of the art in the context of this dissertation. We begin by exploring, in Section 3.1, the current contributions in Subgroup Discovery. Then, in Section 3.2, we go over the state-of-the-art methodologies employed for survival analysis. Finally, in Section 3.3, the most relevant works related to the prediction of Age of Onset (AOO) are discussed.

3.1 Subgroup discovery

In the field of Medical Sciences, traditional subgroup analysis was mostly used in clinical trials as a way to identify and compare treatment groups [60]. However, since a great deal of information related to the subgroups was chosen by the investigators conducting the studies (sometimes even the subgroups themselves), the traditional methods were considered subjective [61]. This and other problems led traditional methods to be deemed as problematic or even not completely trustworthy. [60–63].

As a result of the aforementioned issues, investigation drew apart from these strategies and focused on correcting the subjectivity problem, which led to the emergence of several data-driven approaches. One of the first data-driven approaches, recursive partitioning, was proposed in 1963 by Morgan and Sonquist and consisted in recursively bisecting the predictor space, which proved to be a powerful tool in establishing relationships between the outcomes and the predictors [64]. Even though the use of tree-based methods exploded, especially with the emergence of Classification and Regression Trees (CART) in 1984 [65], they were not used to perform subgroup discovery until 1995. Ciampi et al. [66] demonstrated the potential of tree-based methods in subgroup discovery with Recursive

Partition and Amalgamation (RECPAM), an algorithm which showcased a subgroup in every terminal node of the generated tree. This work paved the way to the acceptance of tree-based models among clinicians, being still used nowadays [67–69].

Although tree-based models are still seen as a viable option, mostly due to their interpretability, it was demonstrated that recursive partitioning can be a double-edged sword, as the action of recursively splitting each node can lead to overgrown trees that lose their interpretability [70]. In order to solve this, several approaches were theorized, all having in common the attachment of parametric models to terminal nodes. Among others, the works of Quinlan [71], Gama [72] and Loh [70, 73, 74] should be noted. In an attempt to unify these strategies and provide a foundation for future studies in this area, the work performed by Zeileis et al. introduced Model-based Recursive Partitioning (MOB) [75], an approach in which every leaf of a tree is associated with a fitted parametric model. The parameters (θ) of each model are estimated by fitting the model *M(Y,θ)*, with Y being the dataset, to the full set of observations in a tree node. The chosen set of parameters is the one which minimizes the objective function ψ (usually the negative log-likelihood). Because of this, subgroups are based on parameters, in opposition to regular tree-based models, where the subgroups are the outcome values of the tree leaves.

MOB was considered an advance because of three main criteria: (a) the complexity of the tree can be controlled, (b) model parameters can be restricted, allowing for an easier interpretation of the relationship between outcomes and predictors, and (c) it is versatile when it comes to different statistical models. In 2016, Seibold et al. [76] proved the value of model-based approaches as a strategy for discovering subgroups, being followed and built-on by several studies, many still being conducted recently. In a work directed by Thomas et al. [77], this methodology is used to access the correct subgroups in dose-finding clinical trials, a core theme in precision medicine. Another study, presented by Tiendrébéogo et al. [78], showed the efficacy of MOB in distinguishing mortality risk profiles for patients infected with Human Immunodeficiency Virus (HIV).

Notwithstanding the global use of model-based approaches in research, some authors state that these models do not account for confounding, a fundamental assumption when dealing with causal relationships [79]. Confounding is a key aspect to have into account when comparing subgroups. This phenomenon is illustrated in Figure 3.1. As it is visible in Subfigure A, confounding happens when there is an inaccuracy in the causal effect, meaning that the actual exposure of interest is mistaken for some other factor which is

related to the outcome, often labeled the disease [80]. If the other factor is only related to the actual exposure (Subfigure B) or to the disease (Subfigure C), then we are not in the presence of confounding. In 2019, van Wie et al. [79] demonstrated the effects of confounding in MOB, subsequently combining it with a confounder detection approach proposed by Wiedermann and Li in 2018 [81].

Algorithm-wise, MOB can be compared to other methodologies, all of them having recursive partitioning or simply trees as a foundation. These are Interaction Trees (IT) [68], Simultaneous Threshold Interaction Modeling Algorithm (STIMA) [82], Subgroup Identification Based on Differential Effect Search (SIDES) [83] and Adaptive Refinement by Directed Peeling (ARDP) [84, 85]. In 2019, Huber et al. presented a thorough comparative study of these four methodologies, alongside with MOB, on Amyotrophic Lateral Sclerosis (ALS) data, additionally proposing a criterion to better select subgroups [86].

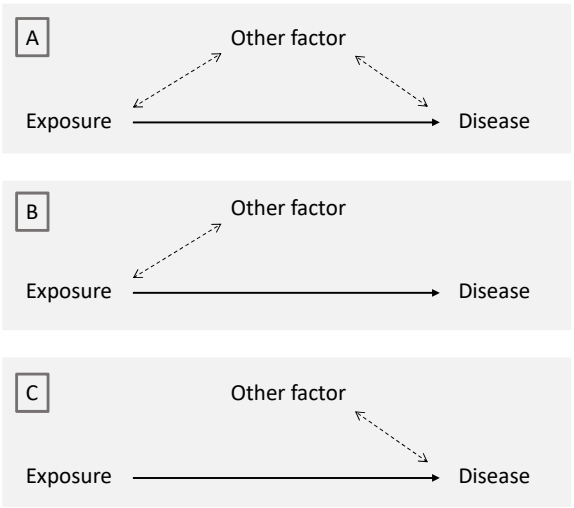


FIGURE 3.1: Example of a confounding (A) and not confounding (B and C) scenarios (adapted from [80]).

Despite the fact that tree and tree-derived methodologies constitute a great portion of the methods currently used to perform subgroup discovery, it is still worth mentioning other techniques that were observed. Among these we can find clustering techniques, with a common approach being the grouping of patients based on their symptoms, as we can observe in the works performed by Allapattu et al. [87] and by Almeida et al. [88]. In 2017, Nezhad et al. defended that previous methodologies had poor generalization behavior, having opted for designing a supervised approach on clustering, Supervised Bi-Clustering (SUBIC) [89]. Even though clustering methods have proven their worth,

these are less used when compared to tree-based methods, as some authors deem them less effective when dealing with Electronic Health Record (EHR) data [90].

3.2 Survival Analysis

Due to censored instances, a limited feature space and the potentially extensive computational time, using traditional Machine Learning (ML) methods to predict time-to-event data has been shown to be infeasible [91]. For this reason, problems with this type of data need to be tackled with survival analysis. As it is visible in Figure 3.2, survival analysis methodologies comprise two types of methodologies, statistical methods and machine learning methods.

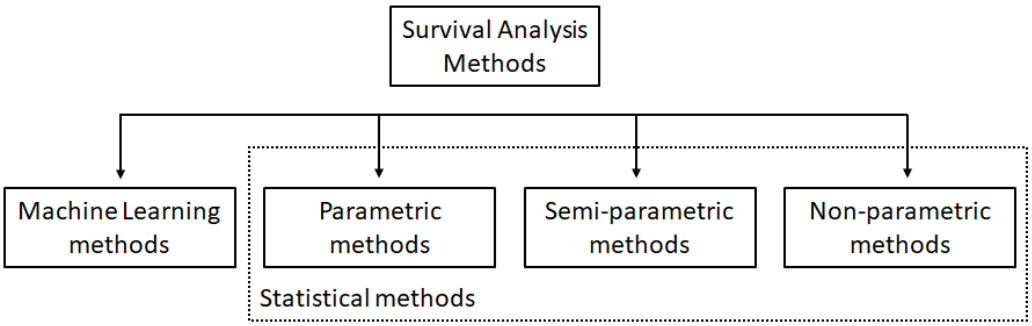


FIGURE 3.2: Types of Survival Analysis methods.

The latter are not the traditional ML methods, but, instead, adapted to be able to deal with the shortcomings caused by survival data. Due to this adaptation, these methodologies have risen in popularity [92]. Some of the most widely used ML methods used for survival analysis nowadays are Bayesian methods, Neural Networks, Survival Forests, Support Vector Machines and Ensemble methods. These will be introduced in the following subsections.

Despite the rising interest of ML, statistical methods are still widely used, with Kaplan-Meier and the Cox Regression methodologies being applied to a large set of areas. Just in 2018, Choi et al. used the Cox regression, in parallel with other methods, with the aim of predicting suicide cases in South Korea [93]. In 2020, the Kaplan-Meier estimate was used by Calabuig et al. to understand the behaviour of COVID-19 in different countries [94].

3.2.1 Bayesian Methods

Two well-known Bayesian methods are Bayesian Networks (BN) and Naïve Bayes (NB), having the Bayes theorem as their foundation. Introduced in 1997, Bayesian Networks are a graphical representation of the relationships between an array of variables, depicting a conditional dependence between variables as edges in an acyclic graph [95]. The potential of this model for survival analysis was immediately seen, with Gustafson et al. highlighting BN's interpretability [96]. It was also in this study that some shortcomings of this method were pointed, namely its stronger ability to obtain qualitative results than quantitative. Although BN are still used, mostly for risk prediction, their main advantage over statistical methodologies is its explainability and graphical output [97].

The NB model is often referred to as a restricted form of the BN, given the fact that the variables have no edges between them (i.e., they are considered to be independent). Despite being a simpler model, it can frequently outperform BN when it comes to quantitative results [95].

3.2.2 Neural Networks

Artificial Neural Networks, an already popular ML method, is being increasingly used in Survival Analysis, mostly due to their ability to provide non-linear modeling of censored data, as pointed out by Biganzoli et al. [98]. According to a survey conducted by Wang et al. [92], there are currently three main methods being used to apply neural networks to survival analysis problems.

1. Applying the neural network to perform survival analysis directly from the inputs provided, providing a pointwise prediction of the time until the event studied.

2. Adapting the neural network to work as an extension of the Cox Proportional Hazards model [99].

3. Using the survival status of the subjects as the output of the neural networks [100].

Specifically in the medical domain, we observe the presence of neural network for survival analysis through the works of Katzman et al. (DeepSurv) [99] and Yao et al. (DeepCorrSurv) [101]. While the former is specifically used to assist in the personalized treatment of patients by estimating the risk of diseases, the latter focus on discovering important markers from different types of patient data.

3.2.3 Random Survival Forests

Random Survival Forests are a non-parametric model widely used in survival analysis [102]. These consist in Random Forests that were adapted to handle right-censored survival data. Overall, this model is built according to the principles of Random Forests:

- Using bootstrapped data, survival trees are grown;
- In order to split tree nodes, random feature selection is used;
- The final prediction of the Survival Forest is calculated by averaging the survival tree predictors;

Its popularity has risen mostly due to its lack of assumptions or prior knowledge of variable interactions, its robustness to outliers and its integrated out-of-bag error, avoiding overfitting. Besides, it has the particularity of dealing well with highly dimensional data, a common problem in EHR data.

3.2.4 Support Vector Machines

Traditional Support Vector Machines are not used in survival analysis due to their poor ability to deal with incomplete data (i.e., censoring). Because of that, adaptations of this method were introduced to face this problem [103]. One of the first proposed methodologies to account for censored data was to apply Support Vector Regressors (SVR) to the survival data, ignoring the censored instances [104]. Building on this approach, Khan and Zubek [105] introduced SVRc - Support Vector Regression for Censored Data, making use of an updated asymmetric loss function to be able take into account both the censored and uncensored instances. Later, Widodo and Yang [106] presented Relevance Vector Machines, which make use of Bayesian inference in order to obtain parsimonious estimates.

3.3 Prediction of age of onset

As early as 1941 we can observe a study trying to find patterns of AOO in degenerative diseases. Julia Belle studied the possible patterns of AOO in hereditary muscular dystrophies by analyzing their correlation coefficients [107]. Several years later, now in the context of other diseases, the foundation for the estimation of AOO still relied on the same basal concept: linear methods. In 1975, Brackenridge et al. used Linear Regression

as a tool to estimate the AOO of Huntington's disease, taking into account the affected parents [108]. Still in relation to Huntington's disease, two formulae were developed to more accurately predict AOO (i.e., formula of Langbehn [109] and formula of Ranen [110]). Despite the fact of being based on linear relationships, these methods only apply to one disease, as they are created according to inner genetic characteristics of Huntington's disease

Although linear methodologies are still in use today, they appear most of the times accompanied by survival analysis methods (e.g., survival curves, the Kaplan-Meier model or Cox regression methods) [111]. In current times, survival analysis methodologies for AOO prediction can be considered a staple of disease management, with studies being conducted using this approach as a foundation for different types of diseases. In 2016, Allport et al. used Cox proportional regression analysis to predict offsprings' AOO of cardiovascular disease, having parental AOO as a predictor [112]. With this approach, a significant relationship between parental and offspring AOO was observed, in the context of cardiovascular disease. In a neurodegenerative disease setting, survival analysis was used to predict AOO in Alzheimer's disease [113], Huntington's disease [114] and Multiple Sclerosis [115], just to name a few.

In regards to TTR-FAP, one work, conducted by Cisneros-Barroso et al., used Pearson's correlation coefficient to estimate the offsprings' AOO based on parental AOO [116]. The majority of studies, however, approached this problem with ML techniques [117–119]. Even though ML methods are reliable and are able to achieve great results, the problem with some of those models is their inability to explain the results presented [120]. This makes them a less viable option in the medical assistance context, since the reasoning behind decisions is not entirely explainable.

To solve this problem, one may resort to the usage of explainable techniques like SD or sensitivity analysis, as a means to support medical decisions. Given this remark, two works are of extreme importance to refer. The first, by Li et al., is able to use deep learning together with patient stratification to identify patient subgroups [90]. The second, and the closest to the one here presented, by Katzman et al., introduces DeepSurv, a Cox proportional hazards deep neural network which is able to predict survival curves of patients [99].

