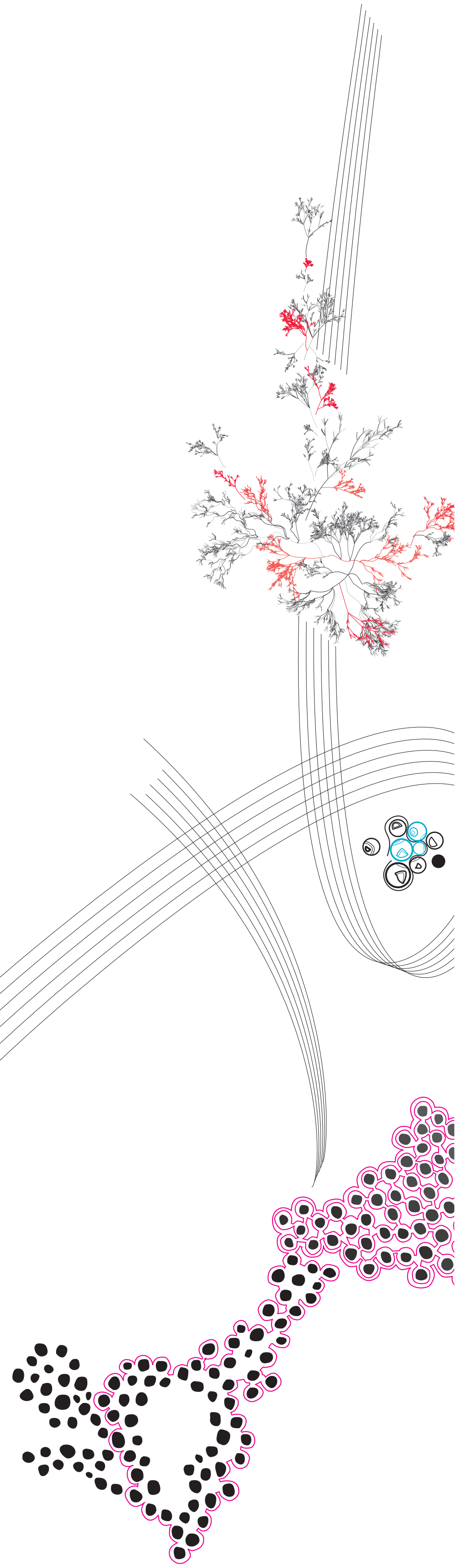
MSc Business Information Technology
Master thesis



Towards explainable machine learning for prediction of disease progression

Stijn Everard Berendse

Supervisor: Dr Faizan Ahmed

September, 2023

Faculty of Electrical Engineering,
Mathematics and Computer Science,
University of Twente

UNIVERSITY OF TWENTE.

Abstract

This research addresses the current problems surrounding interpretability of machine learning techniques in the field of prediction of disease progression. The use of machine learning in the diagnosis of diseases and the prediction of disease progression is a recent and promising development. Such predictions have potential to help physicians to make better informed decisions based on patient data, ultimately improving the patient's quality of life or altering the outcome of treatment. However, the interpretability and transparency of machine learning models aimed at this is lagging behind. This lack of interpretability and transparency is especially problematic in the application of machine learning in prediction of disease progression, as (perceived) trustworthiness is essential for predictive models in the medical domain. To achieve improved interpretability and transparency, we first perform a systematic literature review to identify the state-of-the-art in machine learning for disease progression modelling and challenges related to this context. Based on the review, we design and develop a pipeline consisting of data preparation, prediction and explanation. Predictions are made using a deep recurrent neural network based model which is followed by an integration of the LIME framework to provide explanations for each prediction. We demonstrate our pipeline by applying it to two diverse case studies using datasets on diabetes and Parkinson's disease. Besides this, we perform an experiment to compare the influence of three data imputation methods on predictive model performance in the context of prediction of disease progression. The results of this research show that there is no statistically significant difference in performance as a result of different data imputation methods. Furthermore, we provide a number of concrete recommendations and directions for future research, such as improving input flexibility of the prediction model and improving the visualisation of generated explanations. Based on the results of this research, we conclude that the proposed pipeline achieves the goal of integrating a state-of-the-art prediction model and the LIME framework to make the model more transparent and interpretable.

Keywords: Machine learning, prediction of disease progression, XAI, explainability, interpretability, Parkinson's disease, diabetes

Acknowledgements

After a summer filled with hard work, my student life has come to an end. Over the past years at the University of Twente I have had the opportunity to develop myself academically, but also on a professional and personal level. I look back on an exciting time which has prepared me for the start of my working life, filled with great experiences shared with my peers. As my study career culminates with this final research and thesis, I wish to thank several people that have supported me along the way, and during this research in particular.

First, I would like to express my deep gratitude to my thesis supervisor, Faizan, for his unwavering support and guidance throughout my thesis research. You patiently helped me stay focused on the scope of my goals, and you shared expertise that allowed me to successfully complete my research and write this thesis. Your advice and recommendations were invaluable to me during the past nine months, as were your ever well-prepared answers to any questions I posed.

Next, I wish to thank Abhishta, for making time to serve on my thesis committee as senior examiner.

Furthermore, I want to express my appreciation to Dr Hans Krabbe from Medlon, who provided me with the required data for my second case study and served as the domain expert for this case. You were an incredibly enthusiastic and motivating sparring partner during our meetings.

I also want to thank my friends and peers that I had the pleasure of meeting and working with at the University of Twente. I shared many experiences with you, ranging from training nine times per week to studying together while working towards graduation, which made my student life into the memorable time it has been. You supported me with motivation and guidance, but also distraction when needed over the course of my studies, especially during this research.

Finally, I would be terribly remiss if I did not address my parents and my sister. Thank you for the never-ending support you have given me through highs and lows, whether it be by giving advice or simply hearing me out on the phone. Without you I would not be where I am today, and I cannot express how fortunate I consider myself knowing that you always have my back.

Contents

Chapter 1

Introduction

As machine learning and artificial intelligence (AI) have become widespread in their use [1] and are reaching ever more sensitive domains such as healthcare, a need for responsible AI has arisen [2]. Normal AI systems function essentially as a "black box": the system takes input data, performs some processing, then provides a prediction. These models rely on extreme internal complexity to achieve high performance. However, this non-transparent approach has major drawbacks, such as a lack of understanding of the reasoning behind predictions, leading to decreased trust and doubts about morality [3]. Due to the need for responsible AI, previously waned interest in the field of XAI (eXplainable AI) increased again. In addition to providing predictions based on input data, XAI is designed to provide a way of explaining how and why a certain prediction was made based on the given input [4].

Disease progression modelling is a concept that has proven useful in the treatment of diseases, particularly in chronic disease [5]. Various literature exists on applying disease progression modelling for specific diseases, such as Alzheimer's disease [6] and diabetes [7]. Modelling the progression of a disease allows physicians to more accurately determine whether a patient's condition will worsen, and if so, at what rate this will occur. This is important because inability to accurately determine if and when to start treatment can have dire consequences. For example, medication for multiple sclerosis can have strong adverse effects [8, 9]. In such cases, the advantages and disadvantages of starting treatment must be weighed carefully. Research has shown that machine learning performs well in the field of disease modelling [10]. However, there remain problems with the interpretability of such models, so further research on this subject is important for the field.

For machine learning based prediction of disease progression and subsequent decision making to be a trustworthy, ethical practice, it is essential to improve the transparency and interpretability of models used for this prediction task. In this paper, we first explore the state-of-the-art in the domain of disease detection and prediction of disease progression using machine learning by performing a systematic literature review. Based on this review and further background a prediction model is chosen, which is extended with a degree of explainability using the popular LIME framework and a minor proposed extension of the LIME Python implementation. Consequently, we apply this explainable pipeline on two case studies. The first case study uses a dataset from the Parkinson's Progression Markers Initiative (PPMI). The second case study is performed with a dataset on diabetes patients provided by Medlon, a medical laboratory in Enschede, the Netherlands. For each of these case studies, we start with an experiment using the pipeline up to and including the prediction phase. This experiment consists of comparing various data imputation techniques by evaluating model performance. The best performing variant is then used for testing the full pipeline including the explanation phase.

The research questions formulated for this research are the following:

RQ1 How can we improve the interpretability and transparency of machine learning models aimed at prediction of disease progression?

SRQ1 What are the current trends and state-of-the-art in machine learning for disease detection and prediction of disease progression?

SRQ2 What are common data quality challenges in machine learning for prediction of disease progression?

SRQ3 What are common techniques for explaining machine learning models in the context of disease progression modelling?

1.1 Research contribution

In this research, we design and develop a deep recurrent neural network based pipeline, capable of predicting the progression of a disease based on time-series data and providing explanations for its predictions using an integration of the LIME framework. Furthermore, we provide a script to reverse data normalisation in LIME explanations. We demonstrate the functionality of this pipeline by applying it to two case studies: one case study with a small dataset, the other with a large dataset. This contribution advances the field of machine learning for prediction of disease progression by showcasing the importance of and possibilities for integrating interpretability in state-of-the-art prediction pipelines.

1.2 Thesis structure

In Chapter 2, we explore background on progressive diseases, neural networks, and explainability in machine learning and disease progression modelling. Subsequently, Chapter 3 shows how we applied the design science research methodology in this research, as well as how we design and validate our proposed pipeline using two case studies. Chapter 4 contains the results for the performed experiments per case study. This is followed by Chapter 5 where we discuss limitations of this research and provide recommendations for future research. Finally, in Chapter 6, we summarize our findings and draw conclusions on the research questions.

Chapter 2

Background

In this chapter, we discuss progressive diseases and the characteristics of two such conditions: Parkinson's disease and diabetes. Subsequently, neural networks and various relevant sub-types of neural networks are discussed. Following that, we briefly explore explainability in machine learning (ML), and highlight two explainability frameworks for ML. Finally, we discuss challenges in the field of disease progression modelling using (explainable) ML.

2.1 Progressive diseases

Progressive diseases are medical conditions that worsen over time, often leading to disability, impairment, or death. They can affect various organs and systems in the body. Examples of such diseases include Alzheimer's disease, Parkinson's disease, multiple sclerosis, and diabetes. The causes and mechanisms of progressive diseases are complex and diverse, depending on the type and stage of the disease. Several factors may contribute to disease progression, ranging from genetic predisposition to aging and lifestyle habits.

The diagnosis and treatment of progressive diseases are challenging and often require a multidisciplinary approach. The goals of therapy are to slow down or halt the progression of the disease, to alleviate the symptoms and complications, and to improve the quality of life of the patients. However, there is no cure for most progressive diseases, and they remain a major social and economical burden for individuals and society [11, 12].

2.1.1 Parkinson's disease

Parkinson's disease (PD) is a complex neurological disorder, mainly caused by loss of dopamine producing cells in the brain. Since dopamine is a neurotransmitter that helps control movement, PD can cause a variety of motor symptoms, such as the characteristic tremor, muscle rigidity, slowness of movement, and impaired balance and coordination. However, PD can manifest itself in a wide range of symptoms, including a variety of non-motor impairments that may precede the motor symptoms by more than a decade [13]. These non-motor symptoms range from depression to loss of smell and overall cognitive decline. Because of its complexity, the progression of PD is heterogeneous, which means that treatment goals differ per patient.

Since no cure for PD exists, early recognition of prodromal PD does not have implications for the onset of the disease, but can aid in timely mitigation of some of its symptoms [14]. Management of the disease consists primarily of regulating dopamine, either by increasing dopamine concentrations or directly stimulating dopamine receptors [13].

To monitor the progression and condition of PD patients, various assessments have been developed. The most widely used assessment for PD is the Movement Disorders Society Unified Parkinson Disease Rating Scale (MDS-UPDRS) [15]. The MDS-UPDRS consists of four partitions, assessing both motor and non-motor aspects of the disease. The first part is focused on non-motor experiences in daily life, whereas the second part focusses on the motor experiences in daily life. Part three is a motor examination, and part four is aimed at any motor complications. The cumulative score across all parts of the assessment, can be seen as a proxy for overall disease severity in a patient.

2.1.2 Diabetes

Diabetes is a chronic disease that occurs if the body does not produce sufficient levels of insulin, or is not capable of properly using the insulin it produces. Insulin is a hormone that regulates the blood sugar level. A common effect of uncontrolled diabetes is hyperglycaemia, also known as elevated blood glucose. Hyperglycaemia may lead to severe damage to the body over time, in particular to blood vessels and nerves. According to the World Health Organisation (WHO) [16], the prevalence of diabetes has skyrocketed over the past decades, increasing from 108 million patients in 1980 to 422 million in 2014. Additionally, between 2000 and 2019, a 3% increase in mortality rates by age was observed. Thus, it is clear that diabetes is a world-wide health problem of increasing proportions.

The American Diabetes Association offers a description of diabetes, its causes, and its consequences [17]. Diabetes mainly exists in two types: diabetes type 1 and type 2. Diabetes type 1 is caused by an auto-immune reaction, which leads to the body no longer producing (sufficient) insulin due to the destruction of beta-cells in the pancreas. Type 1 is generally diagnosed at a young age. Around 5% of diabetes patients suffer from this variant. Diabetes type 2 is often developed at a later age, and is caused by the body being unable to use the insulin it produces to regulate blood glucose levels to an adequate degree. This variant has been found to be related to an inactive lifestyle, obesity, and unhealthy food intake. Long term consequences of uncontrolled diabetes include heart disease, vision loss, and kidney disease.