To the best of our knowledge, there is no work conducted so far with the aim of providing a methodology to predict a patients' AOO survival curve that is both capable of

dealing with unseen data and of providing explainable predictions with patient grouping. For this exact reason, the work here presented will focus on the creation of an explainable methodology of AOO prediction. That way, not only can it be used for patients with TTR-FAP but also for the epidemiological purpose of studying the disease progression within offsprings.

Chapter 4

Modeling the Survival Curves

In this chapter, we describe our approach for modeling the Age of Onset (AOO) survival curves of patients with Transthyretin-Related Familial Amyloid Polyneuropathy (TTR-FAP). First, in Section 4.1, we provide an overview of the architecture of the Mixture Density Network (MDN). In Section 4.2, we expose the problems encountered during the training of the model and the strategies used to mitigate them. Then, in Section 4.3, we showcase the survival curves modeled by the MDN. In Section 4.4 we explain the performance evaluation methodology used and finally, in Section 4.5, we present the results obtained.

4.1 Architecture

MDN make use of an Artificial Neural Network (ANN) to estimate the parameters of each component of the mixture model. For the MDN model used, patient data is fed to a Deep Neural Networks (DNN) with two fully-connected layers with a rectified linear activation function (i.e., ReLU). These were chosen due to their capacity to learn complex relationships and avert the vanishing gradient problem. The results of the last layer are then used to calculate the parameter vectors comprising the mixture coefficients (α), means (μ) and variances (σ) for each component of the mixture model. Since each parameter needs to obey to a set of restrictions, three different layers are used to calculate them:

- The mixture coefficients (α) are obtained using a fully connected layer with a Softmax activation function, in order to normalize the given values.
- The means (μ) are obtained with a fully-connected layer with a linear activation.

- The variances (σ) are calculated in a fully-connected layer with a non-negative exponential linear unit activation function (i.e., Nnelu), as the values need to be positive.

Finally, the outputs of these three layers are joined using a concatenation layer, which in its turn will output the set of parameter vectors used to model the distributions, one per component. The architecture of the MDN can be observed in Figure 4.1.

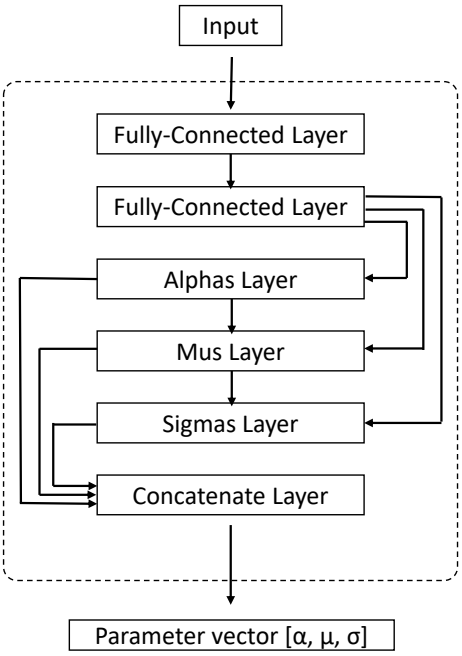


FIGURE 4.1: Architecture of the Mixture Density Network.

For the described MDN, we use three Gaussian components to model a distribution, and the Negative Log-Likelihood as the loss function.

4.2 The NaN problem

While training the above-defined MDN, the loss function often returned NaN values (i.e., Not a Number). According to the literature, this constitutes a common problem with these models and can happen because we have either too large (i.e., arithmetic overflow) or too small (i.e., arithmetic underflow) values to be properly expressed by the corresponding numerical data types, and are thus represented as NaN [121]. This may occur under three specific scenarios:

- The result of an exponential expression is too large for its data type.
- The result of a logarithm expression is extremely close to zero.

- A fraction denominator is extremely close to zero.

In the original proposal of the MDN, Bishop [53] explains that a probability density function $p(x)$ can be defined by a mixture of m probability density functions indexed by j, $p_j(x)$, with the mixture coefficients α_j being the weights. Thus, $p(x)$ can be defined as shown in Equation 4.1.

$$p(x) = \sum_{j=0}^{m-1} \alpha_j p_j(x|\theta_j) \tag{4.1}$$

With θ_j being the set of parameters that describe a distribution. In his work, Bishop [53] also shows that any probability density function can be approximated by the expression defined in Equation 4.2.

$$p(x|\Pi, \Theta) = \sum_{j=0}^{m-1} \alpha_j \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-\left(\frac{(x-\mu_j)^2}{2\sigma_j^2}\right)} \tag{4.2}$$

If we observe Equations 4.1 and 4.2, it is clear that a mixture of densities is the sum of the probability density functions $p_j(x)$ weighted by the mixture coefficients α_j. When applying the loss function, i.e., the Negative Log-Likelihood, we apply a logarithm to an exponential multiplied by a mixing coefficient α_j, which can lead to extremely small values, leading to a numerical underflow.

Upon inspection of the loss function outputs during training, we observed a gradual decrease in value, leading to the eventual case of numerical underflow. In order to contradict this problem, we applied the Adam optimizer [122] with Gradient Clipping, a gradient descent method which, contrary to Stochastic gradient descent, allows for the adaptive estimation of first-order and second-order moments. Gradient Clipping is a technique that is able to control the numerical underflow problem, by simply clipping the derivatives of the loss function according to a given threshold. This means that the values will be clipped if a gradient value is either less than the negative threshold or greater than the positive threshold [123]. In the case of this work, by clipping the gradient value according to a defined minimum, the NaN problem was effectively mitigated.

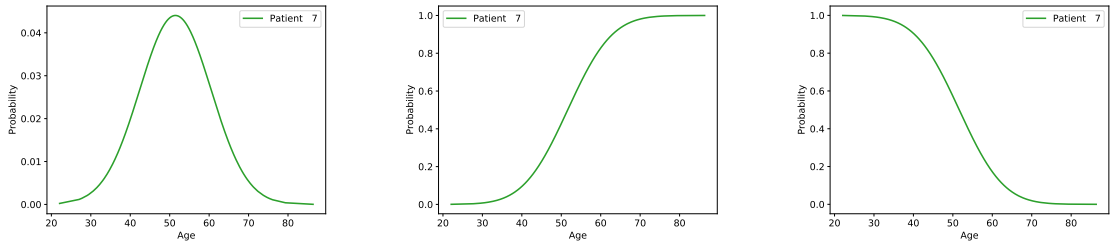
4.3 Survival Curves

In order to obtain the survival curve of a patient using the MDN, the parameters output by the neural network are passed on to the mixture model. The mixture model, in turn, outputs the probability density function for the patient in question, which we then use to calculate the cumulative density function. This function is obtained by integrating the probability density function, and represents the area under the probability density function up to a value x, which translates to the probability of a random variable X being less or equal to x, or $P(X \leq x)$.

Upon having the cumulative density function, we then obtain the patient's survival curve according to the expression in Equation 4.3. The survival curve is also known as the complementary cumulative distribution function, as it translates to the probability of a random variable X being greater than x, or $P(X > x)$.

$$P(X > x) = 1 - P(X < x) \tag{4.3}$$

The three curves mentioned (i.e., probability density function, cumulative density function and survival curve) are illustrated for one patient, in Figure 4.2.



(A) Probability Density Function (B) Cumulative Density Function (C) Survival Curve

FIGURE 4.2: Curves obtained for a patient.

The obtained survival curves of AOO for a group of ten randomly-chosen patients is depicted in Figure 4.3. With the modeled survival curves, and from a medical point of view, it is possible to obtain information in an unequivocal way. Especially when compared to a standard Machine Learning (ML) point prediction, where only a value of AOO is predicted, the survival curve provides much more information. A straightforward observation one can take from a survival curve is the age interval at which a patient can

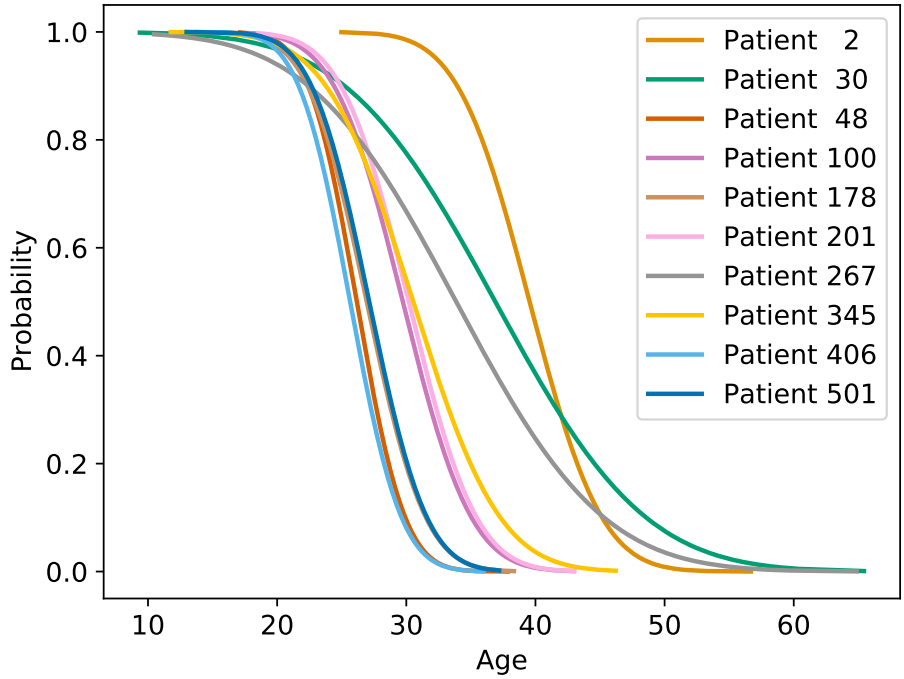


FIGURE 4.3: Survival curves of a selected group of patients.

develop symptoms for a given disease. Then, and the most important piece of information expressed by a survival curve, is the probability that a patient has of developing symptoms at each given age point in the survival curve's age range, or rather, the probability of survival. Note that, unlike most survival problems, where the event at focus is death, the event under study in this work is that of developing symptoms related to a given disease (i.e., more commonly referred to as AOO).

Considering, for instance, Patient 2 from Figure 4.3, we can clearly observe their age range: between 25 and 57 years of age, this patient is likely to present the first symptoms of TTR-FAP. Despite this, we observe through the respective curve that it is unlikely that they will develop symptoms before 35 years of age, as the probability of survival is extremely close to 100%. We can also observe that the patient's curve is very steep, which tells us that between 35 and 45 years of age the patient will, most likely, develop symptoms, as after the age of 45 the probability of survival is almost zero.

4.4 Performance Evaluation

To show the performance of the MDN model used, we applied it to a set of four datasets, each containing patient data related to a given medical condition. The main dataset used in this work consists of Patient Clinical Information available in Electronic Health Record (EHR) format. It was provided by Unidade Corino de Andrade, a department of the Centro Hospitalar do Porto, which is a specialized unit focused on the study and treatment of TTR-FAP. The data contains information about 2793 patients at this hospital and comprises six features: Sex, Birth Year, Age of Onset, Age of Onset of the Parents, Number of Siblings, and the Average Age of Onset of the Patient's Family Tree. The patient's Age of Onset is used as the target variable.

To assess the applicability of the MDN model to other diseases, we also applied it to three public datasets retrieved from Kaggle [124]:

- *Cardio:* The *"Cardiovascular Disease dataset"* is constituted by data of 462 males in a heart-disease high-risk region of the Western Cape, South Africa [125], and the goal is to predict the AOO of Coronary Heart Disease.
- *ALS:* The *"End ALS dataset"* comprises data from 140 patients with Amyotrophic Lateral Sclerosis (ALS) and 30 healthy controls, for a total of 170 instances. Since this work focuses on AOO prediction, the healthy controls were removed. The goal is to predict the AOO of ALS [126].
- *Parkinson:* The *"Early Biomarkers of Parkinson's Disease dataset"* is comprised of 30 patients with early untreated Parkinson's disease, 50 patients with REM Sleep Behavior Disorder which are at high risk of developing Parkinson's disease, and 50 healthy controls [127]. Here, the healthy controls were also removed. The goal is to predict the AOO of Parkinson.

Contrarily to the TTR-FAP data, some of these three datasets required a pre-processing step, mostly to remove patients with a lot of missing values. The characteristics of all the datasets, before and after pre-processing, can be observed in Table 4.1.

Additionally, in order to have a benchmark of where the MDN stands when compared to other models, in terms of pointwise prediction, we compared its Root Mean Squared Error (RMSE) with a set of seven other typically used regression ML models. These were chosen based on the work conducted by Pedroto et al. [119] for prediction of AOO in

TABLE 4.1: Dataset information

	Before pre-processing		After pre-processing	
Dataset	**Number of attributes**	**Number of instances**	**Number of attributes**	**Number of instances**
TTR-FAP	6	2793	6	2793
Parkinson	65	130	33	80
Cardio	10	462	10	462
ALS	45	170	19	106

TTR-FAP. The chosen models considered were Decision Tree Regressors, Random Forest Regressors, Support Vector Regressors, Linear Regression, Lasso Regression, Ridge Regression and Elastic Nets. The used hyperparameters for each model are described in Table 4.2.

In order to avoid any bias or overfitting in the models and to achieve an equivalent comparison, a 10-fold cross validation with no shuffle was performed for all models. To validate this comparison, we performed the Friedman Rank test, as recommended by Demšar [128]. This test has the following hypotheses:

- H_0: The average results for each of the compared algorithms are equivalent.
- H_A: The average results for each of the compared algorithms are not equivalent.

In case the null hypothesis (H_0) is rejected, the Nemenyi post-hoc test is performed to assess which of the algorithms are different [128].

TABLE 4.2: Model hyperparameters

Model	**Hyperparameters**
Decision Tree Regressor	*{criterion: "mse", max_depth: None, min_samples_split: 2, min_samples_leaf: 1}*
Random Forest Regressor	*{n_estimators: 100, max_depth: None, min_samples_split: 2, min_samples_leaf: 1}*
Support Vector Regressor	*{kernel: "rbf", gamma: "scale", C: 1, epsilon: 0.1}*
Linear Regression	*{fit_intercept: True, normalize: False, positive: False}*
Lasso Regression	*{alpha: 1, fit_intercept: True, normalize: False, positive: False, selection: "cyclic"}*
Ridge Regression	*{alpha: 1, fit_intercept: True, normalize: False, solver: "auto"}*
Elastic Net	*{alpha: 1, l1_ratio: 0.5, fit_intercept: True, normalize: False, positive: False, selection: "cyclic"}*

We also compared the performance of the MDN, in terms of survival curve prediction, with that of two state-of-the-art survival analysis algorithms, namely Random Survival Forests [102] and DeepSurv [99]. This was accomplished by comparing the Harrel's C-index of each model, a performance metric which evaluates the discriminative power of a survival model [129]. The C-index, or concordance index, corresponds to the fraction of correctly ordered survival time of subjects (i.e., the probability of concordance between the predicted survival and the real clinical outcome). It constitutes a generalization of the area under the ROC curve (AUC), with the additional ability to account for censored data. Therefore, $c = 1$ means that the model has a perfect prediction. A simplified mathematical expression for this metric can be seen in Equation 4.4.

$$c = \frac{\text{number concordant pairs}}{\text{number of concordant pairs} + \text{number of discordant pairs}} \tag{4.4}$$

Finally, given the fact that this work is enclosed in a medical context, it is of maximal importance for users to be able to understand how reliable the model they are using really is. This need derives from the fact that, even though modern neural networks have an increased performance, they are often crudely calibrated (i.e., the predicted probability estimates do not completely represent the real correctness likelihood), as demonstrated by Guo et al. [130]. For this reason, we observe, through the means of a diagram, the agreement between the probabilities of the predicted events and the error of the model (Figure 4.4).

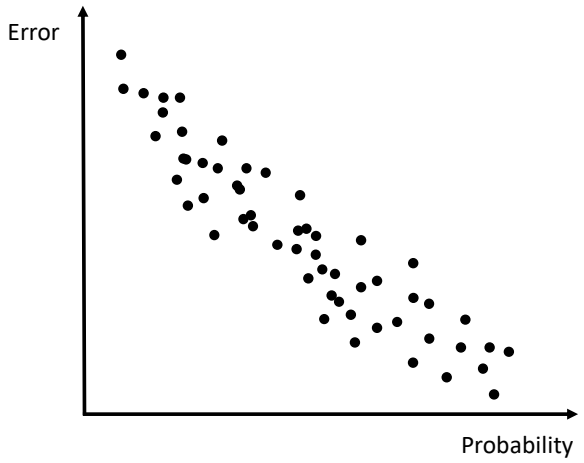


FIGURE 4.4: Calibration measurement approach.

Ideally, a model with 100% reliability, would allow us to observe a very strong inverse correlation between these two components (i.e., the predicted events and the error of the model), which we use as the measure of calibration.

The probabilities of the predicted events, or the confidence measure of the model, were calculated as shown in Equation 4.5, using the mixture parameters for each patient.

$$confidence = \sum_{i=1}^{n} \alpha_i \, cdf(x, \mu_i, \sigma_i) \tag{4.5}$$

With α_i (i.e., mixture coefficient), μ_i (i.e., mean) and σ_i (i.e., variance) being the parameters of a mixture component i, x being the predicted age values for a given patient, and cdf the cumulative density function.

4.5 Results

In this section, we present the results related to the performance comparisons carried out between the MDN and the previously mentioned models, as well as the outcomes of the calibration process.

4.5.1 Pointwise Prediction

The first set of results presented consists in the comparison of performance with the regression ML models. In Table 4.3, we compare the MDN with a Linear Regression (LR), Random Forest Regressor (RF), Support Vector Regressor (SVM), Elastic Net (EN), Lasso Regression (Lasso), Ridge Regression (Ridge) and Decision Tree Regressor (DT). For all the models, the mean RMSE and the standard deviation are presented. The best performing model for each dataset is represented in bold.

TABLE 4.3: Pointwise prediction evaluation of the tested Machine Learning regression models

	MDN	LR	RF	SVM	EN	Lasso	Ridge	DT
TTR-FAP	**6.3 ± 1.4**	7.4 ± 1.4	6.9 ± 1.2	6.8 ± 0.7	7.5 ± 1.4	7.4 ± 1.4	7.4 ± 1.3	8.8 ± 1.5
ALS	**7.3 ± 1.3**	12.2 ± 2.8	10.7 ± 2.1	10.6 ± 2.9	11.7 ± 2.1	11.7 ± 2.0	12.0 ± 2.6	15.3 ± 2.7
Cardio	**6.6 ± 0.88**	10.1 ± 1.0	9.3 ± 0.97	9.6 ± 1.3	10.2 ± 1.0	10.3 ± 1.1	10.1 ± 1.0	13.2 ± 1.2
Parkinson	7.8 ± 1.0	0.3 ± 0.1	2.6 ± 0.8	5.4 ± 2.5	0.3 ± 0.1	0.3 ± 0.1	**0.2 ± 0.1**	2.6 ± 1.0

We can see, in Table 4.3, the better performance of the MDN on all but the Parkinson dataset. However, it is important to assess whether this represents a statistical difference in performance. Using the Friedman Rank test, we are able to prove that this difference was, indeed, significant. Knowing this, we then applied the Nemenyi post-hoc test to be able to discriminate which models presented statistically significant differences. For all datasets, we observed a statistical difference between the MDN and the Linear Regression (LR), Elastic Net (EN), Lasso Regression (Lasso), Ridge Regression (Ridge) and Decision Tree Regressor (DT), which comprises the majority of the models tested against. Note that these statistical tests were not performed in the case of the Parkinson dataset, as the MDN performed considerably lower than the remainder of the models and, therefore, it was deemed unnecessary to follow through with this statistical verification process.

4.5.2 Survival curve prediction

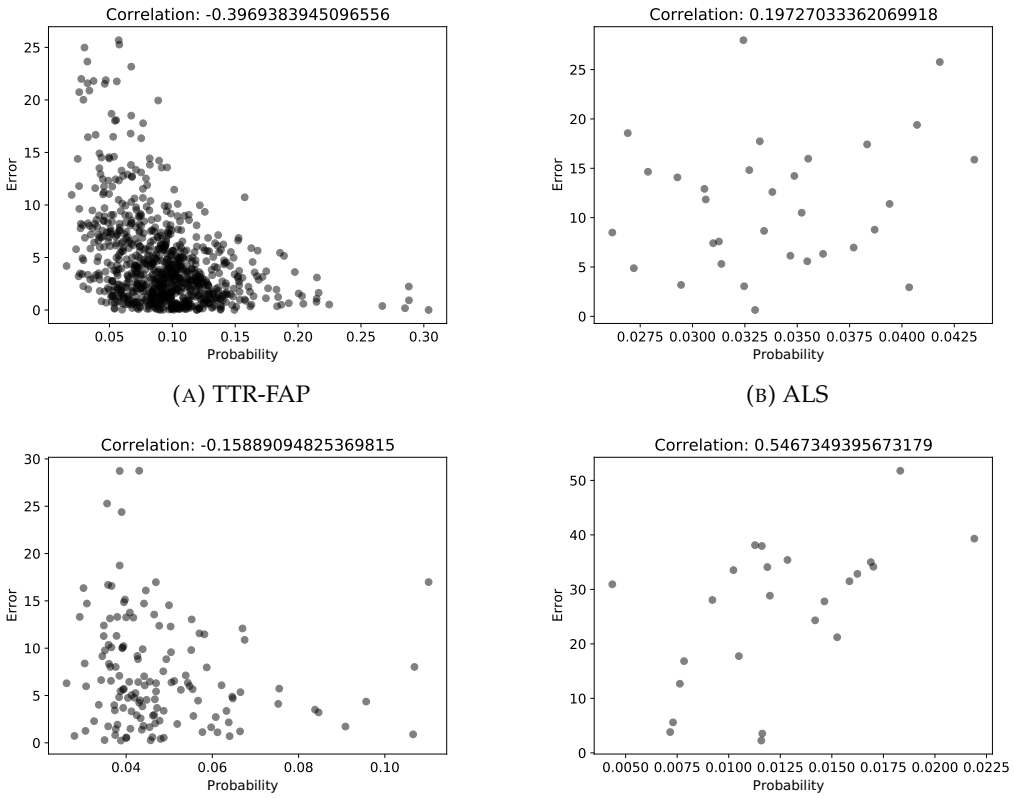
Regarding the results for the comparison between the MDN and the two selected state-of-the-art survival analysis models, the experience set up is as follows:

- **DeepSurv:** For this model we followed the indications in [131]. A 60%/20%/20% split into train, test and validation sets is created. We used 500 epochs and the following hyperparameters: $\{'n\_in' : 5, 'learning\_rate' : 1e-3, 'hidden\_layers\_sizes' : [5,5], 'batch\_norm' : True\}$.

- **Random Survival Forests:** RSFs are used with the scikit-survival package. The hyperparameters used are $\{n\_estimators : 1000, min\_samples\_leaf : 15, max\_features : "sqrt", n\_jobs : -1, min\_samples\_split : 10\}$

The results obtained from this comparison are presented in Table 4.4. Here we display the C-index of the MDN, the Random Survival Forests (RSF) and the DeepSurv models, for each of the tested datasets. As before, the best-performing model is represented in bold.

As we can observe in Table 4.4, the RSF model slightly outperformed both other models in the ALS dataset (wherein the MDN and the DeepSurv models performed equally well). In the case of the Parkinson dataset, DeepSurv was particularly better than RSF and, even more so, than the MDN. However, the MDN model performed fairly well in half of the test datasets used (i.e., TTR-FAP and Cardio), thus demonstrating it is able to



(A) TTR-FAP (B) ALS

(C) Cardio (D) Parkinson

FIGURE 4.5: Calibration measure for each dataset.

outperform two very used and well-established models in survival analysis. This showcases its prediction ability when compared to these state-of-the-art approaches.

Finally, as to measure how reliable the predictions of the MDN are from a probability density function point of view, we obtain a measure of calibration for all the problems under study. This can be observed in Figure 4.5. With this representation, it is possible to distinguish a clear decrease in the confidence measure as the size of the data also decreases, leading to a higher positive correlation between the two measures, far from the ideal strong negative correlation. This was somewhat expected, as a lower number of

TABLE 4.4: C-index of the tested survival analysis models

	MDN	RSF	DeepSurv
TTR-FAP	**0.80**	0.76	0.78
ALS	0.51	**0.53**	0.51
Cardio	**0.77**	0.76	0.74
Parkinson	0.55	0.76	**0.84**

instances presents an obstacle to the proper training of the model, which in turn causes a lower confidence level in its predictions.

Despite this, on the main problem of this work, the TTR-FAP dataset, we can observe an inverse correlation between the confidence measure and the error of the MDN. Thus, as the likelihood of the prediction gets smaller, as the error gets smaller,the prediction error increases, meaning that medical professionals can trust the predictions obtained by this model. Even though the correlation is smaller in the Cardio dataset, we are able to discern the appearance of an inverse tendency between the two measures. Because of this, we believe that by increasing the number of samples in these problems, a better value of calibration could be achieved.

Chapter 5

Beyond Predictions

In this chapter, we present the strategies employed to guarantee the robustness and explainability of the survival curves obtained by the Mixture Density Network (MDN). First, in Section 5.1, we provide a general overview of how these strategies are integrated with the MDN. Then, in Section 5.2, we present the in-detail methodology for Subgroup Discovery (SD) and the results obtained by this approach. Finally, in Section 5.3, we present the methodology for Sensitivity Analysis (SA) and how it can be applied both for the discovery of subgroups and for for the extraction of feature importance.

5.1 Overview

To achieve the goal of having an explainable and robust prediction of the survival curves, we take advantage of the explicative power of SD and SA. On a high-level, the Electronic Health Record (EHR) data is fed to an MDN, which will model the survival curve and probability density function of AOO (AOO) of a specific patient. Simultaneously, all the subgroups in the data will be found, and for each subgroup a probability density function and a survival curve are produced. The survival curves from the MDN and the subgroups will be compared using the Kolmogorov-Smirnov test, and a patient will be attributed to the subgroup that produces the lowest KS metric. Additionally, using SA, we are able to retrieve the global and subgroup-specific risk factors, as well as identifying the patient-specific thresholds within each attribute. The aforementioned pipeline can be observed in Figure 5.1.

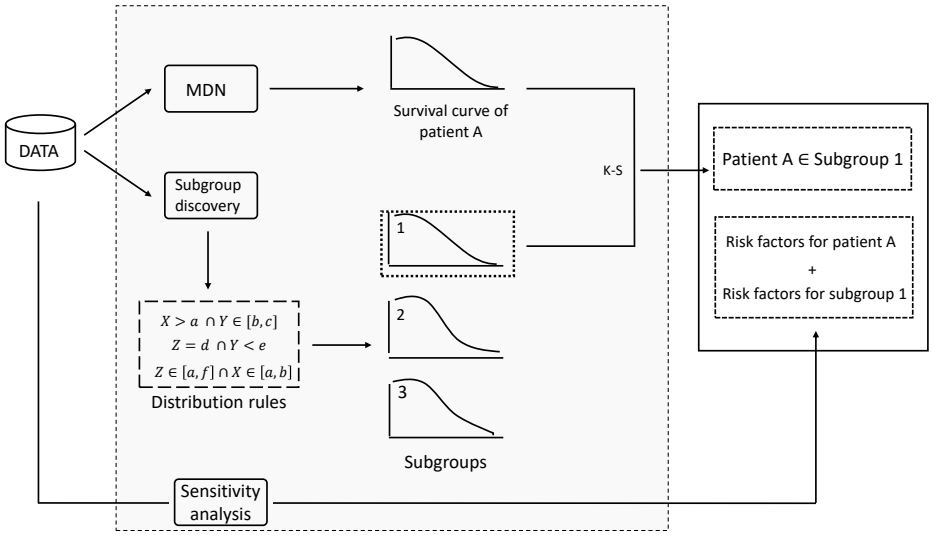


FIGURE 5.1: Exemplification of the proposed model.