As diabetic kidney disease (DKD) develops in 30% of patients with diabetes type 1 and 40% of patients with diabetes type 2, the overall prevalence of DKD is high [18]. On top of this high prevalence, DKD is linked to an increased mortality risk [19]. Currently, diagnosis is often reached only after severe complications of DKD have already manifested [20]. As stated by Hussain et al. [21], early diagnosis is a cost-effective to reduce the economic and humanistic burden of DKD. Thus, if the development of DKD can be identified earlier, physicians can decide on a more effective treatment plan.

There are a number of biomarkers that are related to diabetes, such as blood glucose, and hemoglobin A1c (HbA1c) [22]. One specific biomarker is related to the development of diabetic kidney disease (DKD): the estimated Glomerular Filtration Rate (eGFR) of the kidneys, measured in mL/min/1.73m^2. This biomarker is used to monitor the development of DKD in diabetes patients. A formula to classify chronic kidney failure based on the eGFR was developed in 1999 and refined in 2009, in the form of the Glomerular Filtration Rate estimation equation by the Modification of Diet in Renal Disease study group (GFR-MDRD) [23] and Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) [24] respectively.

2.2 Neural networks

Neural networks (NNs) are a machine learning technique where a network of neurons attempts to solve a problem. As stated by Krogh [25], the structure of NNs is similar to early models of the human brain, with multiple layers of neurons together forming the network. In a simple NN, the first layer (or input layer) receives input, which it attempts to interpret by having each neuron in the layer weigh the input and calculate whether the weighted input exceeds a certain threshold. If it does exceed this threshold, the neuron will output 1. If not, it will output 0. The exact threshold is determined by what is known as the "activation function" of the neuron. This output is then propagated to the next layer of neurons (known as the hidden layer), that attempt to further interpret this intermediate output the same way. The output from this hidden layer is then fed into the final layer, which is the output layer. This output layer gives the final prediction, such as the classification of a potential patient as having a disease.

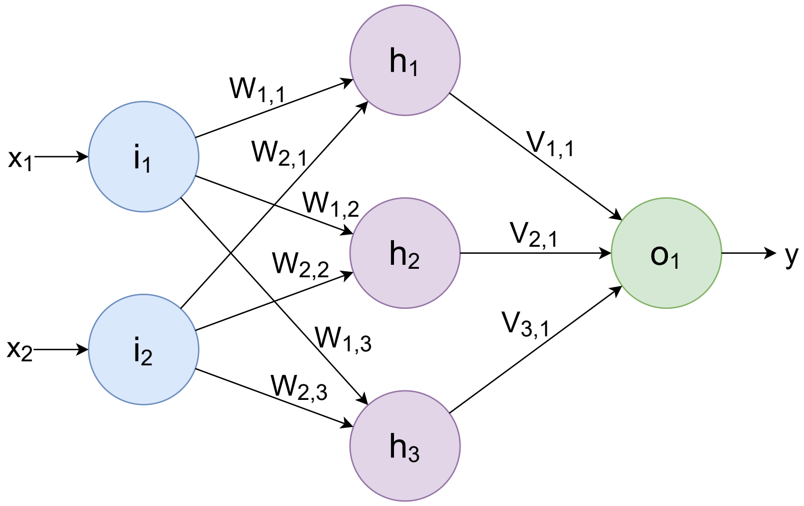


FIGURE 2.1: An example of a simple neural network topology.

For an example, refer to Figure 2.1. In this figure, we see a simple, fully connected NN topology, consisting of an input layer with two neurons, a hidden layer with three neurons, and an output layer with a single neuron. As neurons i_1 and i_2 are the input neurons, their outputs are x_1 and x_2 respectively. These outputs are then propagated to the hidden layer (depicted in purple) via the weighted edges between the two layers. Assuming we have an activation function f(x), the output z_n of a hidden layer neuron h_n can then be calculated using Equation 2.1.

$$z_n = f(x_1 W_{1,n} + x_2 W_{2,n}) \tag{2.1}$$

For each hidden layer neuron h_n, the output z_n will be propagated to the final output layer. Here, the network output y is calculated using Equation 2.2.

$$y = f(z_1 V_{1,1} + z_2 V_{2,1} + z_3 V_{3,1}) \tag{2.2}$$

To be able to predict accurately, an NN must be trained. This is generally done by feeding the network a large amount of input data, along with the ground truth of the prediction target for that data. By "reading" the input and letting this travel through the network, an output is reached and compared to the ground truth. By evaluating how close to the correct answer the network was, the NN can change some of the weights in the network to improve the network's performance for the next prediction attempt [25]. Various variations and subclasses of NNs exist, some of which we will discuss below as they are of relevance to this research.

2.2.1 Deep learning

Deep learning is an extension of NNs, aimed at solving more complex problems than a simple NN. The main difference between deep learning and a regular NN, is that a basic NN generally consists of three layers (an input layer, a hidden layer, and an output layer), while a deep learning model consists of at least two hidden layers [26]. The addition of these hidden layers allows a deep learning model to handle more complex problems, as there are more "paths" through the network, which in turn allow for more sophisticated interpretation of data. There are downsides to this increased complexity however: an increase in the amount of data and time required to train the model [27], as well as a lower degree of transparency in how the model comes to its predictions [28]. Most (deep) NNs are feed-forward networks, meaning that each layer in the network feeds its output forward to the next layer. These types of networks are highly capable for prediction tasks such as classification of images. There is another class of network, however, which cannot only feed layer outputs forward, but also back to the same layer. These networks are known as Recurrent Neural Networks, or RNNs.

2.2.2 Recurrent Neural Networks

Although RNNs were introduced some 35 years ago in the late 1980s [29], they have recently gained traction in a wide variety of prediction tasks. One of these is time-series forecasting [30]. Since RNNs have the ability to back-propagate the output of a cell (sometimes known as a unit) into the system, they are inherently well suited to handling sequential data. This back-propagation mechanism allows RNNs to effectively memorize parts of the input using their hidden state. By combining the hidden state from a previous step with new data from the current step, RNNs can learn relations between datapoints over time. Basic RNNs suffer from a problem known as the vanishing gradient problem [31], caused by the way RNNs use the same weights for each forward and backward propagation steps. In essence, the vanishing gradient problem means that the "further away" a datapoint is, the less it is represented in the model.

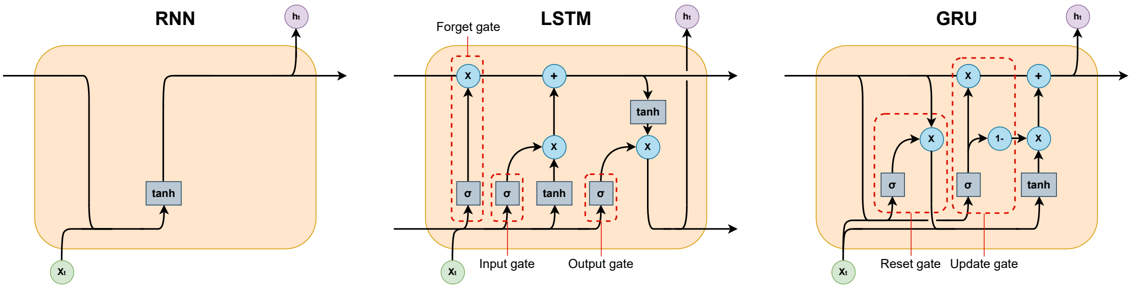


Figure 2.2: A schematic overview the internal cell structure of various RNN types. Adapted from [32].

2.2.3 LSTM and GRU

Since the introduction of RNNs, two popular architectures that mitigate the vanishing gradient problem have been developed: LSTM [33] and GRU [34]. Figure 2.2 shows the difference in internal structure between regular RNN, LSTM, and GRU. Both of the latter architectures make use of internal gates to more strictly control what information is retained in the recurrent units that make up the network layer. In an LSTM unit, for example, keeping the input gate activation near 0 prevents the activation of the cell being overwritten by new inputs in the network. Subsequently, by opening the output gate, the activation of the unit can be made available to the network at a later point in time [35].

The main difference between LSTM and GRU based models is the number of gates per unit. LSTM units have three gates (input, output and forget), while GRU have two (reset and update). The more gates a unit contains, the more precisely it can control the representation of data in the network. LSTM and GRU are highly similar in functionality and performance in most cases. However, due to the less complex internal structure of GRU compared to LSTM, GRU is more computationally efficient and learns quicker for prediction tasks where the amount of sequential data is small or sequences are short. If sequences are long, for example a five-year record of daily stock prices, or a large amount of data and computational resources is available, LSTM's more sophisticated structure generally outperforms GRU, as it can make more complex internal representations of the input [36].

2.3 Explainability in machine learning

To explain complex machine learning models, various frameworks exist. Two of the most well-known frameworks are SHAP and LIME, which are both applicable to a wide range of model types.

2.3.1 SHAP

SHAP (SHapley Additive exPlanations) is a framework for interpreting the predictions of machine learning models based on the concept of Shapley values from cooperative game theory. It was introduced in 2017 by Lundberg and Lee [37]. A Shapley value is the average marginal contribution of a player to all possible coalitions of other players. In other words, it is the fair share of the reward that a player deserves for participating in a game [38]. SHAP applies this idea to machine learning models, where the features are the players and the prediction is the reward. SHAP computes the Shapley values for each feature and each prediction in such a way that the sum of the feature contributions equals the prediction value.

However, some features may not have a fixed contribution based only on the value of the feature itself, but may depend on the values of other features. These interactions between features complicate the interpretation of the model. SHAP handles interactions by using a generalization of Shapley values, called SHAP interaction values. These values measure not only the individual contribution of each feature, but also the pairwise contribution of each pair of features. SHAP interaction values can be visualized as a matrix, where the diagonal elements are the main effects and the off-diagonal elements are the interaction effects. SHAP interaction values can help identify which features are synergistic or redundant in the model, as well as how they affect the prediction.

Furthermore, while SHAP supports a range of complex model types including deep learning models, and there is an example using an LSTM model in the Python library GitHub repository[1], the Python implementation does not offer functionality for time-series forecasting using RNNs at the time of writing.

2.3.2 LIME

The LIME framework is a method for explaining the predictions of any machine learning model, regardless of its complexity or architecture. The framework was introduced in 2016 by Ribeiro et al. [39]. The acronym LIME stands for Local Interpretable Model-agnostic Explanations, which we can explain as follows:

- *Local* refers to local fidelity, meaning that the explanation should reflect the behavior of the model around the instance being predicted, not necessarily globally for all possible instances.

- *Interpretable* refers to the goal of LIME, which is making complex models understandable using simple and intuitive features or representations.

- *Model-agnostic* implies that the framework can be applied to any type of model, from linear regression to deep neural networks, as long as the model can provide predictions for any input. This also means it will be capable of explaining not yet developed models.

- *Explanations* are the output of the framework, which consist of a set of features and values, as well as their corresponding weights that indicate how much each feature contributes to the prediction.

The main concept behind LIME is to create a surrogate model that approximates the original model locally, using a simple, inherently explainable model such as a linear regression or a decision tree. Global interpretation can only be achieved by taking several local explanations, and analysing them manually [40]. The surrogate model is trained on a set of perturbed instances that are generated by randomly tweaking some features of the original instance, and obtaining the predictions of the original model for these instances. These perturbed instances are then weighted by their similarity to the original instance, so that instances that are more similar to the original have more influence on the surrogate model. The features and weights of the surrogate model are then used as the explanation for the prediction of the original model. As such, the explanation offered by LIME is merely an approximation of the true inner process of the model to be explained. However, due to the similarity check between perturbances and the original instance, this approximation holds for the local domain that LIME attempts to explain.

As is the case for SHAP, Ribeiro, the primary author of the LIME paper, offers a Python implementation of their framework on GitHub[2]. In contrast to the SHAP implementation, the LIME Python implementation includes support for time-series forecasting with RNNs, using the Recurrent Explainer classes. This makes the framework suitable for time-series forecasting interpretation. Each data sequence is split in both features and timepoints, so the influence of each feature at each timepoint is calculated and visualised separately.

[1] `https://shap.github.io/shap/notebooks/deep_explainer/Keras%20LSTM%20for%20IMDB%20Sentiment%20Classification.html`

[2] `https://github.com/marcotcr/lime`

2.3.3 Comparing SHAP and LIME

While both the SHAP and LIME frameworks aim to provide explanations for complex model predictions, they differ in the way they generate and weight the features, as well as in their properties and guarantees. As discussed above, LIME constructs its explanations by approximating the model to be explained at a local point, whereas SHAP constructs explanations by considering all possible feature combinations during the calculation of the Shapley values at a local point. As a result, LIME is not capable of offering a global interpretation. The SHAP framework is capable of this, by aggregating information on various local instances. Due to this, SHAP is preferred for global interpretation. For local interpretation, the preferred framework depends on the characteristics of the case that one is trying to implement interpretability for.

2.4 Challenges in (explainable) disease progression modelling

In the literature review included in Appendix C, we explore the state-of-the-art and research gaps in the field of machine learning for disease detection and prediction of disease progression. From this, we are able to precisely determine the prediction task that we set for this research: predicting a continuous value representing disease severity, based on time-series data, in such a way that the prediction can be explained. Along with this, we identify several challenges for the set prediction task. These will be discussed in this subsection.

2.4.1 Time-series data

The first challenge we identify is the use of time-series data. Time-series prediction is a complex task, because the model not only needs to interpret features and their relations at one data point, but also for multiple data points over time. To overcome this challenge, we use the aforementioned RNN type of model for its capability of interpreting this complex type of data. As we have discussed RNNs at length earlier in this chapter, we will not do so again in this section.

2.4.2 Healthcare data

Another challenge that is commonly reported in the literature, is the quality and contents of healthcare data. Various authors mention the high degree of missing data in their datasets, but also the inherent heterogeneity that exists in patient data. The sophisticated nature and associated ability of (deep) NNs to interpret complex data can be seen as a mitigation for the latter challenge. However, for solving missing data issues, a separate approach based on pre-processing is preferred according to Whang and Lee [41].