The in-depth details of both strategies presented are described in the following sections.

5.2 Subgroup Discovery

Given a patient P and survival curve SC_p, the aim is to find a describable subgroup of patients that contains P and has an aggregated survival curve SC_s, similar to SC_p. For that, we explore the vicinity of the point representing the patient P and look for a maximal region with such an aggregated survival curve SC_s. This process is described bellow:

1. For each attribute we obtain, from their value domain, the set of values representative of the 20% quantiles, i.e., the minimum value in the domain (0% quantile), the maximum value in the domain (100% quantile), and all the in-between values for each step of 20%.

2. The quantile values are sorted in ascending order and grouped into pairs, forming an interval composed by a lower bound and an upper bound.

3. We exhaustively combine up to three intervals using the logical `AND` operator, so as to form distribution rules.

4. The distribution rules are applied to the patient data as a filter, which results in the subset of patients whose characteristics are in agreement with the conditions of the rule. Each subset constitutes a subgroup.

5. We obtain a survival curve for each of the generated subgroups, as well as for patient P. For the subgroups, we obtain the survival curve, SC_s, using Kernel Density Estimation. For patient P, the MDN model is responsible for generating its survival curve, SC_p.

6. The Kolmogorov-Smirnov statistical test is used to compare each SC_s to the SC_p, obtaining a KS metric.

7. The pair SC_s-SC_p which resulted in the lowest KS metric is obtained. A schematic representation of this process is depicted in Figure 5.2.

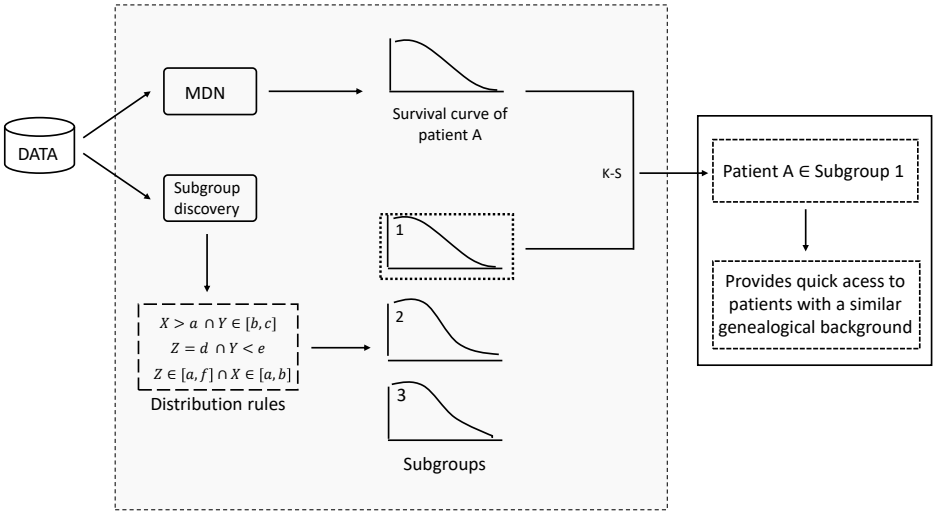
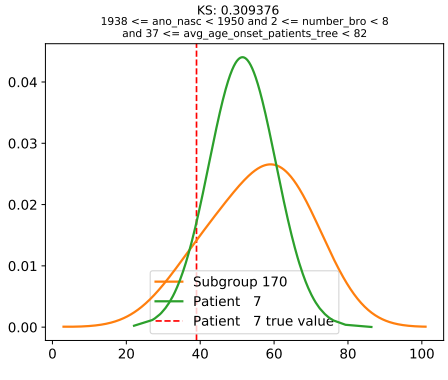


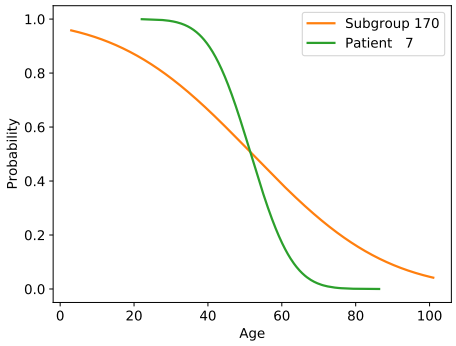
FIGURE 5.2: Process of obtaining the subgroup of a patient.

5.2.1 Subgroup Visualization

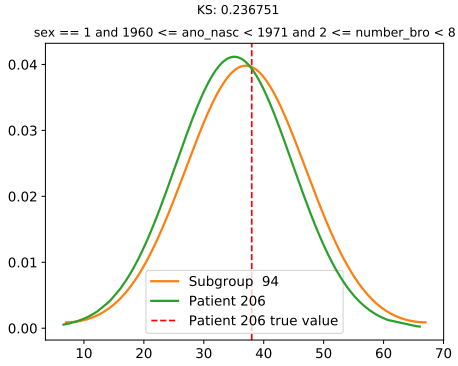
When the most related subgroup for a patient is found, it is important to compare the distribution of AOO of the pair. In Figure 5.3, we present the resulting diagrams when we model the probability density function and survival curve of AOO of a patient. For representation purposes, three patients were randomly selected, namely patients 7, 206 and 709. On the left side of Figure 5.3, we present the probability density function of the AOO for the selected patients and for their corresponding subgroups, and on the right side, we present their survival curves. These diagrams are created with the medical professionals in mind, as they should obtain the information in a straightforward way.



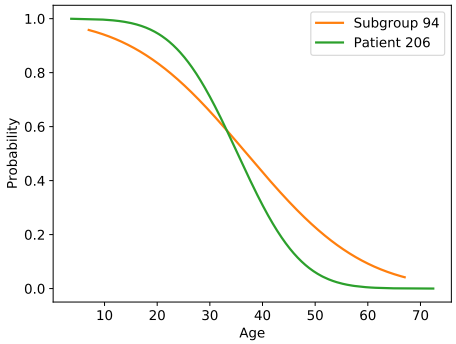
(A) Probability density functions for patient 7.



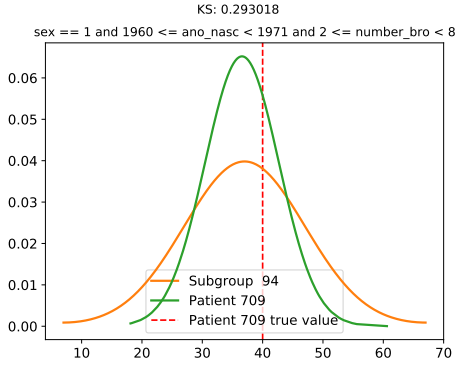
(B) Survival curve for patient 7.



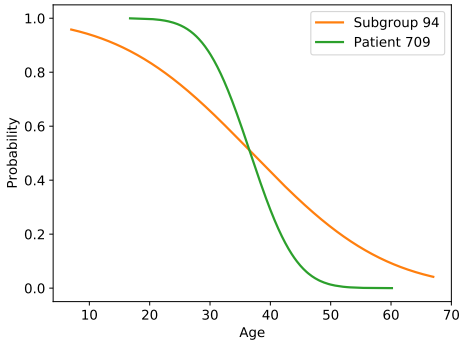
(C) Probability density functions for patient 206.



(D) Survival curve for patient 206.



(E) Probability density functions for patient 709.



(F) Survival curve for patient 709.

FIGURE 5.3: Probability density functions and survival curves of Age of Onset for a selected group of patients.

The probability density functions allow for the identification of the age range in which a patient could expect to experience symptoms for the first time. Here we present a probability density function of the AOO of a patient (i.e., the green curve), the probability density function for the respective subgroup (i.e., the orange curve) and the true AOO of the patient (i.e., the dashed red line). The true AOO would not, of course, be presented in a medical context, as the model would be facing never-before-seen patients. Finally, at the

top, we can observe both the rule that defines a subgroup and the KS metric between the two curves. The rules allow for the quick identification of the characteristics that lead to a specific curve, whereas the KS metric presents a measure of how much a patient fits into a subgroup. The survival curve allows to directly connect an age point with a probability of having the disease at that point.

5.2.2 Results

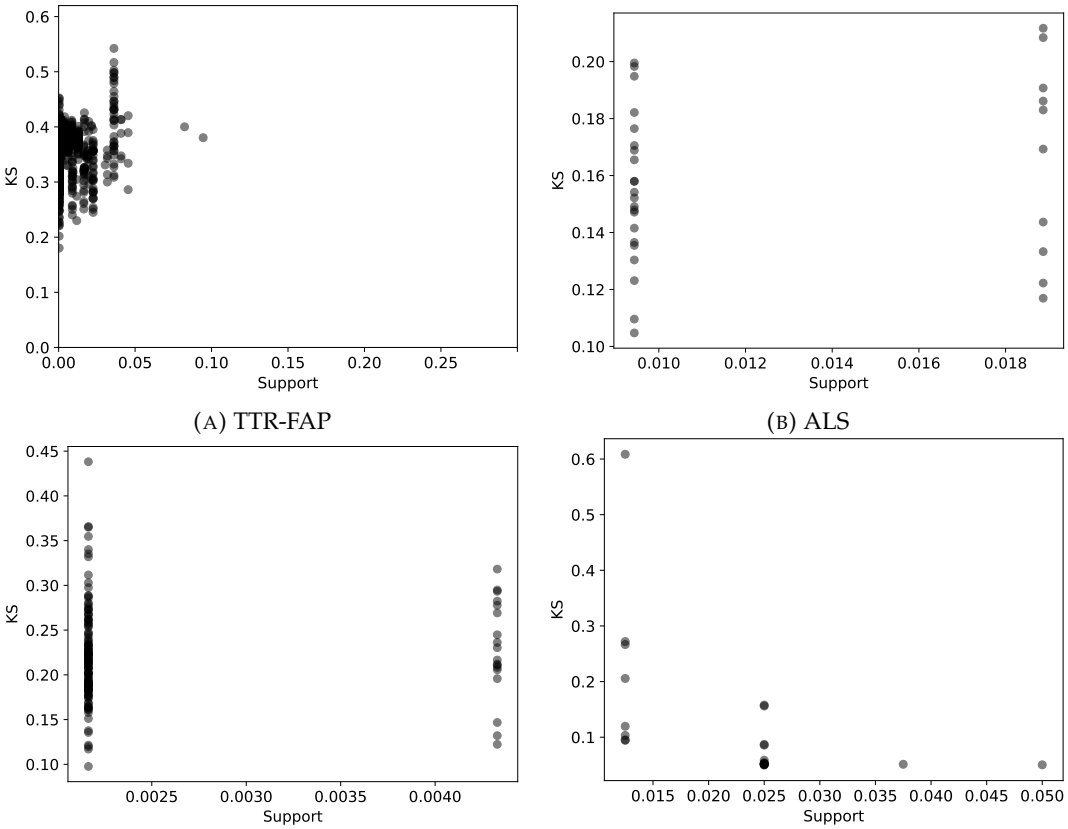
In order to assess how adequate the subgroups are to a patient, it is important to look into two aspects:

- The KS metric between the subgroup chosen and the patient survival curves, as it translates into how much a patient fits in a subgroup.
- The support of the rule that originated the subgroup (i.e., the fraction of the total number of patients in which this rule is observed). With this metric, we measure the representation of a rule, as it allows us to understand how vast a subgroup is.

Ideally, we want the KS metric to be as close to zero as possible (i.e., meaning the subgroup and patient survival curves are close) and the supports to be higher, showing that the rules have a high frequency. To see whether this applied to the subgroups found, the KS metrics and rule supports were analyzed for the same datasets used to fit the MDN, and are represented in Figure 5.4.

The obtained results are not very close to the ideal scenario. Starting with the support of the distribution rules, we observe that these achieve a maximum value of 0.1. If we take into account the TTR-FAP dataset, this does not present a problem, as the subgroup would have 10% of the 1955 patients used for training, constituting 196 patients. From a medical point of view, this constitutes a lot of patients against which to compare the patient under study. However, if we take into account the dataset with the lowest supports, Cardio, which has a maximum support of 0.0045 and only 363 training instances, we obtain subgroups of only two patients, which barely provides any information.

When looking at the KS values, their wide range of values becomes clear. Although some subgroup-patient pairs are able to obtain very low values (e.g., lower than 0.1 for the Parkinson dataset), the values are generally high, especially considering the TTR-FAP dataset. Because of this, in order to guarantee the robustness of the survival curves



(A) TTR-FAP (B) ALS

(C) Cardio (D) Parkinson

FIGURE 5.4: KS metric for all the subgroups obtained, per dataset.

predicted by the MDN it is vital to obtain complementary information regarding the subgroups found. In compliance with this, SA was performed.

5.3 Sensitivity Analysis

Sensitivity Analysis was performed with two different objectives: finding distribution rules and, consequently, subgroups, and providing information on feature importance. When it comes to the feature importance, three scenarios were created. First, we applied global SA in order to obtain the genealogical characteristics that have a higher importance to the whole set of patients. Then, and following the same line of thought, these characteristics are also obtained but on the context of subgroups, using subgroup SA. Finally, to achieve a personalized patient characterization, individual SA was performed, to obtain the patient-specific characteristics that have a higher influence on their outcome. A schematic representation of Sensitivity Analysis is shown in Figure 5.5.

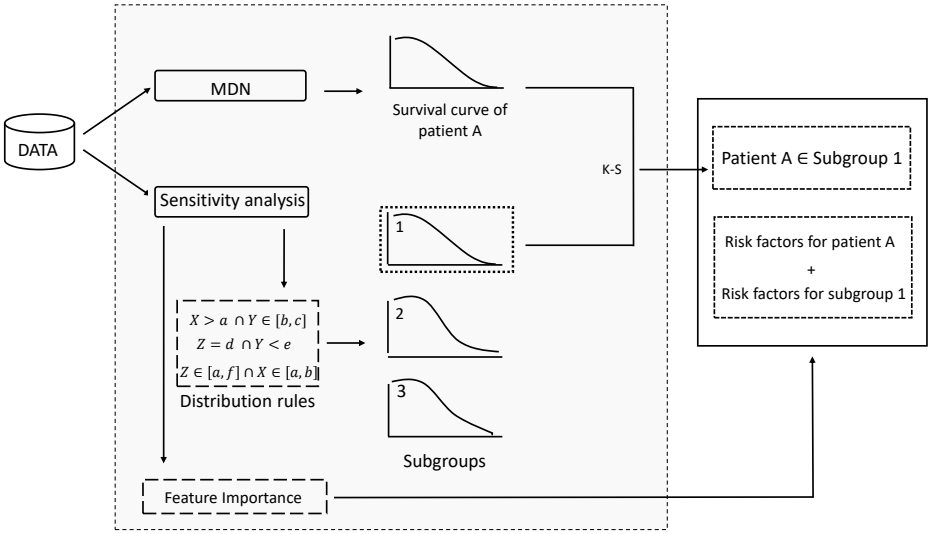


FIGURE 5.5: Approach for sensitivity analysis.

5.3.1 Global and Subgroup sensitivity analysis

For both global and subgroup SA, we used the input perturbation method [132], which is based on applying small perturbations to each feature of the input data in order to discern their respective importance in relation to the target variable. Note that, for global SA the input data is the whole patient set, while for subgroup SA we use a single subgroup of patients. A representation of the process can be observed in Figure 5.6.

Iteratively, we apply a perturbation to the values of each feature of the data while the others are left untouched. Upon each feature change, the altered data is fed to the MDN and the resulting Root Mean Squared Error (RMSE) is registered. The altered feature that results in the highest change in RMSE when compared to the original feature's RMSE,

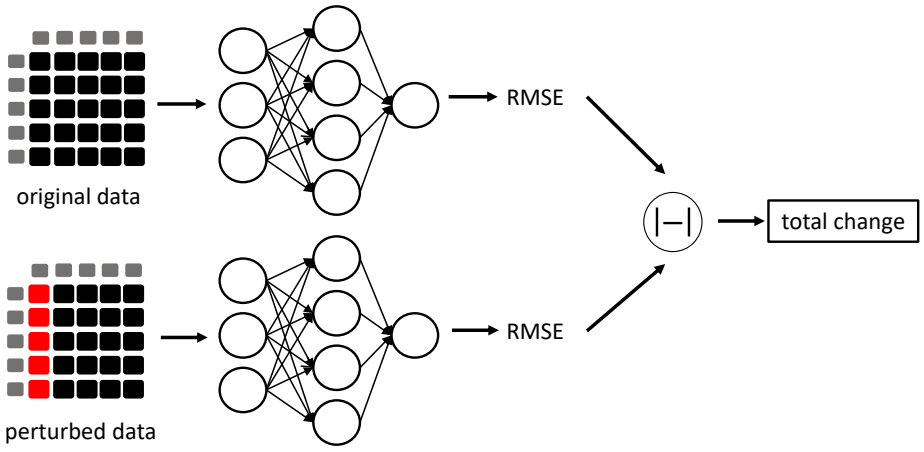


FIGURE 5.6: Global and subgroup sensitivity analysis.

is considered the most sensitive one and, therefore, the most important feature to have into account when looking at the patient's medical history. Note that here we consider change in absolute value, as it is irrelevant whether the RMSE increased or decreased when assessing feature importance.

The resulting feature importance from the global SA was compared to the feature importance of the Random Survival Forest (RSF), given the fact that the MDN proved to be at the same level as this model, performance-wise. This was done using the Kendall rank correlation coefficient, a statistic used to measure the ordinal equivalence between two samples [133]. The two samples are evaluated as presented in Equation 5.1.

$$\tau = \frac{\text{number concordant pairs} - \text{number of discordant pairs}}{\frac{n(n-1)}{2}} \tag{5.1}$$

5.3.2 Individual sensitivity analysis

For individual SA, the goal was to find, for each patient, the set of feature values that would lead to the same prediction as the original values. To accomplish this, the process was divided into two steps:

1. Iteratively, the patient value of a single feature was changed and the distribution was predicted. Using the Kolmogorov-Smirnov test, this distribution was compared to the original patient distribution. If the p-value was greater than 0.05, the feature value was stored. This step is visible in Figure 5.7.

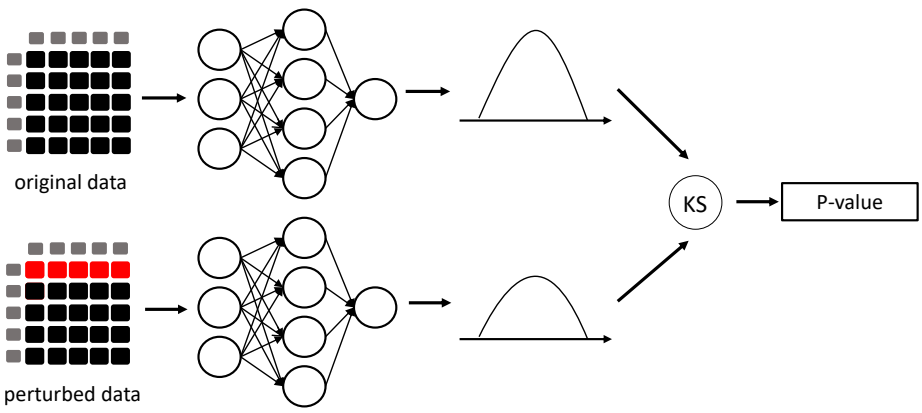
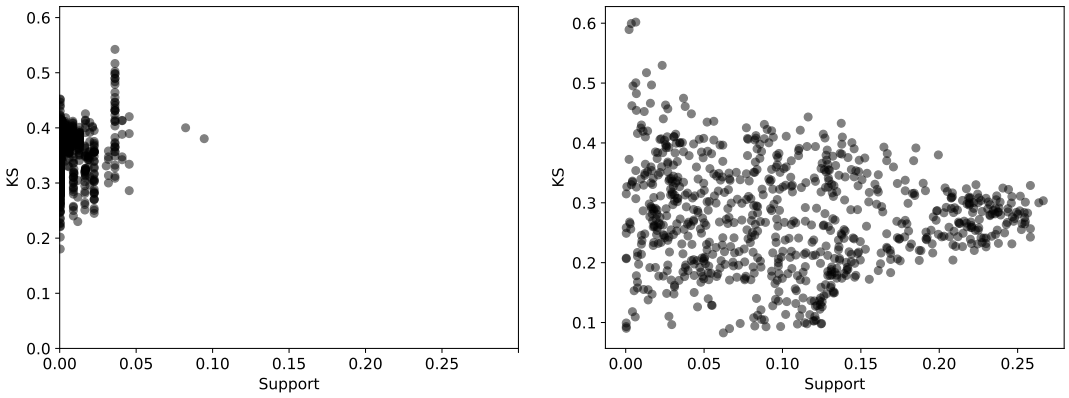


FIGURE 5.7: Individual sensitivity analysis.

2. Using the values stored, the next step consists in making all the possible combinations between them. Then, the previous step is repeated but predicting the distribution of each combination. Finally, the sets of values that result in a p-value greater than 0.05 are kept, as these constitute a distribution rule that, as in SD, is used to generate a probability density function and a survival curve.

5.3.3 Subgroups

In order to evaluate how the subgroups that were found with individual SA compare to the ones found using SD, we focus on the TTR-FAP dataset, given the fact that the SD performed worst in it. Using this approach, we foresee the KS values to decrease, given the fact that we perform an exhaustive search for the best feature combinations. In Figure 5.8 we compare the KS values obtained using this approach and SD, for the TTR-FAP data.



(A) Subgroup Discovery (B) Sensitivity Analysis

FIGURE 5.8: KS values obtained for TTR-FAP subgroup-patient pairs using Subgroup Discovery and Sensitivity Analysis.

As expected, we observe a shift of the KS metric to lower values, with a high concentration of patient-subgroup pairs within the 0.05 to 0.35 range. This is likely to happen due to the fact that SA employs an exhaustive search over all possible values of a feature, while SD uses percentile values only. Additionally, SA obtains much higher support values, reaching as high as 0.25, contrary to the highest support value of 0.09 observed with SD.

Even though there is generally a better performance with this approach, in some sporadic cases KS achieves values of 0.6, higher than any value obtained with SD. Moreover, there are specific patient-subgroup pairs where SA performs better and vice-versa, which highlights the advantage of using SD and SA in symbiotic manner.

5.3.4 Results

In Figure 5.9, we present the expected results when performing the global, subgroup, and individual Sensitivity Analysis. A subgroup and a patient were randomly chosen for the purpose of visualization. For the global SA, the whole dataset was used, and the presented patient characteristics appear by decreasing order of importance, from top to bottom. For subgroup SA, we used subgroup 1, and the patient characteristics are presented as in the global analysis. Finally, for the patient analysis, patient 206 was used, and the possible values for each variable are presented. Note that all variables are under the form of an interval, with the exception of "Sex", which is binary.

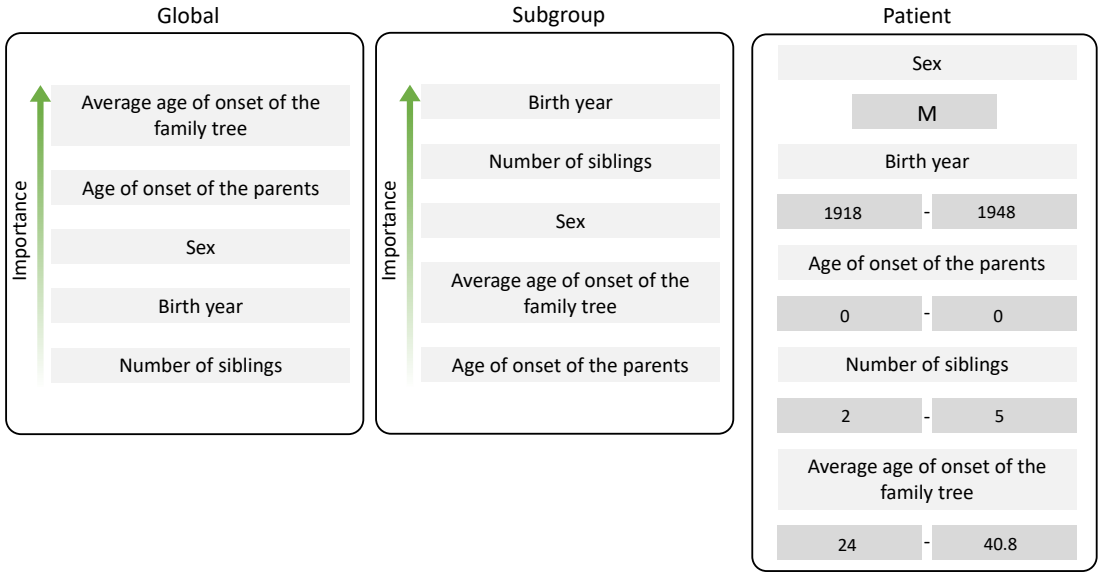


FIGURE 5.9: Global, subgroup (1) and patient (206) sensitivity analysis.

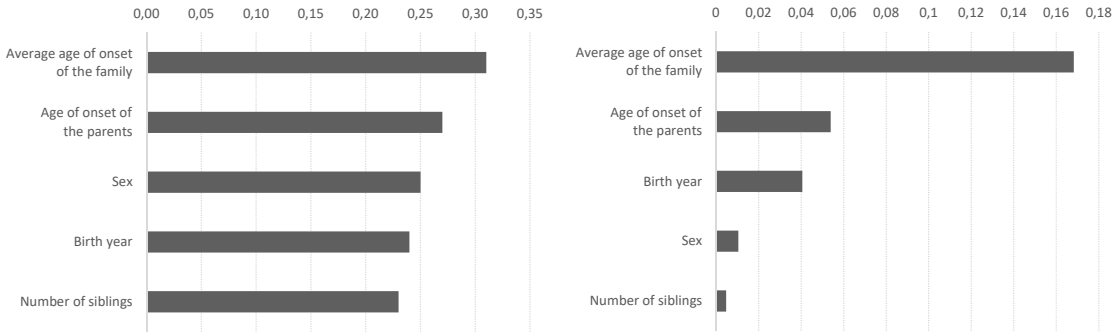
From a clinician's point of view, this representation allows for the clear assessment of the risk factors for each of the three scenarios (i.e., the whole set of patients, a set of patients in a subgroup and the patient being studied). However, it is important to assess the ability of SA to correctly identify the importance a patient characteristics, or at least compare it to other feature importance methodologies. For this, we compared the results of the global SA to the feature importance obtained by the RSF. The Kendall rank coefficients for each model are presented in Table 5.1.

TABLE 5.1: Kendall rank coefficient between the MDN and the Survival Random Forests

	TTR-FAP	ALS	Cardio	Parkinson
T	0.799	0.412	0.516	0.341
Features	6	19	10	33

For the main dataset, TTR-FAP, there is a clear affinity in the way both models rank feature importance, showing the potential of the MDN in obtaining valuable information regarding risk factors for this disease. However, as we move towards more complex datasets (i.e., with a higher number of features and fewer instances) we see this similarity decrease.