One of the methods to handle missing data is imputation. This practice consists of using some, often statistical, method to fill missing values based on surrounding values. Examples of such methods are mean imputation, where the missing value is filled with the mean value across other records, and forward/backward filling, where missing values are filled with the previous or following record's value [42].

In the case of time series data, mean imputation can be done in two ways: either the entire population mean is used across all time points, or we can use the "visit mean". When using the latter variant, for each time point in the sequence (say visit number three), a mean is calculated across the values recorded at the third visit of each patient. It is possible that in the case of predicting progression of diseases, the event mean is more representative for missing values than the population mean. The hypothesis is that the event mean accommodates the change of a value over time as a result of the progressive nature of the disease influencing the value, in turn leading to a more representative imputed value. We note that the event mean imputation method assumes that progression of patients is homogeneous across the population, even though that may not be the case.

Forward/backward filling is interesting in the sense that it works completely different from the aforementioned mean-based imputation methods. Forward/backward filling possibly represents progression over time less strongly than event mean imputation because it simply copies the last or next value in a sequence, and thus does not capture intermediate values. However, by doing so forward/backward filling does accommodate the highly heterogeneous nature of patient data: other patients' data does not have any influence on the value that will be filled. As a result, a patient that always had very high values for a certain biomarker, will have a missing value imputed with a similarly high value that is representative of the patient, rather than a lower value that is representative of the population.

2.4.3 Explainability for disease progression modelling

As mentioned earlier in this section, one of the drawbacks of deep learning is the lack of transparency that the models offer. While this is a known issue, there is not a lot of literature that is aimed at mitigating this issue in the context of predicting progression of diseases. Of all the 64 pieces of literature included in the review performed in Appendix C, only a single paper mentions explainability or transparency of the models they propose or review. Kendrick et al. [43] explicitly state that neural network based models offer low, if any, explainability. Their conclusion is that if a degree of explainability is desired, one should apply Support Vector Machine based models because these offer comparable performance. However, the prediction task that the models reviewed by Kendrick et al. are applied to is significantly less complex than the prediction task we have set for this research: there is no longitudinal data or trajectory prediction involved. Because of that, their advice cannot be taken for this research. The pipeline proposed in this research addresses that, while using existing techniques for both prediction and explanation.

Chapter 3

Methodology

This section is split in four sections. The first section discusses the use of the design science research methodology to perform this research. In the second section we discuss the general approach, model architecture, and a number of settings that are used to predict disease progression. As we have two case studies, each further section will describe the approach for that specific case study. We opted for this structure, because while the model architecture remains the same, the dataset, the data pre-processing steps and model hyperparameters are adapted to each case study.

Each case study section is divided in three parts. In the first part of each section, we will give a general overview of the dataset that is used and subsequently discuss the contents of the dataset. The second part will discuss the data preparation process, followed by a description of the model setup as the third and final part.

As mentioned in Chapter 1, before including the explanation phase in our pipeline, we first perform an experiment where we compare the influence of various imputation techniques on the performance of the predictive model. The imputation techniques that will be compared are population mean imputation, event mean imputation, and forward/backward filling. In this comparison, population mean imputation serves as the baseline due to its simplicity and lack of accommodation for any specific characteristics of patient data. The way each technique works is discussed in Chapter 2. The best performing imputation technique is used with the full pipeline, including explanation of predictions.

3.1 Design Science Research Methodology

This section will briefly describe the design science research (DSR) paradigm and the design science research methodology (DSRM). Subsequently, we show how the DSRM is applied in this research.

3.1.1 Design science research

According to Wieringa [44], DSR is the practice of designing and investigating an artefact within its context. The artefact in context is the object of study. This explicit combination of artefact and context is deliberate, as the interaction between the artefact and the problem context is what allows the artefact to solve the problem. There are two types of research problems in DSR: knowledge questions and design problems. The first is focused on gaining theoretical knowledge about the artefact in the problem context, the latter is focused on designing an artefact in such a way that it improves or evolves in the problem context.

In this research, the main research question RQ1 formulated in Chapter 1 serves as our design problem. There are some artefacts (machine learning models) and a context (the field of predicting progression of diseases) that we are trying to improve (by making the models more transparent and interpretable). This design problem raises several knowledge questions about the artefact and context: our subquestions. These questions are answered in the systematic literature review in Appendix C and Chapter 2. Using the knowledge and precisely defined design problem that we gain from answering the knowledge questions, we design and develop the artefact in context with the aim of improving the artefact in context. The artefact that we design and develop in this research is the proposed machine learning pipeline.

3.1.2 The design science research methodology

To provide guidance for design science researchers, the DSRM was published in 2007 by Peffers et al. [45]. According to their methodology, a DSR should consist of the following six activities:

1. *Problem identification and motivation.* Define the exact research problem that must be solved.
2. *Defining the objectives for a solution.* Define what the objectives for the solution are, while respecting (technical) feasibility.
3. *Design and development.* Determine the required functionality and architecture of the artefact and create it.
4. *Demonstration.* Demonstrate the artefact by solving an instance of the problem.
5. *Evaluation.* Observe to what degree the artefact is a solution to the problem. This can be done by comparing the results with the objectives defined in step 2.
6. *Communication.* Communicate all elements of the research: the problem, the importance of solving it, the artefact, why the artefact is designed the way it is, and how useful the artefact is.

Note that there is no strict ordering required when one applies the DSRM. The above steps can be ordered according to the specific needs of the problem at hand. For example, researchers may observe a problem which triggers DSR, leading to a problem-centered approach that starts at step 1. Similarly, it is possible that an artefact exists, but is not yet tried as a solution for the problem context. This could be the case if an artefact is created for another domain where it solves another problem. If research starts at this point, this leads to a design- and development-centered approach.

3.1.3 Applying the DSRM

In this research we adopt a problem-centered approach, as shown in Figure 3.1. There is a lack of transparency and interpretability that is observed in the field of machine learning (ML) for prediction of disease progression. The observation is made for ML in general as well, but as mentioned in Chapter 1, applications of such systems in the healthcare domain require a relatively high level of trustworthiness. This trustworthiness can be increased by providing more interpretability and transparency with regards to how ML models work when applied to predict the progression of a disease.

The objective of the research is to develop a ML solution that is capable of predicting a continuous value representing progression of a disease based on time-series data using a state-of-the-art ML model, while also offering an explanation why the model comes to this prediction.

The artefact we create in this research is the proposed ML pipeline. The pipeline includes three main stages: data preparation, prediction of progression, and explanation of the prediction. For each of these stages, a separate knowledge question was answered to find the best possible design of the artefact in context.

For the demonstration of the artefact, we perform two case studies in which we apply our artefact to a healthcare dataset. One of these case studies concerns a small dataset, the other a large dataset, to ensure we demonstrate the use of our artefact in the range of context in which it is applicable.

The evaluation step as it is described in the DSRM is not yet applicable to research. The DSRM makes a clear distinction between demonstration and evaluation, where demonstration can be done using application of the artefact to some sample instance, while evaluation requires the artefact to see real use in the intended context. Thus, until the pipeline is applied in a real-world situation, we cannot claim to have performed the evaluation step.

The final step, communication of the research, is done via this manuscript as part of a graduation thesis. The research is publicly available via the University of Twente website.

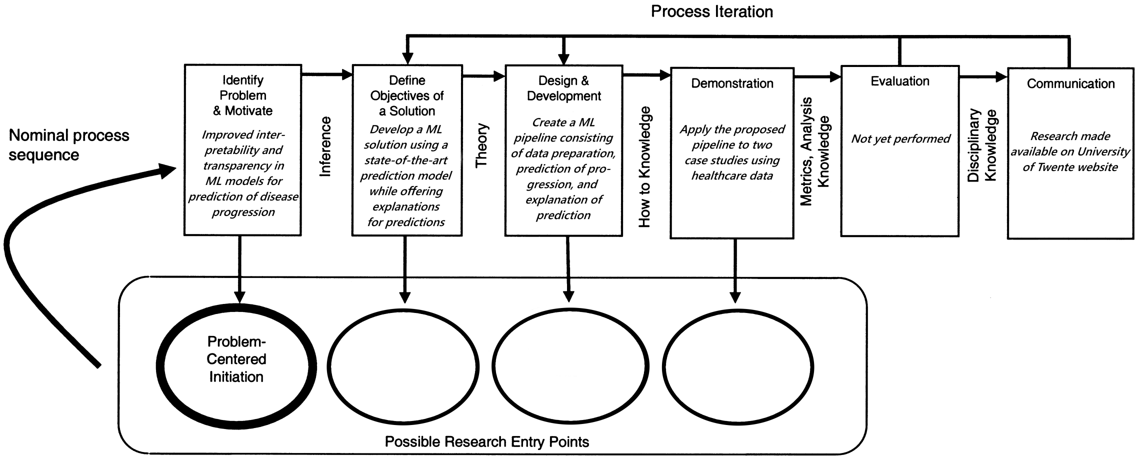

Process Iteration
Nominal process sequence
Identify Problem & Motivate
Improved interpretability and transparency in ML models for prediction of disease progression
Inference
Define Objectives of a Solution
Develop a ML solution using a state-of-the-art prediction model while offering explanations for predictions
Theory
Design & Development
Create a ML pipeline consisting of data preparation, prediction of progression, and explanation of prediction
How to Knowledge
Demonstration
Apply the proposed pipeline to two case studies using healthcare data
Metrics, Analysis Knowledge
Evaluation
Not yet performed
Disciplinary Knowledge
Communication
Research made available on University of Twente website
Problem-Centered Initiation
Possible Research Entry Points

FIGURE 3.1: The DSRM taken from Peffers et al. [45] adapted to show the process for this research.

3.2 Pipeline architecture

The architecture that we propose for our machine learning pipeline consists of three phases. The first phase of the pipeline is data pre-processing. This phase is followed by a prediction phase. Finally, the third phase is the explanation generation. For the prediction and explanation phases, we use existing techniques and frameworks, such as GRU and LIME. Figure 3.2 shows a high level overview of the pipeline. For an in depth explanation of the datasets and pre-processing of said datasets, refer to the case study-specific Section 3.3 and Section 3.4 below.

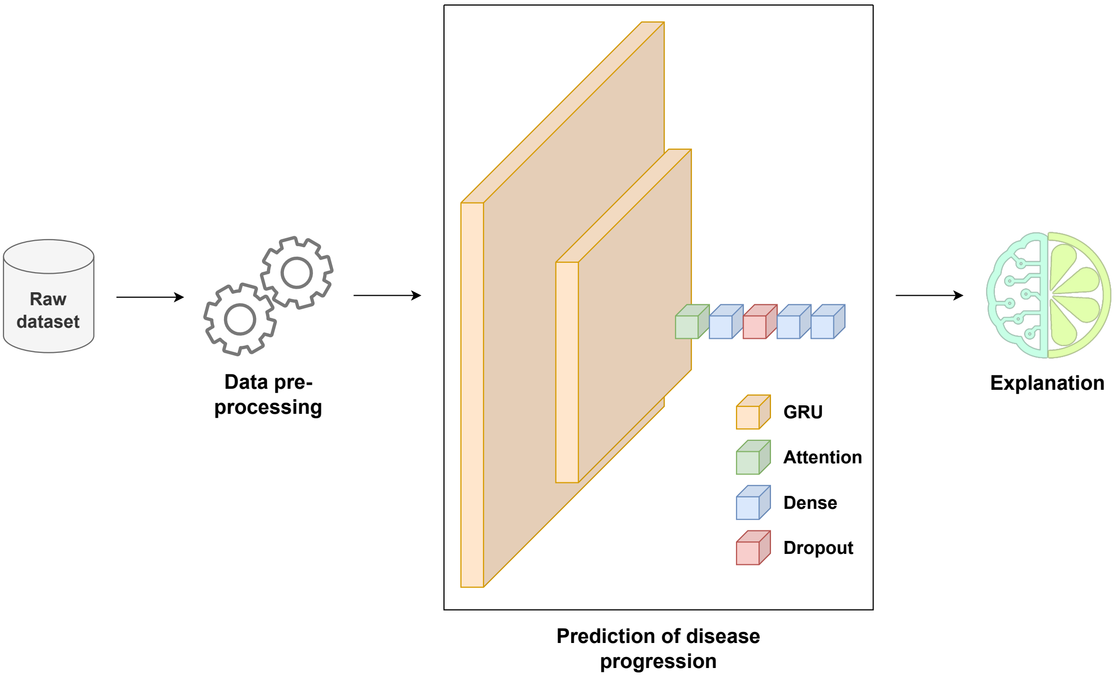


FIGURE 3.2: A high level overview of the proposed explainable pipeline.

3.2.1 Prediction model

The prediction model consists of seven layers in total. The first layer serves a dual purpose, as both input layer and first processing layer. This is the first of two GRU layers. Following the input layer, there is a second GRU layer to further interpret and capture temporal information in the data. We choose GRU over LSTM because the data sequences in our case studies are relatively short. After these two GRU layers, the shape of the data is still three-dimensional: the first dimension being the number of samples in the batch of data, the second being the length of the sequence, and the third being the number of features in each data point. Because this is a very high amount of "effective" features (sequence length $*$ number of features), it may be difficult for the following layers to properly determine what features and what timepoints are important. To aid in this, we add an attention layer. In a nutshell, an attention layer is used to generate an attention vector that emphasises what the most important parts of a piece of data are. After the attention layer, we place the first dense layer. A dense layer is a simple layer of neurons, each of which works as explained in Chapter 2. These are generally used towards the end of a network, to "convert" and connect the intermediate representations of the input to a final output. Because we want to prevent overfitting (disproportionately high performance on training data) of the predictive model, we follow the first dense layer with a dropout layer. This layer randomly drops neurons from the previous layer at a customisable rate, hence removing "fitment" from the network to seen data and generally improving performance on unseen data. Following the dropout layer, there are two more dense layers. The final dense layer functions as our output layer.

3.2.2 Implementation and setup

The implementation of the pipeline is done in Python. For the first part of the pipeline, data preparation, we use the Pandas library, which is a data analysis tool. For creating and training a predictive model, we use the Keras ML library, as it offers ready to use implementations of advanced techniques such as GRU layers. For the attention layer, we use a Keras-based implementation that is available on GitHub[1]. To convert the pre-processed data to a format that is compatible with the implementation of the predictive model, two actions are performed. The first is adding a column with "target" values by taking the value of the target variable at the next visit. The second is normalising the training features.

Finally, for the explainability phase of the pipeline, we use the Python implementation of the LIME framework as provided by Ribeiro, the primary author of the LIME paper. Because this LIME implementation is not capable of reverting normalised values to human-readable values, we created an extension script for the Python LIME implementation to perform this inversion before plotting the explanations.

A common issue we find in the systematic literature review in Appendix C is that literature is published without code or details on implementation, hindering reproducibility. For the sake of transparency, we provide access to the code used to perform this research. Because the data we use is not freely available, it is not possible to share this along with the code. All notebooks and scripts used or written for this research can be found on Github[2].

The optimiser used during the training of the predictive model is Keras' Adam optimiser. This highly popular optimiser is based on the paper by Kingma and Ba [46]. The choice of optimiser influences how the adjustment of internal attributes such as weights in the NN is done during training. We choose Adam because of its relatively low computation time and low amount of required hyperparameter tuning.

The only setup required for using the Python LIME implementation consists of choosing whether continuous feature values should be discretised, and how many features should be included in the explanation. We choose to enable discretisation, because this leads to better explanations according to Dieber and Kirrane [40]. For the number of features, we choose to include the top 20 features.

3.2.3 Model evaluation

The metric used to evaluate the model was the mean squared error, or MSE. The reason why we chose for the MSE rather than mean absolute error (MAE), is that MSE is harsher on larger errors. In the healthcare domain, the more wrong a prediction is, the greater the (adverse) effects can be. As such, we decided to use a metric that would penalise such mistakes harder than smaller mistakes.

To be able to properly evaluate the predictive model, we use a train-test split, combined with K-fold cross validation in the training phase. The test partition consists of 15% of the pre-processed data and is used to estimate the performance of the model on new or unseen data. In the training phase, 5-fold cross validation is used to mitigate overfitting, and to increase visibility of any data imbalance. Each fold combination that is used results in a trained model that is applied to the test set. Hence, for each experiment, five loss values will be reported: one per trained model. Along with these five loss values per experiment, we report the average loss across the five models as well as the standard deviation. Furthermore, because the splitting of the data and fold creation is random, seeds are used to ensure the research is reproducible. To account for possible imbalance in the data beyond the K-fold cross validation, we also perform our comparative experiment with the predictive model on four different seeds. These seeds are 787, 1998, 25 and 959143, in no particular order.

[1] `https://github.com/philipperemy/keras-attention`

[2] `https://github.com/StijnBerendse/MSc_thesis`

3.3 Case study 1: PPMI

3.3.1 Dataset description

The first case study is performed using the PPMI dataset. This dataset is part of an observational clinical study that aims to identify biomarkers of Parkinson's disease progression. The dataset is accessible to accredited researchers via a web-portal, and contains clinical, imaging and biological data on Parkinson's disease patients and various other participant groups. Participation in the PPMI is restricted to people in the United States, Europe, Israel, and Australia. The study covers participant visits from July 2010 and is still ongoing. The dataset we use in this research includes visits up to July 7th 2023.

The PPMI is a structured study project, as a study of this size requires extensive planning of when the participants visit and what assessments and measurements are taken during the appointments. This structured nature is beneficial for this research, as visit intervals are fairly constant, and visit record contents are pre-determined. As such, we do not account for time irregularity in this dataset. Visits may be a few weeks off from perfectly yearly, but we explicitly assume that this irregularity is not significant for the time scale at which the PPMI operates.

The assessments that we choose to include are all parts of the MDS-UPDRS, as well as the Epworth Sleepiness Scale. The latter of these is a questionnaire on sleepiness during daily situations. The large majority of the MDS-UPDRS assessment values may range between 0 and 4, where 0 indicates a fully normal or asymptomatic performance and 4 indicates very severe disability. For the Epworth Sleepiness Scale, all eight assessments are rated between 0 and 3, where 0 indicates no sleepiness and 3 indicates a high chance of falling asleep.

At the time of downloading the PPMI dataset for this research, the raw dataset contains 3199 participants, of whom 1558 (48.7%) are diagnosed with Parkinson's disease. In this population, there is basic demographic information for 3076 participants, of whom 1474 (47.9%) are PD patients. Out of the 3076 participants we have information on, 57.2% is male and 42.7% is female. For 0.16% of the participants, the sex is unknown. Figures 3.3 and 3.4 show the distribution of age for the full participant population and PD patient population in the raw dataset. Between these plots, we see little difference for the male population. When looking at the female population, we see a more uniform spread of ages for the PD patients compared to the plot containing all female PPMI participants. We note that the change in distribution appears to be caused by a change in the 55 to 70 year old age group. This may be caused by a relatively high number of control participants in this age range. Because the target variable (UPDRS total score) is an aggregation of various columns that is created during the data preparation process, we do not have information on this before dataset preparation is completed.

The dataset used in the PPMI case study is constructed by taking information from a collection of CSV files that each contain a different partition of the required information. Five of these CSVs contain records of MDS-UPDRS assessments performed during patient visits, one CSV contains Epworth Sleepiness Scale assessments, and another three CSVs contain patient information such as patient sex and age.

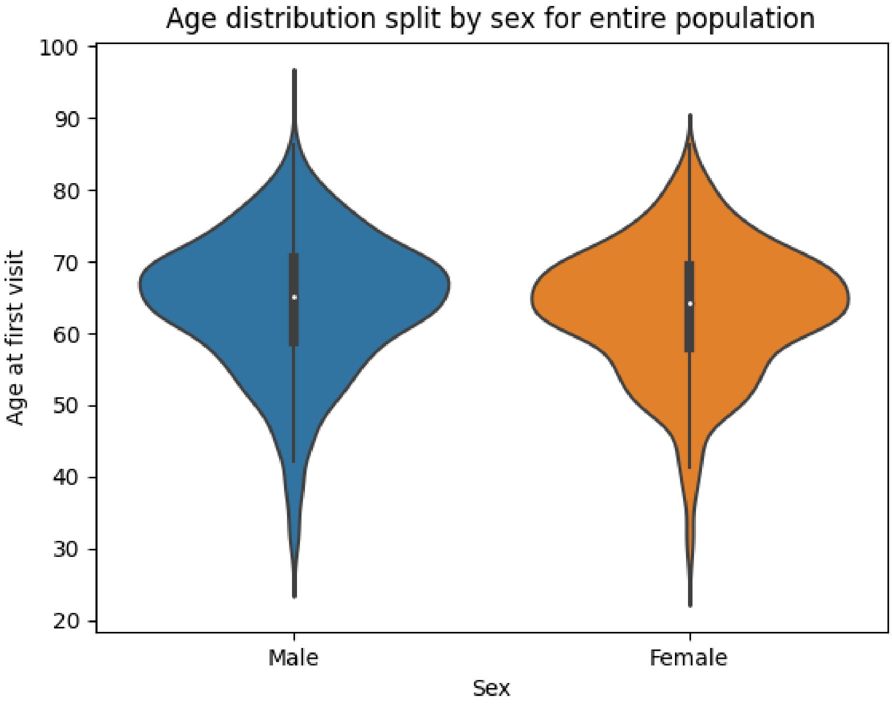


Figure 3.3: The age distribution for all participants in the raw dataset.

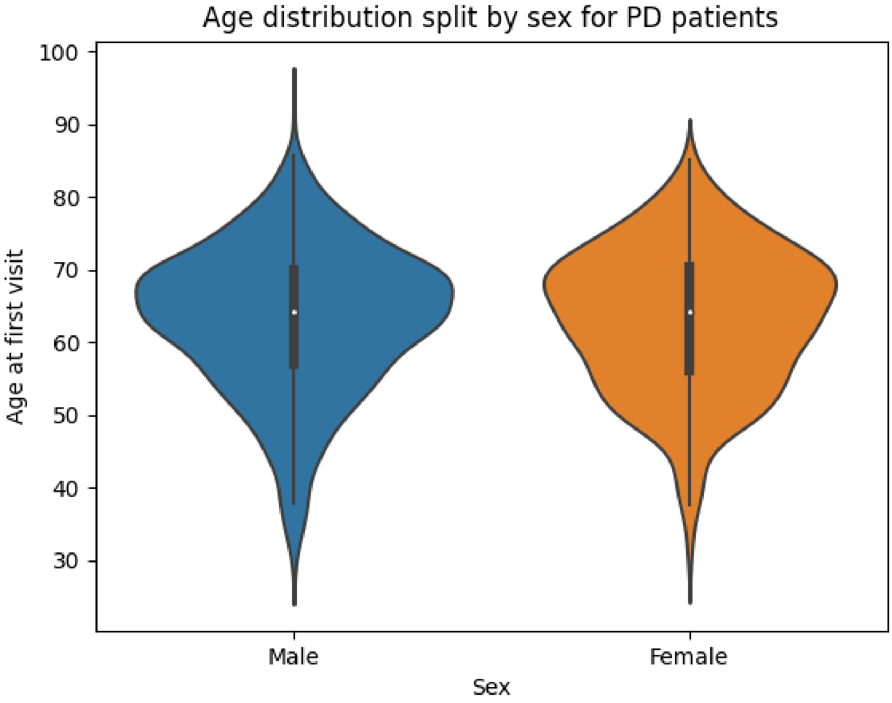


Figure 3.4: The age distribution for PD patients in the raw dataset.

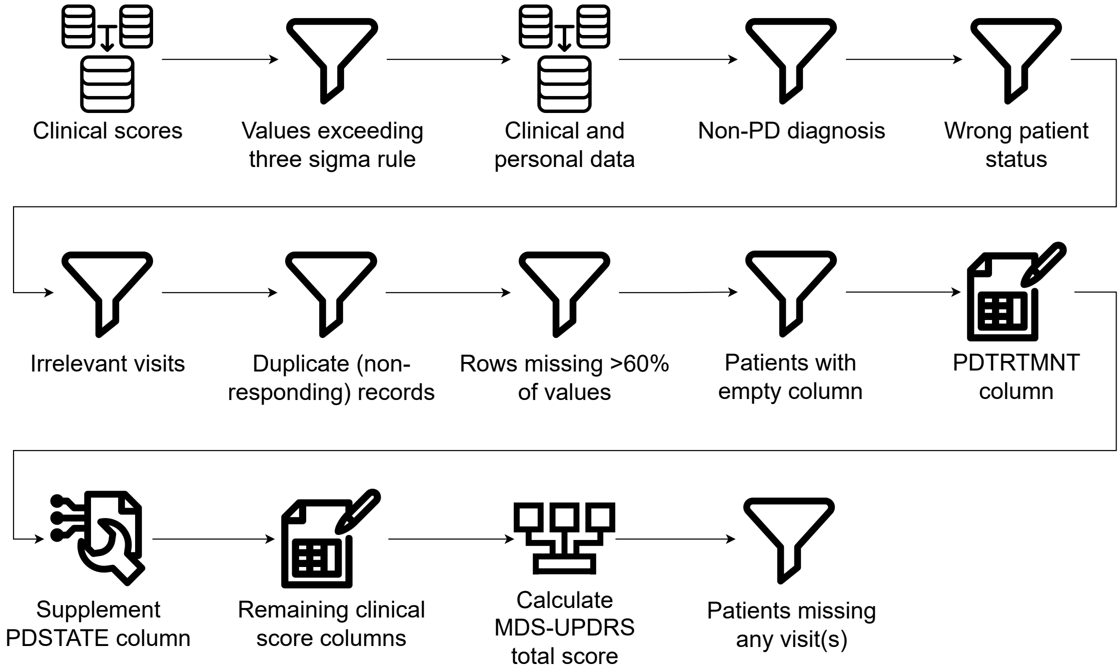

Clinical scores
Values exceeding three sigma rule
Clinical and personal data
Non-PD diagnosis
Wrong patient status
Irrelevant visits
Duplicate (non-responding) records
Rows missing >60% of values
Patients with empty column
PDTRTMNT column
Supplement PDSTATE column
Remaining clinical score columns
Calculate MDS-UPDRS total score
Patients missing any visit(s)

FIGURE 3.5: A flowchart of the pipeline's data pre-processing phase for the PPMI dataset. Icons taken from [47].

3.3.2 Data preparation

A high level overview of the data preparation phase of our pipeline is shown in figure 3.5.

Due to the structured nature of the PPMI project, we can start by merging horizontally on a combination of patient ID and event ID for the six assessment CSVs. After merging, five columns containing the exact visit dates are removed. This results in a dataframe containing 22749 rows and 82 columns, covering 2774 participants. Once this is done, we remove any incorrectly registered values that fall outside the possible scale for the included assessments. Furthermore, any statistical outliers in the MDS-UPDRS partition total score columns are removed by applying the three sigma rule (see [48]). This reduces the number of rows to 21810 and the number of participants to 2725.