Because of this strong similarity, it is important to look closely into the differences between how these models rank feature importance, given that the RSFs obtain the feature importance differently from the approach used in SA. Instead of perturbing the values of a feature, RSF remove them completely, measuring the difference in the C-index between both scenarios (i.e., before and after the removal). Besides, RSF does not measure the absolute change, only the decrease in C-index. For comparison purposes, we also performed the global SA by registering the decrease in the C-index. The resulting bar charts of feature importance for both models are presented in Figure 5.10.



(A) Sensitivity Analysis (B) Random Survival Forests

FIGURE 5.10: Feature importance obtained with Sensitivity Analysis and with Random Survival Forests.

Analyzing both charts, we can notice a clear difference between both approaches. While RSFs are able to obtain clearer differences between the features, with a more noticeable gap between the most important feature and the remaining set, the same does not happen with the global SA approach. Taking the medical perspective into consideration, it would be beneficial to assess with clarity what are the risk factors involved. For this

reason, we believe that, for a future implementation of the global SA, the feature removal approach used in the RSF should be tested.

Chapter 6

Discussion

In this Chapter, we discuss the results obtained during our experiments. We do so while placing emphasis on answering the research questions presented earlier (which are recalled here when necessary).

We begin by discussing **RQ-1 How to obtain a human readable characterization of the patient subgroup, given a prediction?**, for which we focus on the obtained KS statistic between the survival curves of the patient and the subgroups generated both by Subgroup Discovery and Sensitivity Analysis, as this metric assesses how well a patient integrates a subgroup. When it comes to Subgroup Discovery, results show that, in all datasets, the KS values are scattered across a wide range of values, in some cases achieving values greater than 0.5. This happens with more frequency in the TTR-FAP dataset, wherein we find a high concentration of values between the acceptable range of 0.15 to 0.35, as well as pairs achieving values of up to 0.55. On the other hand, for other datasets, namely Parkinson and ALS, almost all the pairs obtained a KS within the 0.0 to 0.30 range, meaning that, for some problems this approach provides viable subgroups.

Looking at the KS values obtained with Sensitivity Analysis, we were able to observe not only a higher concentration of values in lower ranges of the KS spectrum, but also higher supports, achieving values of almost 0.3. Ideally, when looking at the KS versus support, we want the values to be on the fourth quadrant, representing a high support and low KS. However, the third quadrant also represents good results, especially when focusing on how much a patient fits into a subgroup, since it represents a low support but still low KS. Looking at the subgroups generated by Sensitivity Analysis, the vast majority of the values are distributed across these two quadrants, which translates into a good ability of this approach to find adequate subgroups for a patient.

In continuity with RQ-1, we reflect on **RQ-2 On their own, can Subgroup Discovery or Sensitivity Analysis provide enough information to obtain a robust model of Age of Onset?**, for which we take into account all the information provided by Subgroup Discovery and Sensitivity Analysis. An important aspect to have into consideration is that, as observed in the previous results, it is always more beneficial to a model if we use these two approaches synchronously. However, if only one were to be applied, Sensitivity Analysis is the methodology which is able to extract the most clinical information, given the fact that it can perform the same task as Subgroup Discovery, with the additional advantage of providing clinical risk factors both for a patient and for a subgroup. Taking all of this into account, and from a medical perspective, this method constitutes an advantage to the modeling of survival curves of Age of Onset (AOO).

We now focus on **RQ-3 How to produce a Machine Learning model able to accurately predict personalized survival curves for the Age of Onset and similar problems?**, where we analyze the predictive modeling capability of the Mixture Density Networks (MDN). To answer this, we take into account the two types of methodologies to which we compared this model: the regression Machine Learning (ML) methods and the survival analysis ML methods. Regarding the first type, and comparing the Root Mean Squared Error (RMSE) of the models, we observe that the MDN was able to outperform all the models in three of the four problems (i.e., TTR-FAP, Cardio and ALS). In addition, in these three problems, we also confirm a statistically significant difference in the performance of the MDN and the majority of the models tested, namely the Logistic Regression, Elastic Net, Lasso Regression, Ridge Regression and Decision Tree Regressor.

However, on the Parkinson problem, we observe a high discrepancy between the performances of the MDN and the remaining models, with the MDN being the worst performer. One of the reasons for this phenomenon could be the fact that the Parkinson dataset has the lowest amount of instances and the highest dimensionality (i.e., amount of variables), even after pre-processing, which for the MDN constitutes an obstacle in finding an adequate set of weights to achieve a locally-good loss value. Furthermore, the poor performance of the MDN raises the question of why the performance of DeepSurv in this problem is not only good, but the best achieved by the model in all problems. A proposed explanation is the extensive optimization process DeepSurv endures, where solutions like the Nesterov Momentum and Weight Decay Regularization are applied. These allow neural networks to, respectively, avoid overshooting the minimum value of the loss function

and to generalize better in high-complexity contexts. The latter seems particularly helpful when dealing with the Parkinson dataset, which has an increased complexity. In order to verify this, however, more tests are required, either using a larger volume of data or using the same optimization process applied in DeepSurv.

Regarding the survival analysis ML algorithms, we compare the C-indexes between the MDN, Random Survival Forests (RSF) and DeepSurv. In this comparison, we observe the best performance of the MDN in the TTR-FAP and the Cardio datasets, the two problems with a higher volume of data. In the ALS problem, even though the MDN is not the best performing model, it achieves the same C-index as DeepSurv. Like with the ML pointwise prediction models, we observe the worst performance when using the Parkinson dataset. Here, the same reasoning used before applies as to why two neuronal-based models have such different performances.

Although these two comparisons already provide valuable information to answer RQ-2, one more aspect should be taken into consideration, which is the differences in feature importance ranking between the Sensitivity Analysis and the RSF. Note, however, that two different methodologies to rank features are used. Because of this, we cannot translate these results into a best performance from either one of the approaches without having medical verification of which features are, indeed, more important. Therefore, this comparison is used simply to assess the concordance between both methodologies. Having into account the Kendall rank coefficient, it is clear that in the TTR-FAP data both approaches rank the features in a similar way. However, as we move into higher-dimensional problems, like the Parkinson dataset, that similarity fades. This was expected, as the higher the number of features, the harder it is for a model to discern between their importance.

Overall, with the obtained results, it is notable that the MDN can perform as well as or better than most tested models, with its performance decreasing in problems with a high-dimensionality and low volume of data.

Finally, we discuss **RQ-4 Is it possible to use Mixture Density Networks together with Subgroup Discovery and Sensitivity Analysis to estimate survival curves that are human readable?**, which constitutes the main focus of this study. For this, we analyze all the results obtained throughout the course of this work, mainly by joining conclusions obtained from the past three research questions. First, we take into consideration the overall predictive modeling ability of the MDN. As discussed already, we observe an

undoubtable performance of the MDN in modeling the AOO of patients. Moreover, the calibration curves indicate that not only is their performance good, but it is also reliable, especially in the case of TTR-FAP, which acts in the favor of this model in the medical context.

However, error and calibration metrics on their own do not suffice when it comes to this context, as decisions need to be made with caution and clarity. Because of this, we take into account the joint performance of Subgroup Discovery and Sensitivity Analysis. The combined information of the subgroups obtained by both models, especially by Sensitivity Analysis, proved to be able to find groups of genealogically-similar patients with a similar disease history to the patient at study. From a medical point-of-view, this allows physicians to zoom into the individuals of the found subgroup and compare clinical records and outcomes to a specific patient. Additionally, Sensitivity Analysis, especially the individual type, was able to provide an even more focused characterization of the patient by providing genealogical risk factors that lead to the prediction made by the MDN. This information, also provided for the subgroups, allows for the quick identification of common genealogical factors between a group of patients.

Chapter 7

Conclusion

Transthyretin-Related Familial Amyloid Polyneuropathy (TTR-FAP) is a devastating degenerative disease with almost no treatment unless it is diagnosed very early. This constitutes a problem, as TTR-FAP is extremely costly to diagnose. Because of that, it is of utmost importance to provide the medical professionals with a possible age range for the patient to develop symptoms, as to help them build a treatment plan.

This work proposes an approach for modeling the Age of Onset survival curves of patients with TTR-FAP while providing an understandable characterization of the prediction obtained, using Mixture Density Networks, Subgroup Discovery and Sensitivity Analysis. Contrarily to some of the current methodologies in practice in the medical field, which provide a point-prediction of Age of Onset, Mixture Density Networks are able to provide a range of ages in which a patient is more likely to experience symptoms. Furthermore, by joining it with the discovery of subgroups and Sensitivity Analysis, the survival curves predicted by the Mixture Density Network are more robust from a medical point of view. This robustness is attributed to the fact that medical professionals are able to obtain information of the patients genealogically similar to the one being studied (i.e., the subgroups), and information regarding the genealogical risk factors of the patient and the subgroup they are in.

From the results obtained, we verify that, not only are Mixture Density Networks able to keep up with standard Machine Learning approaches used in survival analysis, but also outperform them in some occasions. It is also clear the usefulness of Subgroup Discovery and Sensitivity Analysis in the retrieval of information from a survival curve modeled, and consequent interpretability of a black-box model such as the Mixture Density Networks. Furthermore, from the point-of-view of applicability to other diseases, the

obtained results look promising when using the proposed approach on other datasets. However, to be completely certain of this, more data would be necessary.

Overall, and given all the information obtained throughout this work, we verify that it is possible to turn a black-box model such as Mixture Density Networks into a grey-box approach, by combining it with methodologies that focus on information mining, such as Subgroup Discovery and Sensitivity Analysis. On a closing note, and to summarize, we go over the main findings obtained through the course of this study:

1. Mixture Density Networks are computational models with a lot of potential in the modeling of Age of Onset survival curves for patients with TTR-FAP.

2. In the context of the state of art in survival analysis, Mixture Density Networks are able to model survival curves with the same predictive ability as top Machine Learning approaches.

3. With only few improvements made, it is possible to extend the usage of Mixture Density Networks to a range of other diseases.

4. Subgroup Discovery and Sensitivity Analysis are valuable tools in the context of information mining.

5. Sensitivity Analysis proves itself as a valuable tool not only to obtain information regarding feature importance, but also to define new subgroups.

6. It is possible to turn a black-box model into a grey-box model by combining Subgroup Discovery with Sensitivity Analysis tools.

As for the future directions of this work, we admit the importance of obtaining a higher volume of data to assess the applicability of this approach in the context of other diseases. Then, and probably the most important aspect for the continuity of this study, it is vital to obtain medical feedback on the usability of this approach in a hospital setting, as well as possible improvements. It would also be interesting to compare the performance of the Mixture Density Networks with the statistical models used for survival analysis. Finally, and constituting a secondary objective, we also see potential in studying the possibility of the generalization of this approach to other fields, such as detection of machine failure.

Bibliography

[1] A. Cakar, H. Durmuş-Tekçe, and Y. Parman, "Familial amyloid polyneuropathy," *Archives of Neuropsychiatry*, vol. 56, no. 2, p. 150, 2019. [Cited on pages 1, 5, and 6.]

[2] V. Plante-Bordeneuve, "Update in the diagnosis and management of transthyretin familial amyloid polyneuropathy," *Journal of neurology*, vol. 261, no. 6, pp. 1227–1233, 2014. [Cited on page 1.]

[3] V. Planté-Bordeneuve and G. Said, "Familial amyloid polyneuropathy," *The Lancet Neurology*, vol. 10, no. 12, pp. 1086–1097, 2011. [Cited on pages 1 and 5.]

[4] T. Coelho, B.-G. Ericzon, R. Falk, D. Grogan, S.-i. Ikeda, M. Maurer, V. Plante-Bordeneuve, O. Suhr, P. Trigo, and M. Benson, "A guide to transthyretin amyloidosis," 2016. [Cited on pages 1 and 6.]

[5] H. C. Boshuizen and S. Greenland, "Average age at first occurrence as an alternative occurrence parameter in epidemiology." *International journal of epidemiology*, vol. 26, no. 4, pp. 867–872, 1997. [Cited on pages 1 and 7.]

[6] F. Escolano-Lozano, A. P. Barreiros, F. Birklein, and C. Geber, "Transthyretin familial amyloid polyneuropathy (ttr-fap): Parameters for early diagnosis," *Brain and behavior*, vol. 8, no. 1, p. e00889, 2018. [Cited on page 1.]

[7] P. N. Hawkins, Y. Ando, A. Dispenzeri, A. Gonzalez-Duarte, D. Adams, and O. B. Suhr, "Evolving landscape in the management of transthyretin amyloidosis," *Annals of medicine*, vol. 47, no. 8, pp. 625–638, 2015. [Cited on pages 1 and 6.]

[8] M. Ines, T. Coelho, I. Conceicao, F. Duarte-Ramos, M. de Carvalho, and J. Costa, "Epidemiology of transthyretin familial amyloid polyneuropathy in portugal: a nationwide study," *Neuroepidemiology*, vol. 51, no. 3-4, pp. 177–182, 2018. [Cited on page 2.]

[9] LIIAD. Github repository. [Online]. Available: https://github.com/LIAAD/XSurvPred [Cited on page 3.]

[10] C. Andrade, "A peculiar form of peripheral neuropathy: familiar atypical generalized amyloidosis with special involvement of the peripheral nerves," *Brain*, vol. 75, no. 3, pp. 408–427, 1952. [Cited on page 5.]

[11] V. Planté-Bordeneuve, A. Ferreira, T. Lalu, C. Zaros, C. Lacroix, D. Adams, and G. Said, "Diagnostic pitfalls in sporadic transthyretin familial amyloid polyneuropathy (ttr-fap)," *Neurology*, vol. 69, no. 7, pp. 693–698, 2007. [Cited on pages 5 and 6.]