The next step is merging the participant data with the clinical scores, which is followed by the removal of any participants that do not have a PD diagnosis or that do not have a status "enrolled", "withdrawn", or "complete". The two columns containing this information are then discarded. Once more, the structured nature of PPMI is used, this time to select visits: only the first six yearly visits including the baseline visit are retained as these should register the same assessments for each patient. These steps significantly reduce the size of the data, but add two columns. At this point 5812 rows and 84 columns remain, representing 1143 patients. Depending on the state of the patient during the assessment, which may be responding or non-responding to treatment, motor assessment scores especially may differ strongly. Because more patients have records where they are responding, and because some patients have two assessments linked to a single visit, one responding, one not responding, we choose to remove non-responding assessment records if a responding assessment record exists for the same visit. This leaves 4112 rows of data.

Following this, we move towards handling any missing data, starting with some cleanup based on missingness. First, any records that are missing over 60% of values are removed as these records contain only a single, if any, full MDS-UPDRS partition. Next, we remove any patients from the dataset that have a fully empty column across all visit records, because these patients cannot have this column imputed using forward/backward filling. Doing so removes another 880 rows, bringing the total down to 3232 rows and 716 patients. After this, we impute the column signifying whether a patient was undergoing treatment, and set their responsiveness according to this column. If a patient is not in treatment during a visit, their state is set to 2 instead of an empty value. Once this is done, the remaining missing values are imputed using the chosen imputation method. Finally, the MDS-UPDRS partition total scores are summed in a new column to give the prediction target MDS-UPDRS total score and patients that do not have all six visits are removed to form the final pre-processed dataset.

The full raw dataset (if we were to start by merging all CSVs) contains 26571 rows and 91 columns of data, covering 2774 participants. Once data preparation is completed, we are left with 1746 rows and 85 columns of data, covering 290 patients.

In Figure 3.6, the distribution of age and sex in the prepared dataset is shown. We see that in the male patient population, there is a fairly strong skewness towards higher age, whereas female patients have a fairly symmetrical age spread. Furthermore, we observe that the age range for the female patients is greater than for male patients, although this appears to be caused by a very small number of young female patients. When we compare this distribution to the non-processed age distribution for PD patients shown in Figure 3.4, it appears that a relatively large number of patients just below the age of 70 is removed during the data preparation process, especially in the female population. The cause for this phenomenon could be that the raw dataset includes patients that have only had a screening for PPMI eligibility, but no baseline visit yet, whereas the prepared dataset does not. It would be logical for a relatively large amount of screenings to be performed on patients around the age of 65 to 70, as this is the most common age at which PD is diagnosed [49]. Furthermore, we see that the age range for male patients shrinks quite strongly, with the maximum age reducing from nearly 100 years to just over 85 years. This decrease appears to be caused primarily by an extremely high age outlier in the raw dataset that is removed in the preparation process.

A number of examples of patient MDS-UPDRS total score values over time are shown in Figure 3.7. We can see here that the progression is very heterogeneous: some patients do not seem to worsen over these first five years, whereas others display a very clear upward trajectory in disease severity, such as patients 193 and 270. Aside from this, Figure 3.7 clearly visualises that progression may be highly erratic, as is the case with patients 40 and 274, for example.

3.3.3 Machine learning model

For the PPMI case study, three hyperparameters must be set: the batch size, the learning rate and the dropout rate. Because the prepared dataset in this case study is small, we choose for a relatively low batch size of 32, while leaving the model learning rate at the default 0.001. The dropout rate we choose for this case study is 0.2. As we have a small dataset in this case, it is likely that the model will overfit to the limited data it is trained on. By including the dropout at 0.2, we can reduce the degree to which the model (over)fits to the training data and increase the performance on unseen data.

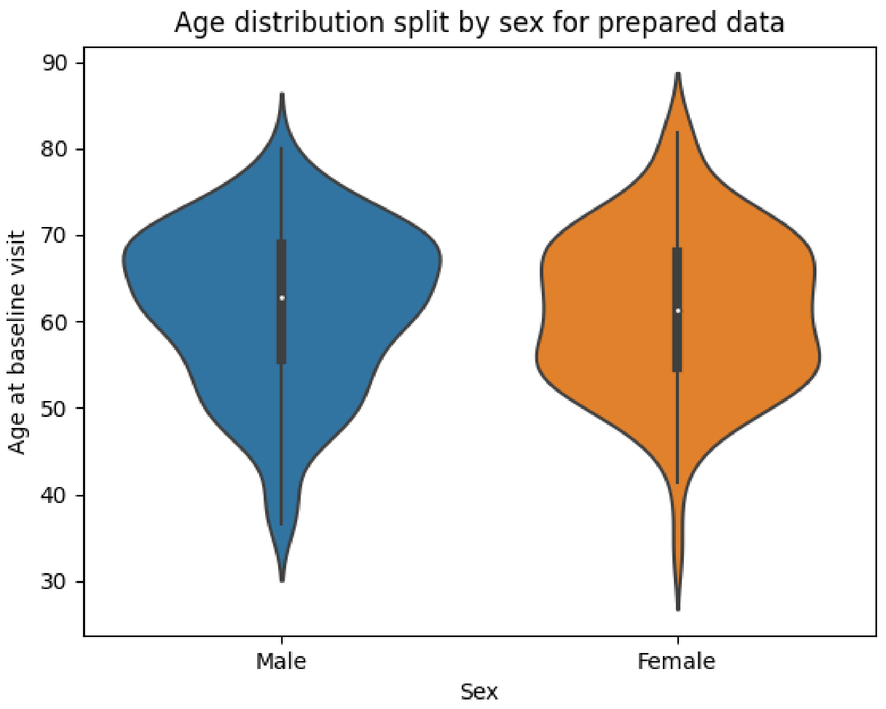


FIGURE 3.6: The distribution of patient sex and age at the baseline visit in the prepared dataset.

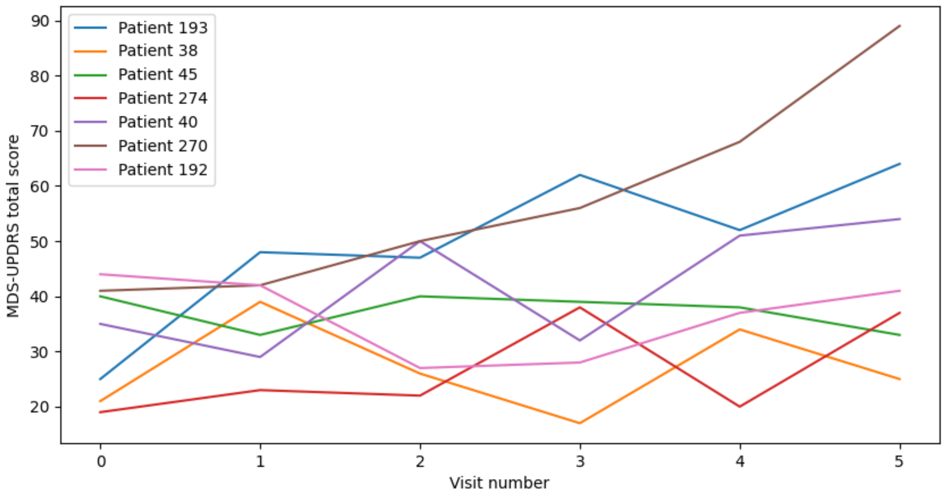


FIGURE 3.7: Several examples of patient MDS-UPDRS total score assessments over time.

3.4 Case study 2: diabetes

3.4.1 Dataset description

For the second case study, we use a dataset containing biomarkers from diabetes patients. This dataset is provided by Medlon, a medical laboratory based in Enschede, the Netherlands. To acquire this dataset, a data transfer agreement was signed. Furthermore, all personal information of patients is encrypted using a key to prevent any traceability and ensure anonimity of the patients before we received the data. The time period covered by the dataset is from the start of 2015 to the first half of 2023. To gather the data for this dataset, records in which a HbA1c measurement was ordered in the laboratory request were selected from primary care providers such as general practitioners. Since HbA1c is used in routine monitoring of diabetes patients, assessed every three to six months, the time interval between the records is generally homogeneous. As such, we assume once again that temporal irregularity for this dataset is limited, and we do not account for such irregularity between data points. The dataset consists of the biomarkers that are included in these laboratory request, such as the eGFR and blood glucose.

Due to the anonimisation of this dataset, less demographic information is available than in our other case study. The sex of patients, for example, is not included in the data. Another piece of information that is not included in the dataset, is whether patients underwent medical interventions between visits that may strongly influence the eGFR either positively or negatively. Hence, it may be the case that the dataset contains large differences between visits that cannot directly be linked to patient biomarker data. The dataset is constructed by taking information from a collection of CSV files that contain a certain timespan, rather than a part of the information. Each CSV contains a year of data for the CSVs covering the years 2015 to 2018. The data from the start of 2019 onward is stored in CSVs per six months.

The raw dataset on diabetes patients contains 128793 patients. Figure 3.8 shows the distribution of age for the full population in the dataset. We see a strong peak in the number of patients around the age of 70. The prevalence of high age patients suggests that our dataset contains predominantly diabetes type 2 patients. Type 2 diabetes is usually developed at a later age, as we mention in Chapter 2. In the same chapter, we note that around 95% of diabetes patients suffer from type 2, which supports our hypothesis on the distribution of diabetes types and the distribution of age in the dataset. In Figure 3.9, several random patient sequences are shown. It is clearly visible that the raw dataset is extremely sparse, with nearly half of the patients having no GFR-MDRD records at all, and only a single patient having more than two records. For this patient, with identifier 86663, we see an irregular trend in eGFR, with one larger jump between visits 3 and 4.

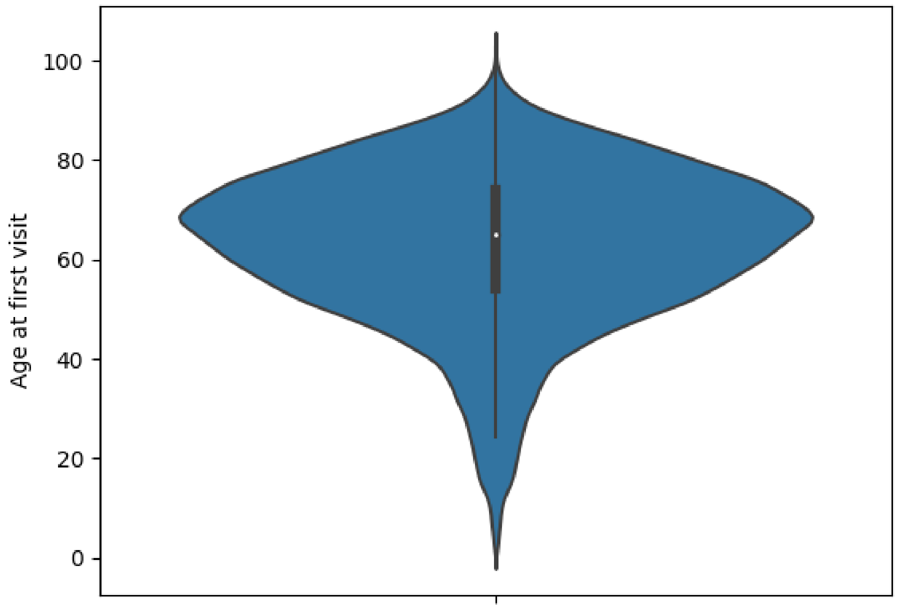


FIGURE 3.8: The distribution of patient age at the first visit in the raw dataset.

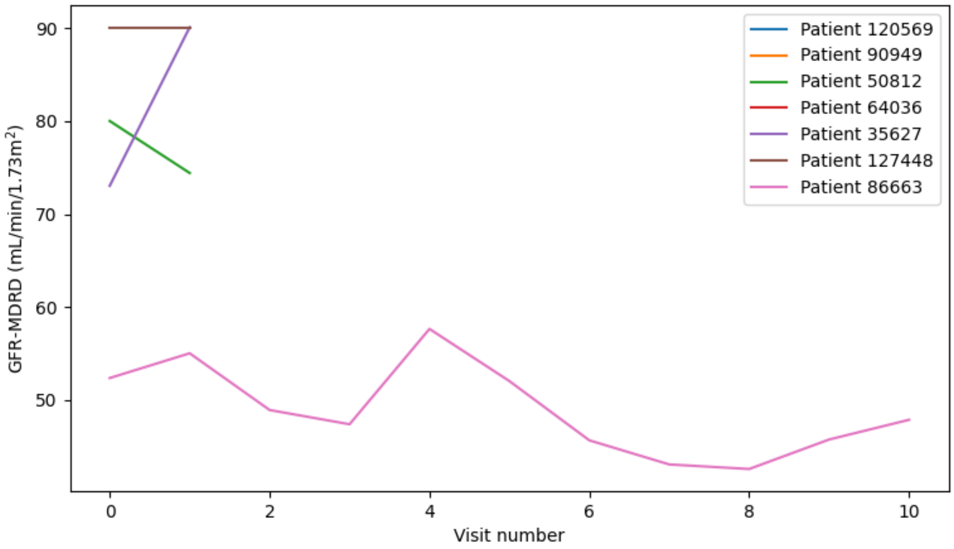


FIGURE 3.9: Several examples of patient eGFR assessments over time.

3.4.2 Data preparation

Similar to the previous case study, a high level overview of the data preparation pipeline we use is shown in figure 3.10. As the dataset for this case study is not fragmented over various files with different assessments, we do not have to merge CSVs horizontally. Since the dataset is split across files containing a year or half a year of data we start by stacking the CSVs so they form one chronologically ordered dataframe. By doing so, we create a dataframe consisting of 737010 rows and 861 columns. A total of 128793 patients are included in this dataframe. Next, we remove any columns that are not relevant for progression of diabetes or specifically the prediction of DKD. Examples of such columns include whether the patient is allergic to a specific fruit, and the patient's desipramine blood content. As some clinical assessments have been scattered across different across columns due to column name differences between CSV files, we merge these back into a single column per assessment. We also need to generate a unique patient ID by using two encrypted PIN columns and the encrypted date of birth of patients to identify them throughout the pipeline. These steps reduce the number of columns to 26.

Next, we remove any records where the target variable is missing and any patients that do not have at least 10 visits recorded in the data. Also, if multiple records were created on a single day, we discard all records on that day except the last one. Because this may have put some patients under the 10 visit threshold, we once again remove any patients that do not have at least 10 visits recorded in the data. This is done twice because of the high computational cost for discarding same-day records. This reduces the dataframe length from 737010 rows to 86185 rows, covering 6793 patients. Because we have no information on the age at each visit of patients, we calculate this using the date of birth and visit time. This age is stored in a new column and the date of birth column is dropped. Next, we remove any visits past the tenth visit for each patient, so we have equal length data sequences. This brings the number of rows down to 67930.

Next, we start handling missing data. First, we remove any columns that are missing over 60% of their values. This removes 12 highly sparse columns, bringing the total down to 14. Next, we remove patients that have a fully empty column across all visits, as we cannot perform forward/backward filling on such columns, and reset the patient IDs. Doing so reduces the number of rows and patients to 66440 and 6644 respectively. After this, we change the visit dates to visit numbers and finalise the preparation of the data by imputing any remaining missing values using the chosen imputation method.

The full raw dataset (created by stacking the CSVs) contains 737010 rows and 861 columns of data, covering 128793 patients. After completing the data preparation phase, we are left with 66440 rows and 14 columns of data, covering 6644 patients.

In Figure 3.11, we show the distribution of patient age after data preparation. Similar to the raw dataset, it is clearly visible that the distribution is skewed heavily towards elderly patients, with little observable change in the age group of 50 years and older. However, when looking at the lower age group, we see that the number of patients decreases drastically during data preparation. Under the age of 30, barely any patients are included. We could not find a clear cause for this change.

A number of examples of what the patient eGFR values look like over time are shown in Figure 3.12. Similar to the PPMI case study data, we see heterogeneity between patients and erratic trajectories. At the same time, we also see that the overall trend of most patients is fairly stable. Patient 4099 stands out in that regard, displaying a very strong decrease in eGFR. When comparing these sequences to the raw dataset example sequences, we see a large difference. Many patients in the raw dataset have either very few or no GFR-MDRD records, whereas the prepared data contains only patients with a full sequence of 10

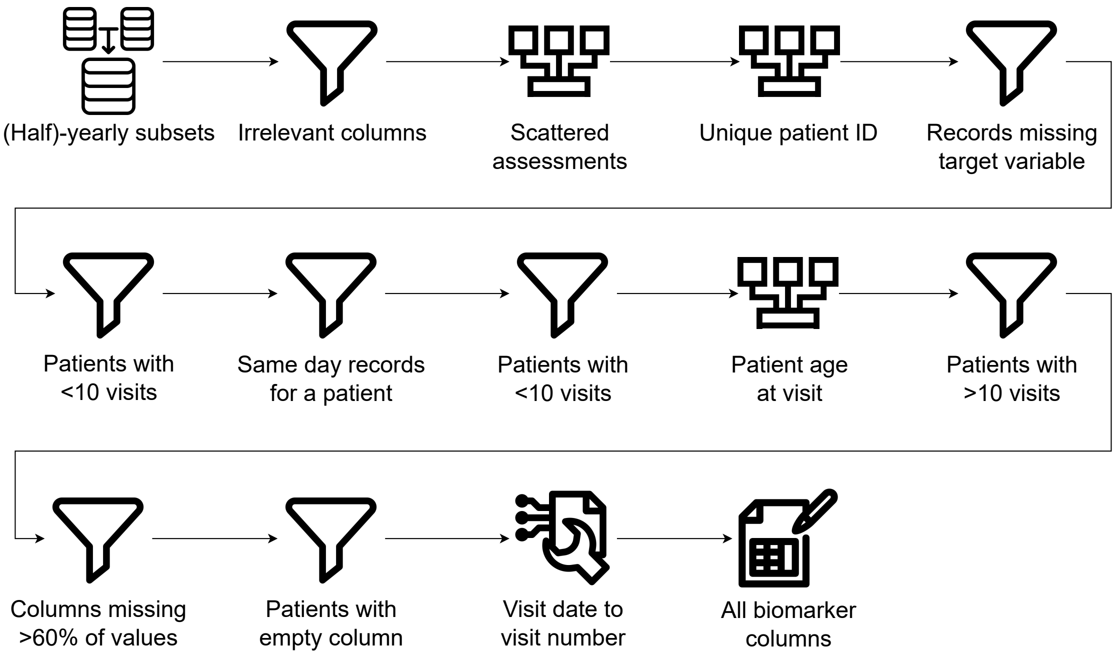


Figure 3.10: A flowchart of the pipeline's data pre-processing phase for the diabetes dataset. Icons taken from [47].

visits. We also see a larger range for the eGFR in the prepared dataset, but this is caused by the lack of representative patient sequences in the raw dataset sequence example as a result of data sparsity.

As the features for this case study are laboratory measurements rather than assessments that follow a set scale, as is the case for the PPMI dataset, we plot the distribution of each of the biomarkers included in the dataset. These are shown in Figure 3.13. We see that most patients have an eGFR around 90, which is the maximum possible value, and another large amount of patients around 60 to 70 mL/min/1.73m^2. Furthermore, we see that some biomarkers related to creatinine (Kreat, AlbKr and Alb_U_kw), have an extremely large range that appears to be caused by a small number of outliers.

3.4.3 Machine learning model

For the diabetes case study, we have the same three hyperparameters to optimise: batch size, learning rate and dropout rate. However, in this case study we have a much larger dataset. Due to this, we choose to increase the batch size 128, while decreasing the learning rate from the default 0.001 to 0.0005. In contrast to the PPMI case study, there is no issue with small datasets for this case study, so we do not account as strongly for overfitting to training data that might not be representative of unseen data. Hence, we choose to decrease the dropout rate by 50% to 0.1.

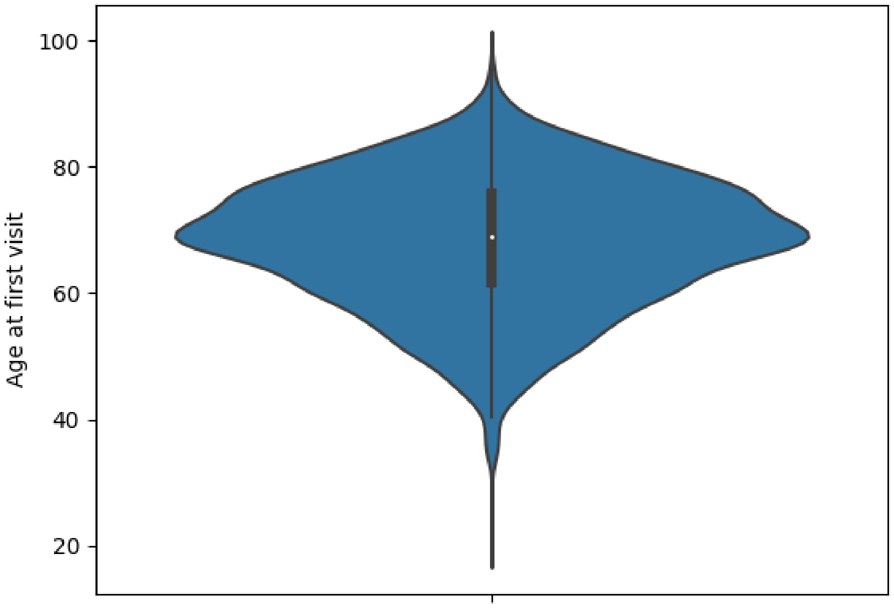


FIGURE 3.11: The distribution of patient age at the first visit.

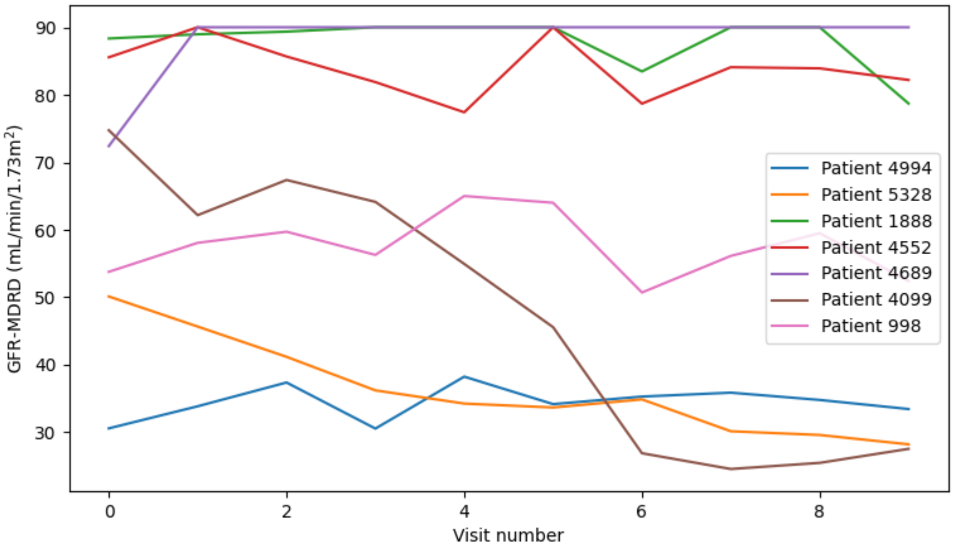


FIGURE 3.12: Several examples of patient eGFR assessments over time.

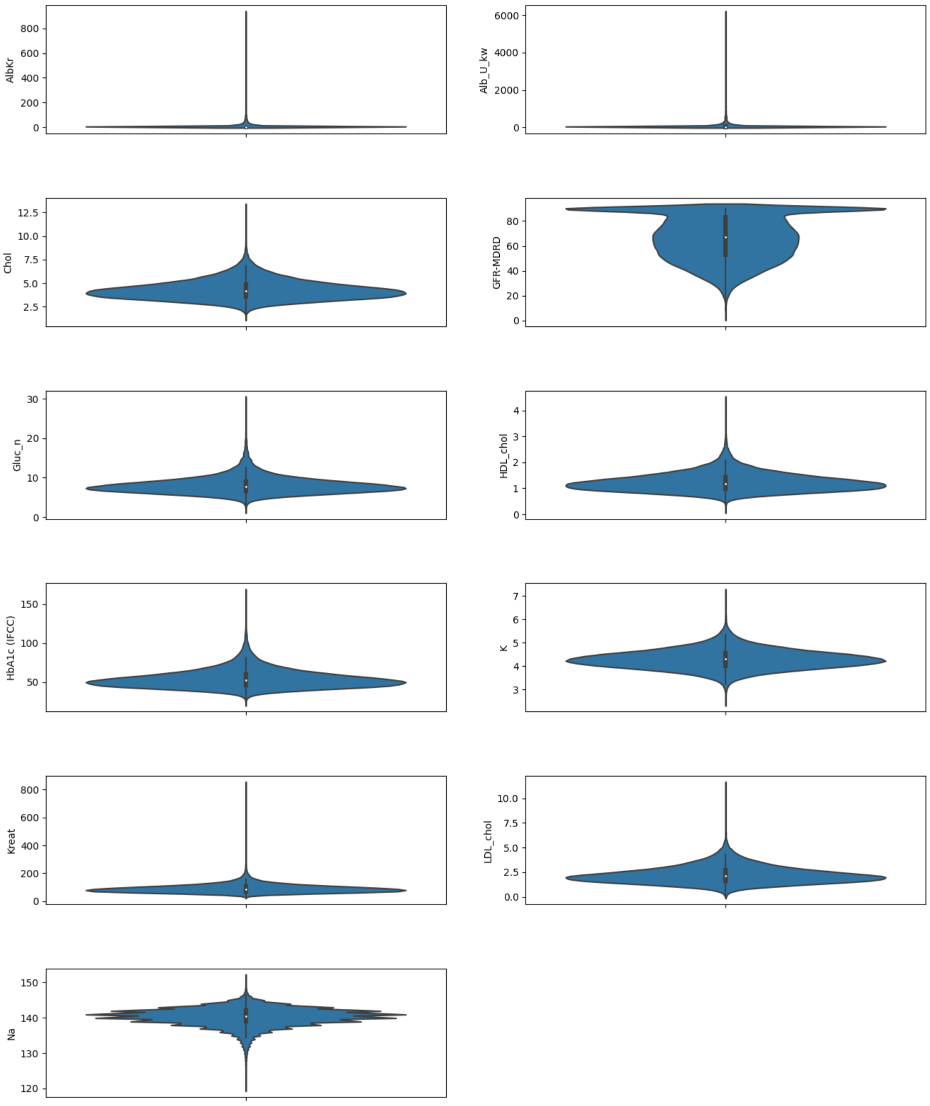


FIGURE 3.13: The distribution of the biomarkers included in the diabetes dataset.

Chapter 4

Results

This chapter is divided in two sections, each covering the results for one case study. Within these sections, we first present the results of the comparison between our three imputation methods without the explanation phase. These results are shown for four data splitting seeds and five cross-validated model versions. We do not discuss the predictive performance itself in depth, as development of a best-in-class predictive model is not the aim of this research. After the results of the imputation comparison, we show the output of the pipeline using the best performing imputation method for a number of example patients, including both the prediction and explanation phases.