[12] V. Plante-Bordeneuve, "Transthyretin familial amyloid polyneuropathy: an update," *Journal of neurology*, vol. 265, no. 4, pp. 976–983, 2018. [Cited on page 5.]

[13] K. J. Rothman, S. Greenland, and T. L. Lash, *Modern epidemiology*. Lippincott Williams & Wilkins, 2008. [Cited on page 6.]

[14] I. Conceição, A. González-Duarte, L. Obici, H. H.-J. Schmidt, D. Simoneau, M.-L. Ong, and L. Amass, ""red-flag" symptom clusters in transthyretin familial amyloid polyneuropathy," *Journal of the Peripheral Nervous system*, vol. 21, no. 1, pp. 5–9, 2016. [Cited on pages xi and 6.]

[15] A. Carvalho, A. Rocha, and L. Lobato, "Liver transplantation in transthyretin amyloidosis: issues and challenges," *Liver Transplantation*, vol. 21, no. 3, pp. 282–292, 2015. [Cited on page 7.]

[16] G. Herlenius, H. E. Wilczek, M. Larsson, B.-G. Ericzon *et al.*, "Ten years of international experience with liver transplantation for familial amyloidotic polyneuropathy: results from the familial amyloidotic polyneuropathy world transplant registry," *Transplantation*, vol. 77, no. 1, pp. 64–71, 2004. [Cited on page 7.]

[17] M. T. P. M. Coelho, "Disease modifying therapies for attr amyloidoses: clinical development of new drugs and impact on the natural history of the disease," 2019. [Cited on page 7.]

[18] Y. Parman, D. Adams, L. Obici, L. Galán, V. Guergueltcheva, O. B. Suhr, T. Coelho *et al.*, "Sixty years of transthyretin familial amyloid polyneuropathy (ttr-fap) in europe: where are we now? a european network approach to defining the epidemiology and management patterns for ttr-fap," *Current opinion in neurology*, vol. 29, no. Suppl 1, p. S3, 2016. [Cited on page 7.]

[19] S. Bandyopadhyay, J. Wolfson, D. M. Vock, G. Vazquez-Benitez, G. Adomavicius, M. Elidrisi, P. E. Johnson, and P. J. O'Connor, "Data mining for censored time-to-event data: a bayesian network model for predicting cardiovascular risk from electronic health record data," *Data Mining and Knowledge Discovery*, vol. 29, no. 4, pp. 1033–1069, 2015. [Cited on page 7.]

[20] M. Ghassemi, T. Naumann, P. Schulam, A. L. Beam, I. Y. Chen, and R. Ranganath, "A review of challenges and opportunities in machine learning for health," *AMIA Summits on Translational Science Proceedings*, vol. 2020, p. 191, 2020. [Cited on page 7.]

[21] A. Shinozaki, "Electronic medical records and machine learning in approaches to drug development," *Artificial Intelligence in Oncology Drug Discovery and Development*, p. 51, 2020. [Cited on page 8.]

[22] W. Klösgen, "Explora: A multipattern and multistrategy discovery assistant," in *Advances in knowledge discovery and data mining*, 1996, pp. 249–271. [Cited on pages 8 and 9.]

[23] S. Wrobel, "An algorithm for multi-relational discovery of subgroups," in *European symposium on principles of data mining and knowledge discovery*. Springer, 1997, pp. 78–87. [Cited on page 8.]

[24] M. Atzmueller, "Subgroup discovery," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 5, no. 1, pp. 35–49, 2015. [Cited on page 8.]

[25] S. Helal, "Subgroup discovery algorithms: a survey and empirical evaluation," *Journal of Computer Science and Technology*, vol. 31, no. 3, pp. 561–576, 2016. [Cited on pages 8 and 9.]

[26] N. Lavrač, B. Cestnik, D. Gamberger, and P. Flach, "Decision support through subgroup discovery: three case studies and the lessons learned," *Machine Learning*, vol. 57, no. 1, pp. 115–143, 2004. [Cited on pages 8 and 10.]

[27] P. K. Novak, N. Lavrač, and G. I. Webb, "Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining." *Journal of Machine Learning Research*, vol. 10, no. 2, 2009. [Cited on page 8.]

[28] D. Gamberger, D. Lučanin, and T. Šmuc, "Analysis of world bank indicators for countries with banking crises by subgroup discovery induction," in *2013 36th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, 2013, pp. 1138–1142. [Cited on page 8.]

[29] C. J. Carmona, V. Ruiz-Rodado, M. J. del Jesús, A. Weber, M. Grootveld, P. González, and D. Elizondo, "A fuzzy genetic programming-based algorithm for subgroup discovery and the application to one problem of pathogenesis of acute sore throat conditions in humans," *Information Sciences*, vol. 298, pp. 180–197, 2015. [Cited on page 8.]

[30] A. Y. Noaman, J. M. Luna, A. H. Ragab, and S. Ventura, "Recommending degree studies according to students' attitudes in high school by means of subgroup discovery," *International Journal of Computational Intelligence Systems*, vol. 9, no. 6, pp. 1101–1117, 2016. [Cited on page 8.]

[31] F. Herrera, C. J. Carmona, P. González, and M. J. Del Jesus, "An overview on subgroup discovery: foundations and applications," *Knowledge and information systems*, vol. 29, no. 3, pp. 495–525, 2011. [Cited on pages xiii and 9.]

[32] M. Van Leeuwen and A. Knobbe, "Diverse subgroup set discovery," *Data Mining and Knowledge Discovery*, vol. 25, no. 2, pp. 208–242, 2012. [Cited on page 9.]

[33] H. Grosskreutz and S. Rüping, "On subgroup discovery in numerical domains," *Data mining and knowledge discovery*, vol. 19, no. 2, pp. 210–226, 2009. [Cited on page 9.]

[34] M. J. Del Jesus, P. González, F. Herrera, and M. Mesonero, "Evolutionary fuzzy rule induction process for subgroup discovery: a case study in marketing," *IEEE Transactions on Fuzzy Systems*, vol. 15, no. 4, pp. 578–592, 2007. [Cited on page 9.]

[35] M. Atzmueller, F. Puppe, and H.-P. Buscher, "Towards knowledge-intensive subgroup discovery." in *LWA*. Citeseer, 2004, pp. 111–117. [Cited on page 9.]

[36] A. M. Jorge, P. J. Azevedo, and F. Pereira, "Distribution rules with numeric attributes of interest," in *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer, 2006, pp. 247–258. [Cited on page 10.]

[37] B. Widrow and M. E. Hoff, "Adaptive switching circuits," Stanford Univ Ca Stanford Electronics Labs, Tech. Rep., 1960. [Cited on page 10.]

[38] Y. Dimopoulos, P. Bourret, and S. Lek, "Use of some sensitivity criteria for choosing networks with good generalization ability," *Neural Processing Letters*, vol. 2, no. 6, pp. 1–4, 1995. [Cited on page 11.]

[39] M. Gevrey, I. Dimopoulos, and S. Lek, "Review and comparison of methods to study the contribution of variables in artificial neural network models," *Ecological modelling*, vol. 160, no. 3, pp. 249–264, 2003. [Cited on page 11.]

[40] J. Montano and A. Palmer, "Numeric sensitivity analysis applied to feedforward neural networks," *Neural Computing & Applications*, vol. 12, no. 2, pp. 119–125, 2003. [Cited on page 11.]

[41] R. Singh and K. Mukhopadhyay, "Survival analysis in clinical trials: Basics and must know areas," *Perspectives in clinical research*, vol. 2, no. 4, p. 145, 2011. [Cited on page 11.]

[42] A. Indrayan and A. Bansal, "The methods of survival analysis for clinicians," *Indian pediatrics*, vol. 47, no. 9, pp. 743–748, 2010. [Cited on page 11.]

[43] D. G. Kleinbaum and M. Klein, *Survival analysis*. Springer, 2010. [Cited on page 12.]

[44] S. Prinja, N. Gupta, and R. Verma, "Censoring in clinical trials: review of survival analysis techniques," *Indian journal of community medicine: official publication of Indian Association of Preventive & Social Medicine*, vol. 35, no. 2, p. 217, 2010. [Cited on page 12.]

[45] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American statistical association*, vol. 53, no. 282, pp. 457–481, 1958. [Cited on page 12.]

[46] D. G. Altman, *Practical statistics for medical research*. CRC press, 1990. [Cited on page 12.]

[47] M. K. Goel, P. Khanna, and J. Kishore, "Understanding survival analysis: Kaplan-meier estimate," *International journal of Ayurveda research*, vol. 1, no. 4, p. 274, 2010. [Cited on page 12.]

[48] J. T. Rich, J. G. Neely, R. C. Paniello, C. C. Voelker, B. Nussenbaum, and E. W. Wang, "A practical guide to understanding kaplan-meier curves," *Otolaryngology—Head and Neck Surgery*, vol. 143, no. 3, pp. 331–336, 2010. [Cited on page 12.]

[49] W. N. Dudley, R. Wickham, and N. Coombs, "An introduction to survival statistics: Kaplan-meier analysis," *Journal of the advanced practitioner in oncology*, vol. 7, no. 1, p. 91, 2016. [Cited on page 13.]

[50] D. G. Kleinbaum and M. Klein, "Kaplan-meier survival curves and the log-rank test," in *Survival analysis*. Springer, 2012, pp. 55–96. [Cited on page 13.]

[51] E. T. Lee and J. Wang, *Statistical methods for survival data analysis*. John Wiley & Sons, 2003, vol. 476. [Cited on page 13.]

[52] G. J. McLachlan and K. E. Basford, *Mixture models: Inference and applications to clustering*. M. Dekker New York, 1988, vol. 38. [Cited on page 14.]

[53] C. M. Bishop, "Mixture density networks," 1994. [Cited on pages xi, 14, and 29.]

[54] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, 1943. [Cited on page 15.]

[55] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain." *Psychological review*, vol. 65, no. 6, p. 386, 1958. [Cited on page 15.]

[56] A. Cauchy *et al.*, "Méthode générale pour la résolution des systemes d'équations simultanées," *Comp. Rend. Sci. Paris*, vol. 25, no. 1847, pp. 536–538, 1847. [Cited on page 16.]

[57] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," California Univ San Diego La Jolla Inst for Cognitive Science, Tech. Rep., 1985. [Cited on page 16.]

[58] D. A. Reynolds, "Gaussian mixture models." *Encyclopedia of biometrics*, vol. 741, pp. 659–663, 2009. [Cited on page 16.]

[59] M. Sugiyama, *Introduction to statistical machine learning*. Morgan Kaufmann, 2015. [Cited on page 18.]

[60] D. I. Cook, V. J. Gebski, and A. C. Keech, "Subgroup analysis in clinical trials," *Medical Journal of Australia*, vol. 180, no. 6, p. 289, 2004. [Cited on page 19.]

[61] L. Cui, H. James Hung, S. J. Wang, and Y. Tsong, "Issues related to subgroup analysis in clinical trials," *Journal of biopharmaceutical statistics*, vol. 12, no. 3, pp. 347–358, 2002. [Cited on page 19.]

[62] S. F. Assmann, S. J. Pocock, L. E. Enos, and L. E. Kasten, "Subgroup analysis and other (mis) uses of baseline data in clinical trials," *The Lancet*, vol. 355, no. 9209, pp. 1064–1069, 2000.

[63] S. W. Lagakos *et al.*, "The challenge of subgroup analyses-reporting without distorting," *New England Journal of Medicine*, vol. 354, no. 16, p. 1667, 2006. [Cited on page 19.]

[64] M. A. Ali, P. Hickman, and A. Clementson, "The application of automatic interaction detection (aid) in operational research," *Journal of the Operational Research Society*, vol. 26, no. 2, pp. 243–252, 1975. [Cited on page 19.]

[65] L. Breiman, J. Friedman, R. Olshen, and C. Stone, "Classification and regression trees. statistics/probability series," 1984. [Cited on page 19.]

[66] A. Ciampi, A. Negassa, and Z. Lou, "Tree-structured prediction for censored survival data and the cox model," *Journal of clinical epidemiology*, vol. 48, no. 5, pp. 675–689, 1995. [Cited on page 19.]

[67] A. Negassa, A. Ciampi, M. Abrahamowicz, S. Shapiro, and J.-F. Boivin, "Tree-structured subgroup analysis for censored survival data: validation of computationally inexpensive model selection criteria," *Statistics and computing*, vol. 15, no. 3, pp. 231–239, 2005. [Cited on page 20.]

[68] X. Su, C.-L. Tsai, H. Wang, D. M. Nickerson, and B. Li, "Subgroup analysis via recursive partitioning." *Journal of Machine Learning Research*, vol. 10, no. 2, 2009. [Cited on page 21.]

[69] W.-Y. Loh, X. He, and M. Man, "A regression tree approach to identifying subgroups with differential treatment effects," *Statistics in medicine*, vol. 34, no. 11, pp. 1818–1833, 2015. [Cited on page 20.]

[70] K.-Y. Chan and W.-Y. Loh, "Lotus: An algorithm for building accurate and comprehensible logistic regression trees," *Journal of Computational and Graphical Statistics*, vol. 13, no. 4, pp. 826–852, 2004. [Cited on page 20.]

[71] J. R. Quinlan, *C4. 5: programs for machine learning*. Elsevier, 2014. [Cited on page 20.]

[72] J. Gama, "Functional trees," *Machine learning*, vol. 55, no. 3, pp. 219–250, 2004. [Cited on page 20.]

[73] W.-Y. Loh, "Regression tress with unbiased variable selection and interaction detection," *Statistica sinica*, pp. 361–386, 2002. [Cited on page 20.]

[74] H. Kim and W.-Y. Loh, "Classification trees with unbiased multiway splits," *Journal of the American Statistical Association*, vol. 96, no. 454, pp. 589–604, 2001. [Cited on page 20.]