4.1 Case study 1: PPMI

4.1.1 Imputation comparison

As mentioned in Chapter 3, we start by comparing three imputation methods. This experiment is performed on four different data split seeds. In the appendix, Table A.1 shows the full results of the experiment. Each seed represents a different data split, and as such, a different distribution of sequences across test and train partitions, as well as cross validation folds.

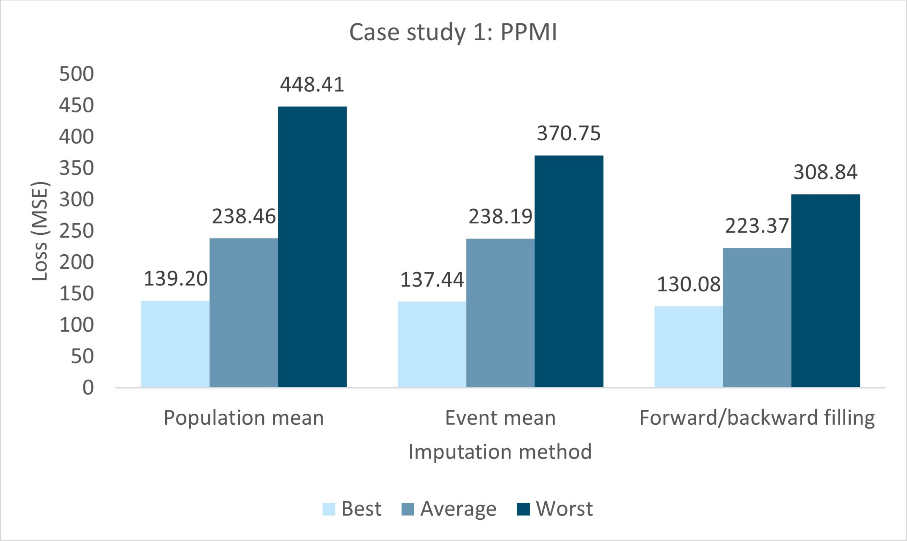


Figure 4.1: The predictive performance results for the PPMI dataset, taken across all cross-validations and data split seeds.

Figure 4.1 shows a direct comparison between the MSE scores across for each imputation method. Across all cross-validations and data split seeds, we show the best performing model, the worst performing model, and the average performance. We show the best and worst model performance to provide an indication of the variability in loss values. From the figure, it is clearly visible that the average loss of the two mean-based imputation methods is almost identical, with event mean imputation performing less than 0.01% better than population mean imputation. Forward/backward filling performs best on average, outperforming both other methods by 6.3%.

To better interpret the meaning of the results, the statistical significance of the difference in performance is verified. By performing a statistical significance test, we can validate whether the results provide conclusive evidence on which, if any, imputation technique leads to the best predictive performance. The significance test used is the Paired t-test, as the samples cannot be considered independent: the majority of the data in the dataset is the same for each trained model, regardless of the data split seed, the cross-validation folds, and the imputation method. For more information about the Paired t-test and how the test is performed, refer to Wilkerson [50]. The significance level for which we test in all comparisons is 0.05, the null hypothesis is that there is no difference in mean performance.

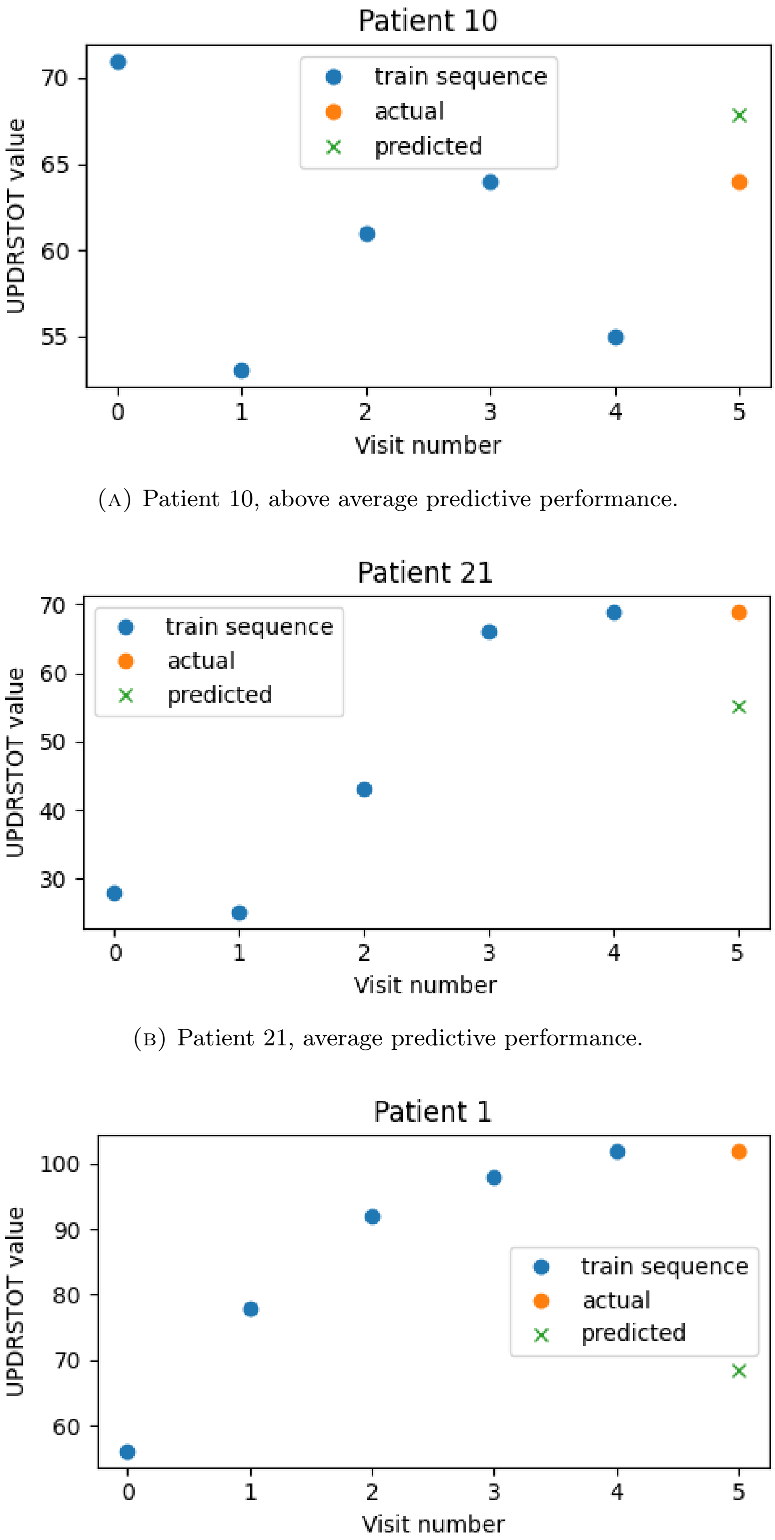
We first apply the Paired t-test to determine whether there is a significant difference using population mean imputation and event mean imputation. From the two samples containing 20 loss values per method, as shown in Table A.1, the two-tailed p-value is calculated to be 0.988. As this value is considerably larger than the significance level of 0.05, we cannot reject the hypothesis. Similarly, the comparison between the population mean technique versus foward/backward filling yields a two-tailed p-value of 0.293. While this p-value is substantially closer than the p-value for the previous comparison, it is still far from our significance level of 0.05, meaning rejection of the null hypothesis fails once more.

From these Paired t-tests, we conclude that neither event mean imputation or forward/backward filling offers a statistically significant performance increase over using population mean imputation. It should be noted that the lack of statistical significance of our observed difference does not mean that the choice of imputation method has no influence on the performance, but that from the results we obtained there is no conclusive evidence that the observed difference in performance is not caused by chance. For the remainder of this case study, we consider forward/backward filling to be the best performing imputation method and use this method for the generation of our further results.

4.1.2 Pipeline output

For the generation of the pipeline outputs in this subsection, we use the fold 1 model trained using seed 1998 because this model's loss approximates the average loss for the imputation method we use within 1%. As such, we assume the chosen model is representative of the pipeline's general performance. In Figure 4.2, we showcase three example predictions, along with the preceding sequence of the target variable and the ground truth for the predicted value. The first example, patient 10, is an instance of a relatively good prediction, the second (patient 21) an average prediction, and the third (patient 1) a relatively poor prediction.

When looking for an explanation for the predictions shown, the last phase of the pipeline is of interest. For the same patients we discussed above, the explanations offered by LIME are visible in Figures 4.3, 4.4, and 4.5. In each instance, the explanation consists of three elements. The leftmost element of the explanation, the "Predicted value" bar, shows the range of predictions, as well as the prediction for the instance that is being explained. The middle element contains a discretised view of the top 20 features, in descending order of



(A) Patient 10, above average predictive performance.

(B) Patient 21, average predictive performance.

(C) Patient 1, below average predictive performance.

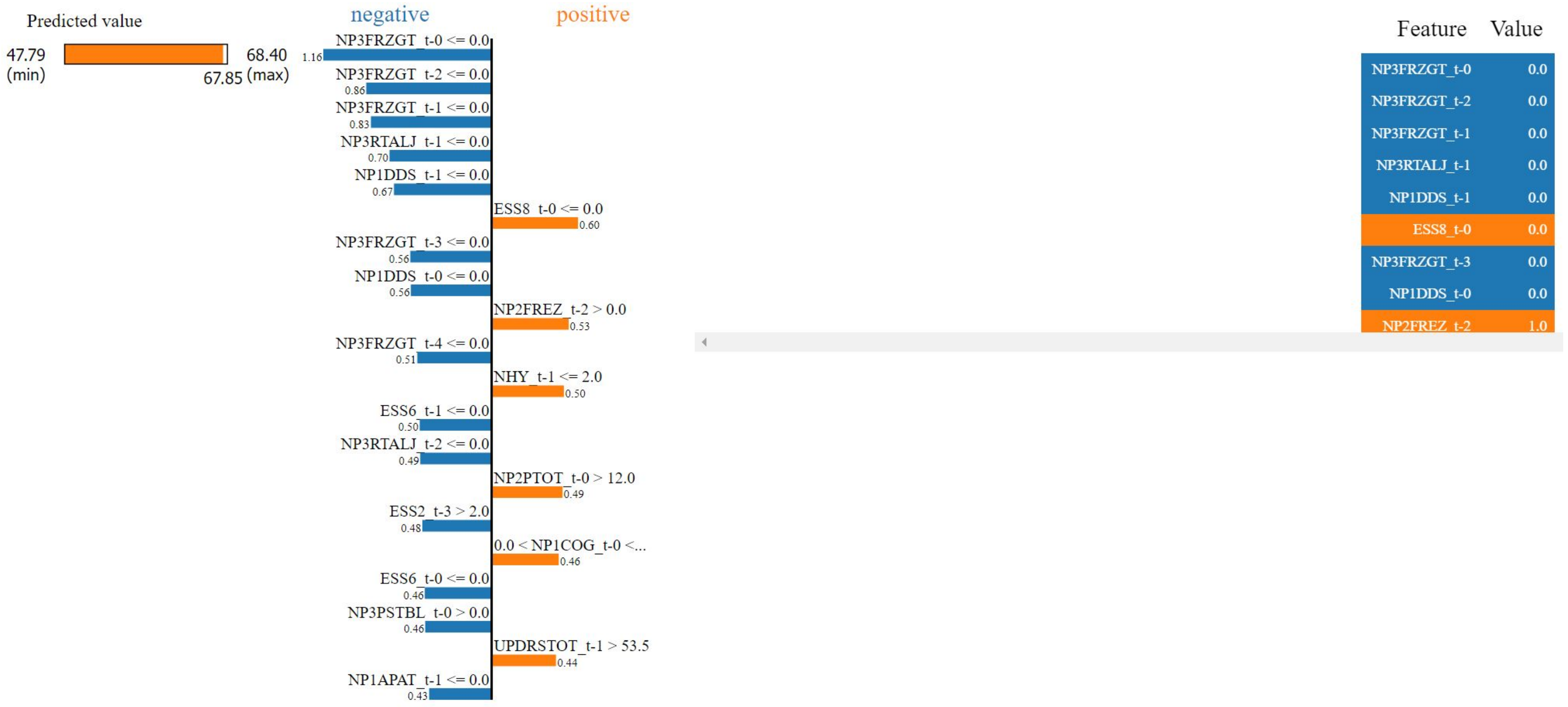
FIGURE 4.2: Three plots containing the MDS-UPDRS total score sequence of various patients in the PPMI dataset.

importance or influence. Negative influence (depicted in blue) means that the predicted value is decreased by the feature, positive influence (depicted in orange) means an increase in predicted value. Each feature name consists of an assessment and the timepoint at which the assessment was done. On the right, a table showing the top 20 features is included. This table contains the feature name, as well as the exact value of that feature at that timepoint, ranked by descending feature importance. In the environment in which the explanations are generated and presented, this table can be scrolled through to show all 20 features.

During analysis of the explanations provided by the pipeline for the PPMI case, we observe something peculiar. The explanations shown in Figures 4.3, 4.4, and 4.5 show rather small feature influences. Furthermore, looking at Figure 4.3 specifically, we see in the leftmost element that the predicted value is roughly the maximum possible prediction, but the five most important features in the middle element are shown to have a negative effect on the prediction. As such, the explanation points towards a relatively low prediction value, while the prediction value in reality is very high. It is possible that this is a consequence of the low differences in influence between the various features and timepoints. As the influence of each feature is small, it could be that outside the top 20 features shown, there are relatively many features that have a positive influence that bring the total prediction value up.