[75] A. Zeileis, T. Hothorn, and K. Hornik, "Model-based recursive partitioning," *Journal of Computational and Graphical Statistics*, vol. 17, no. 2, pp. 492–514, 2008. [Cited on page 20.]

[76] H. Seibold, A. Zeileis, and T. Hothorn, "Model-based recursive partitioning for subgroup analyses," *The international journal of biostatistics*, vol. 12, no. 1, pp. 45–63, 2016. [Cited on page 20.]

[77] M. Thomas, B. Bornkamp, and H. Seibold, "Subgroup identification in dose-finding trials via model-based recursive partitioning," *Statistics in medicine*, vol. 37, no. 10, pp. 1608–1624, 2018. [Cited on page 20.]

[78] S. Tiendrébéogo, B. Some, S. Kouanda, and S. Dossou-Gbété, "Survival analysis of data of hiv infected persons receiving antiretroviral therapy using a model-based binary tree approach," *Journal of Mathematics and Statistics*, 2019. [Cited on page 20.]

[79] M. P. van Wie, X. Li, and W. Wiedermann, "Identification of confounded subgroups using linear model-based recursive partitioning," *Psychological Test and Assessment Modeling*, vol. 61, no. 4, pp. 365–387, 2019. [Cited on pages 20 and 21.]

[80] R. McNamee, "Confounding and confounders," *Occupational and environmental medicine*, vol. 60, no. 3, pp. 227–234, 2003. [Cited on pages xi and 21.]

[81] W. Wiedermann and X. Li, "Direction dependence analysis: A framework to test the direction of effects in linear models with an implementation in spss," *Behavior research methods*, vol. 50, no. 4, pp. 1581–1601, 2018. [Cited on page 21.]

[82] E. Dusseldorp, C. Conversano, and B. J. Van Os, "Combining an additive and tree-based regression model simultaneously: Stima," *Journal of Computational and Graphical Statistics*, vol. 19, no. 3, pp. 514–530, 2010. [Cited on page 21.]

[83] I. Lipkovich, A. Dmitrienko, J. Denne, and G. Enas, "Subgroup identification based on differential effect search—a recursive partitioning method for establishing response to treatment in patient subpopulations," *Statistics in medicine*, vol. 30, no. 21, pp. 2601–2621, 2011. [Cited on page 21.]

[84] S. Patel, S. W. Hee, D. Mistry, J. Jordan, S. Brown, M. Dritsaki, D. R. Ellard, T. Friede, S. E. Lamb, J. Lord *et al.*, "Identifying back pain subgroups: developing and applying approaches using individual patient data collected within clinical trials," *Programme Grants for Applied Research*, vol. 4, no. 10, 2016. [Cited on page 21.]

[85] M. LeBlanc, J. Moon, and J. Crowley, "Adaptive risk group refinement," *Biometrics*, vol. 61, no. 2, pp. 370–378, 2005. [Cited on page 21.]

[86] C. Huber, N. Benda, and T. Friede, "A comparison of subgroup identification methods in clinical drug development: Simulation study and regulatory considerations," *Pharmaceutical statistics*, vol. 18, no. 5, pp. 600–626, 2019. [Cited on page 21.]

[87] M. Alappattu, G. Lamvu, J. Feranec, K. Witzeman, M. Robinson, and A. Rapkin, "Vulvodynia is not created equally: empirical classification of women with vulvodynia," *Journal of pain research*, vol. 10, p. 1601, 2017. [Cited on page 21.]

[88] S. C. Almeida, S. Z. George, R. D. Leite, A. S. Oliveira, and T. C. Chaves, "Cluster subgroups based on overall pressure pain sensitivity and psychosocial factors in chronic musculoskeletal pain: differences in clinical outcomes," *Physiotherapy theory and practice*, 2018. [Cited on page 21.]

[89] M. Z. Nezhad, D. Zhu, N. Sadati, K. Yang, and P. Levi, "Subic: A supervised bi-clustering approach for precision medicine," in *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2017, pp. 755–760. [Cited on page 21.]

[90] X. Li, D. Zhu, and P. Levy, "Predicting clinical outcomes with patient stratification via deep mixture neural networks," *AMIA Summits on Translational Science Proceedings*, vol. 2020, p. 367, 2020. [Cited on pages 22 and 25.]

[91] B. Zupan, J. Demšar, M. W. Kattan, J. R. Beck, and I. Bratko, "Machine learning for survival analysis: a case study on recurrence of prostate cancer," *Artificial intelligence in medicine*, vol. 20, no. 1, pp. 59–75, 2000. [Cited on page 22.]

[92] P. Wang, Y. Li, and C. K. Reddy, "Machine learning for survival analysis: A survey," *ACM Computing Surveys (CSUR)*, vol. 51, no. 6, pp. 1–36, 2019. [Cited on pages 22 and 23.]

[93] S. B. Choi, W. Lee, J.-H. Yoon, J.-U. Won, and D. W. Kim, "Ten-year prediction of suicide death using cox regression and machine learning in a nationwide retrospective cohort study in south korea," *Journal of affective disorders*, vol. 231, pp. 8–14, 2018. [Cited on page 22.]

[94] J.-M. Calabuig, L.-M. García-Raffi, A. García-Valiente, and E.-A. Sánchez-Pérez, "Kaplan-meier type survival curves for covid-19: a health data based decision-making tool," *arXiv preprint arXiv:2005.06032*, 2020. [Cited on page 22.]

[95] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine learning*, vol. 29, no. 2, pp. 131–163, 1997. [Cited on page 23.]

[96] P. Gustafson, "Large hierarchical bayesian analysis of multivariate survival data," *Biometrics*, pp. 230–242, 1997. [Cited on page 23.]

[97] P. Arora, D. Boyne, J. J. Slater, A. Gupta, D. R. Brenner, and M. J. Druzdzel, "Bayesian networks for risk prediction using real-world data: a tool for precision medicine," *Value in Health*, vol. 22, no. 4, pp. 439–445, 2019. [Cited on page 23.]

[98] E. Biganzoli, P. Boracchi, L. Mariani, and E. Marubini, "Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach," *Statistics in medicine*, vol. 17, no. 10, pp. 1169–1186, 1998. [Cited on page 23.]

[99] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, "Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network," *BMC medical research methodology*, vol. 18, no. 1, pp. 1–12, 2018. [Cited on pages 23, 25, and 34.]

[100] P. J. Lisboa, H. Wong, P. Harris, and R. Swindell, "A bayesian neural network approach for modelling censored data with an application to prognosis after surgery for breast cancer," *Artificial intelligence in medicine*, vol. 28, no. 1, pp. 1–25, 2003. [Cited on page 23.]

[101] J. Yao, X. Zhu, F. Zhu, and J. Huang, "Deep correlational learning for survival prediction from multi-modality data," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 406–414. [Cited on page 23.]

[102] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, M. S. Lauer *et al.*, "Random survival forests," *Annals of Applied Statistics*, vol. 2, no. 3, pp. 841–860, 2008. [Cited on pages 24 and 34.]

[103] Z. Ding, "The application of support vector machine in survival analysis," in *2011 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC)*. IEEE, 2011, pp. 6816–6819. [Cited on page 24.]

[104] A. J. Smola and B. Schölkopf, *Learning with kernels*. Citeseer, 1998, vol. 4. [Cited on page 24.]

[105] F. M. Khan and V. B. Zubek, "Support vector regression for censored data (svrc): a novel tool for survival analysis," in *2008 Eighth IEEE International Conference on Data Mining*. IEEE, 2008, pp. 863–868. [Cited on page 24.]

[106] A. Widodo and B.-S. Yang, "Application of relevance vector machine and survival probability to machine degradation assessment," *Expert Systems with Applications*, vol. 38, no. 3, pp. 2592–2599, 2011. [Cited on page 24.]

[107] J. Bell, "On the age of onset and age at death in hereditary muscular dystrophy with some observations bearing on the question of antedating," *Annals of Eugenics*, vol. 11, no. 1, pp. 272–289, 1941. [Cited on page 24.]

[108] C. Brackenridge and B. Teltscher, "Estimation of the age at onset of huntington's disease from factors associated with the affected parent." *Journal of medical genetics*, vol. 12, no. 1, pp. 64–69, 1975. [Cited on page 25.]

[109] D. R. Langbehn, R. R. Brinkman, D. Falush, J. S. Paulsen, M. Hayden, and an International Huntington's Disease Collaborative Group, "A new model for prediction of the age of onset and penetrance for huntington's disease based on cag length," *Clinical genetics*, vol. 65, no. 4, pp. 267–277, 2004. [Cited on page 25.]

[110] N. G. Ranen, O. C. Stine, M. H. Abbott, M. Sherr, A.-M. Codori, M. L. Franz, N. I. Chao, A. S. Chung, N. Pleasant, C. Callahan *et al.*, "Anticipation and instability of it-15 (cag) n repeats in parent-offspring pairs with huntington disease," *American journal of human genetics*, vol. 57, no. 3, p. 593, 1995. [Cited on page 25.]

[111] L. Almaguer-Mederos, N. Falcón, Y. Almira, Y. Zaldivar, D. C. Almarales, E. Gongora, M. Herrera, K. Batallan, R. Arminan, M. Manresa *et al.*, "Estimation of the age at onset in spinocerebellar ataxia type 2 cuban patients by survival analysis," *Clinical genetics*, vol. 78, no. 2, pp. 169–174, 2010. [Cited on page 25.]

[112] S. A. Allport, N. Kikah, N. Abu Saif, F. Ekokobe, and F. D. Atem, "Parental age of onset of cardiovascular disease as a predictor for offspring age of onset of cardiovascular disease," *PloS one*, vol. 11, no. 12, p. e0163334, 2016. [Cited on page 25.]

[113] A. Sinai, C. Mokrysz, J. Bernal, I. Bohnen, S. Bonell, K. Courtenay, K. Dodd, D. Gazizova, A. Hassiotis, R. Hillier *et al.*, "Predictors of age of diagnosis and survival of alzheimer's disease in down syndrome," *Journal of Alzheimer's disease*, vol. 61, no. 2, pp. 717–728, 2018. [Cited on page 25.]

[114] J. L. Schultz, L. A. Harshman, D. R. Langbehn, and P. C. Nopoulos, "Hypertension is associated with an earlier age of onset of huntington's disease," *Movement Disorders*, 2020. [Cited on page 25.]

[115] A. Mohammadbeigi, M. Kazemitabaee, and M. Etemadifar, "Risk factors of early onset of ms in women in reproductive age period: survival analysis approach," *Archives of women's mental health*, vol. 19, no. 4, pp. 681–686, 2016. [Cited on page 25.]

[116] E. Cisneros-Barroso, J. González-Moreno, A. Rodríguez, T. Ripoll-Vera, J. Álvarez, M. Usón, A. Figuerola, C. Descals, C. Montalá, M. A. Ferrer-Nadal *et al.*, "Anticipation on age at onset in kindreds with hereditary attrv30m amyloidosis from the majorcan cluster," *Amyloid*, vol. 27, no. 4, pp. 254–258, 2020. [Cited on page 25.]

[117] M. Pedroto, A. M. Jorge, J. Mendes-Moreira, T. Coelho, U. C. de Andrade, and C. H. do Porto, "Regression based approaches for predicting age at onset," *XXIV Jornadas de Classificação e Análise de Dados (JOCLAD 2017).*, p. 53, 2017. [Cited on page 25.]

[118] M. Pedroto, A. Jorge, J. Mendes-Moreira, and T. Coelho, "Impact of genealogical features in transthyretin familial amyloid polyneuropathy age of onset prediction," in *International Conference on Practical Applications of Computational Biology & Bioinformatics*. Springer, 2018, pp. 35–42.

[119] ——, "Predicting age of onset in ttr-fap patients with genealogical features," in *2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 2018, pp. 199–204. [Cited on pages 25 and 32.]

[120] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *arXiv preprint arXiv:1708.08296*, 2017. [Cited on page 25.]

[121] A. Brando Guillaumes, "Mixture density networks for distribution and uncertainty estimation," Master's thesis, Universitat Politècnica de Catalunya, 2017. [Cited on page 28.]

[122] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014. [Cited on page 29.]

[123] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *International conference on machine learning*. PMLR, 2013, pp. 1310–1318. [Cited on page 29.]

[124] "Kaggle," https://www.kaggle.com. [Cited on page 32.]

[125] d. P. J. Rossouw JE, J. P. Benade AJS, and K. JP, "Coronary risk factor screening in three rural communities-the coris baseline study," *South African medical journal*, vol. 64, no. 12, pp. 430–436, 1983. [Cited on page 32.]

[126] A. group. Als challenge. [Online]. Available: https://www.kaggle.com/alsgroup/end-als [Cited on page 32.]

[127] J. Hlavnička, R. Čmejla, T. Tykalová, K. Šonka, E. Růžička, and J. Rusz, "Automated analysis of connected speech reveals early biomarkers of parkinson's disease in patients with rapid eye movement sleep behaviour disorder," *Scientific reports*, vol. 7, no. 1, pp. 1–13, 2017. [Cited on page 32.]

[128] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *The Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006. [Cited on page 33.]

[129] F. E. Harrell Jr, *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer, 2015. [Cited on page 34.]

[130] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1321–1330. [Cited on page 34.]

[131] J. Katzman. Deepsurv github. [Online]. Available: https://github.com/jaredleekatzman/DeepSurv [Cited on page 36.]

[132] R. Novak, Y. Bahri, D. A. Abolafia, J. Pennington, and J. Sohl-Dickstein, "Sensitivity and generalization in neural networks: an empirical study," *arXiv preprint arXiv:1802.08760*, 2018. [Cited on page 45.]

[133] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1/2, pp. 81–93, 1938. [Cited on page 46.]