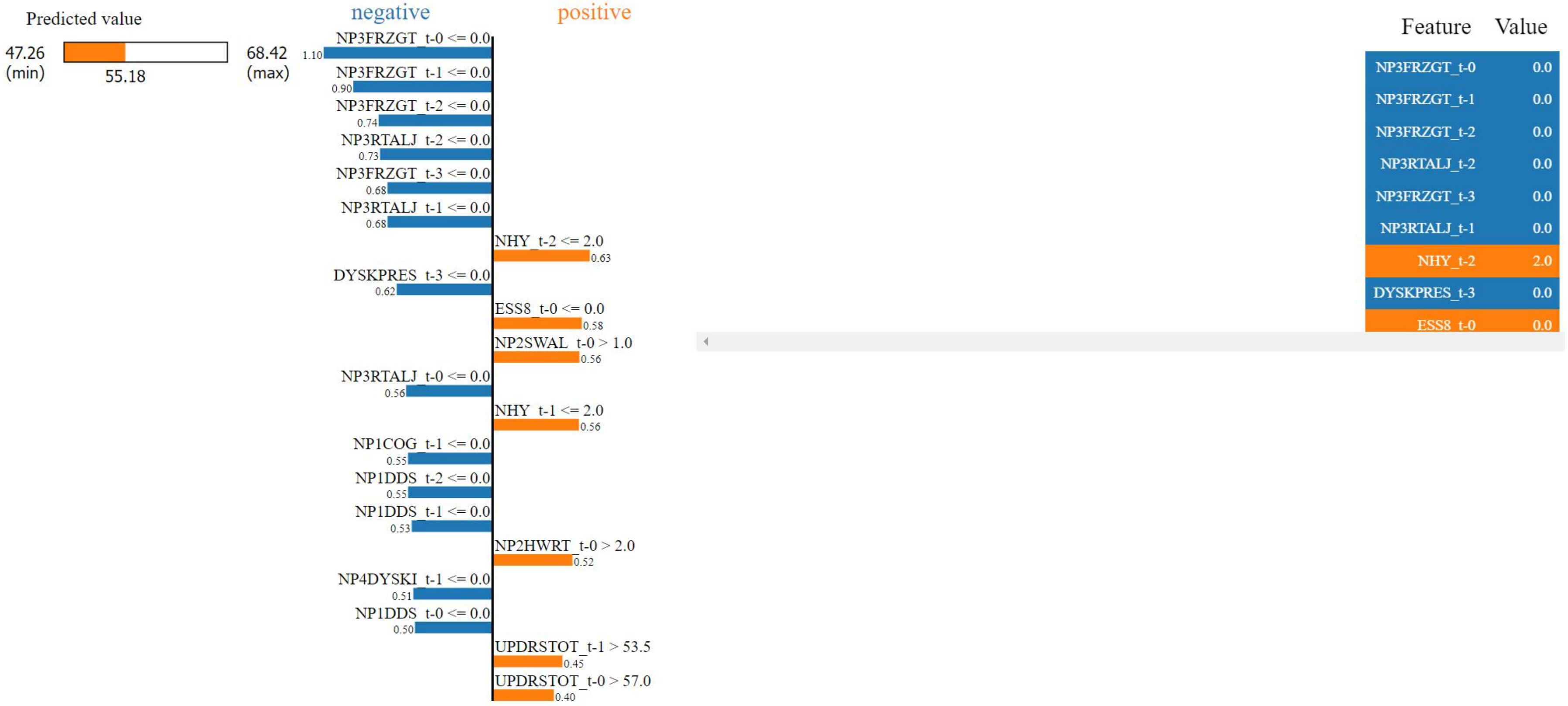
Another highly interesting observation is that for all examples, the earlier values of the target variable are not in any of the top 15 feature lists. Instead, we see a variety of highly specific (motor) scores that are assigned the highest influence. In the explanations for the prediction of both patients 10 and 21, depicted in Figures 4.3 and 4.4 respectively, the UPDRS total score is ranked nineteenth among the most important features. In the explanation for patient 1, shown in Figure 4.5, the UPDRS total score is not included in the top 20 features at all. This is striking, because we would expect the progressive nature of Parkinson's disease to lead to a strong influence of previous UPDRS total scores on the predictions. Looking at the correlation between various features in the prepared data and the UPDRS total score shown in Appendix B, we see that by far the strongest correlations exist between the UPDRS total score and the partition total scores NP2PTOT and NP3TOT. This is logical, as the UPDRS total score is a summation of all partition total scores. However, similar to the UPDRS total score, the partition total scores are barely, if at all, included in the explanations. Only Figure 4.5, shows the NP2PTOT score in the top 20 most important features, at rank eighteen.

Another remarkable finding from the explanations, is that the maximum MDS-UPDRS total score that can be predicted is reported to be just under 70. However, when looking at the sequence for patient 1 in Figure 4.2 plot (C), we see that this is not an adequate range. Patient 1 has only a single visit during which the score was below that boundary. Similarly, in Figure 4.2 plot (A), we see that patient 10 has a score above 70 for their baseline visit as well. As such, it appears that the predictive model is not able to account for the full range in which the target variable exists.



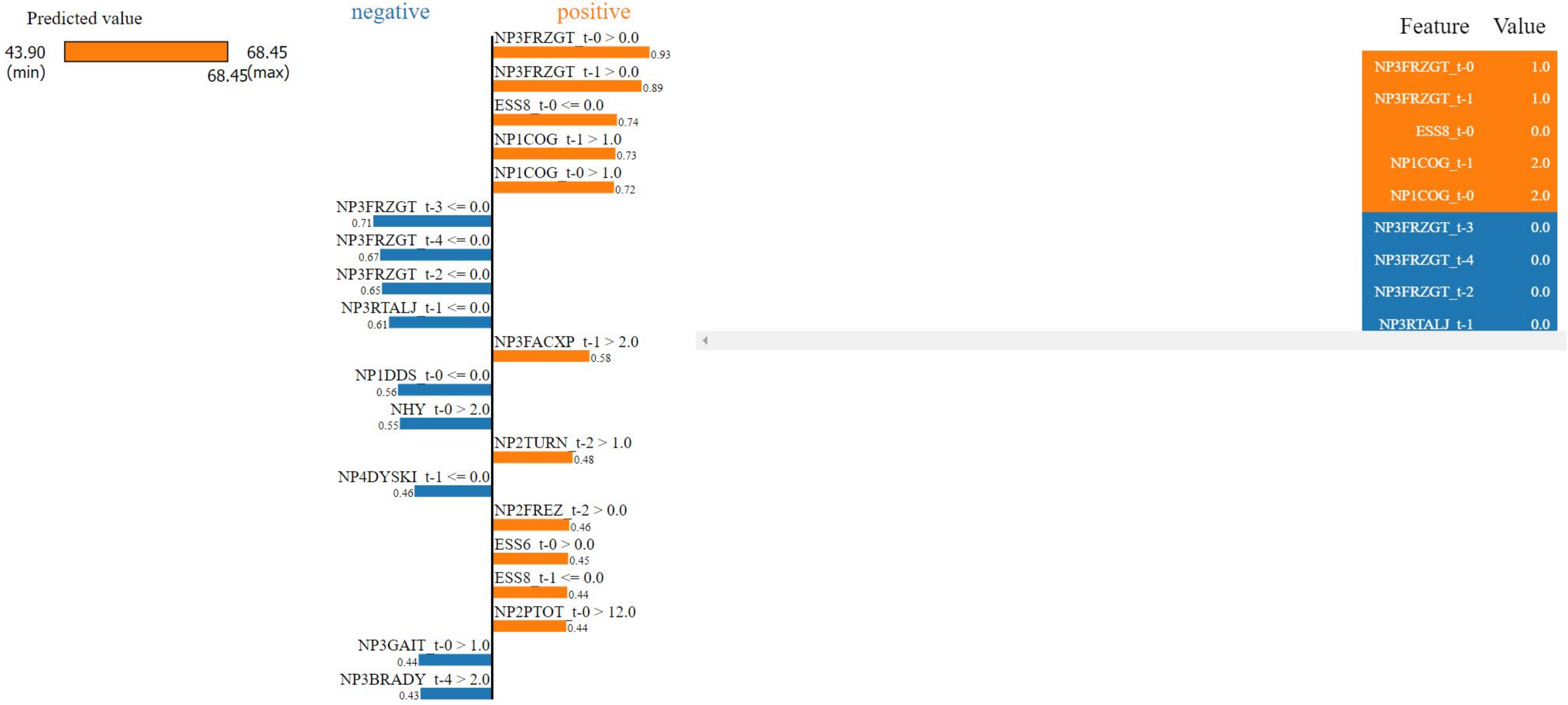
Feature	Value
NP3FRZGT_t-0	0.0
NP3FRZGT_t-2	0.0
NP3FRZGT_t-1	0.0
NP3RTALJ_t-1	0.0
NP1DDS_t-1	0.0
ESS8_t-0	0.0
NP3FRZGT_t-3	0.0
NP1DDS_t-0	0.0
NP2FREZ_t-2	1.0

FIGURE 4.3: The explanation offered by LIME for the progression prediction of patient 10.



Feature	Value
NP3FRZGT_t-0	0.0
NP3FRZGT_t-1	0.0
NP3FRZGT_t-2	0.0
NP3RTALJ_t-2	0.0
NP3FRZGT_t-3	0.0
NP3RTALJ_t-1	0.0
NHY_t-2	2.0
DYSKPRES_t-3	0.0
ESS8_t-0	0.0

FIGURE 4.4: The explanation offered by LIME for the progression prediction of patient 21.



Feature	Value
NP3FRZGT_t-0	1.0
NP3FRZGT_t-1	1.0
ESS8_t-0	0.0
NP1COG_t-1	2.0
NP1COG_t-0	2.0
NP3FRZGT_t-3	0.0
NP3FRZGT_t-4	0.0
NP3FRZGT_t-2	0.0
NP3RTALJ_t-1	0.0

FIGURE 4.5: The explanation offered by LIME for the progression prediction of patient 1.

4.2 Case study 2: diabetes

4.2.1 Imputation comparison

Again, we start by comparing our three imputation methods. Table A.2 in the appendix contains the full results of this comparison. Figure 4.6 shows a direct comparison between the best, worst and average performance across all cross-validations and data split seeds to provide an indication of the variability in loss values. Once more, it is clearly visible that the average loss of the two mean-based imputation methods is almost identical, with event mean imputation performing less than 0.01% better than population mean imputation. Forward/backward filling performs best on average, but the difference in performance for this method is less than 0.01% compared to the other two methods as well.

Similar to Subsection 4.1.1, we apply the Paired t-test to determine whether any of the observed performance differences are statistically significant. For the first comparison, population mean imputation versus event mean imputation, we obtain a two-tailed p-value of 0.924. Hence, the null hypothesis cannot be rejected at a significance level of 0.05. The second comparison, population mean imputation and forward/backward filling, gives a two-tailed p-value of 0.539. This exceeds the significance level of 0.05 and therefore fails to reject the hypothesis.

From these test results, we conclude that no tested imputation technique offers a statistically significant performance increase over population mean imputation. However, because forward/backward filling offers the lowest MSE in the comparison, we consider that technique to be the best performing imputation method and use this method for the generation of our further results in this section, as we did for the other case study.

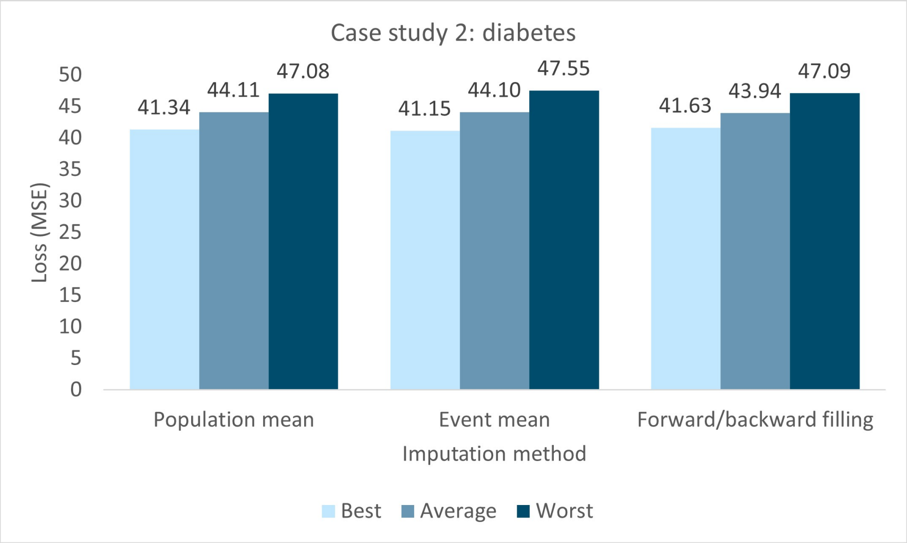


Figure 4.6: The predictive performance results for the diabetes dataset, taken across all cross-validations and data split seeds.

4.2.2 Pipeline output

For the generation of the pipeline outputs in this subsection, we use the fold 5 model trained using seed 25 because this model's loss approximates the average loss for the imputation method we use within 1%. As such, we assume the chosen model is representative of the pipeline's general performance. In Figure 4.7, we showcase three example predictions, along with the preceding sequence of the target variable and the ground truth for the predicted value. The first example, patient 6, is an instance of a relatively good prediction, the second (patient 11) an average prediction, and the third (patient 24) a relatively poor prediction. In plot (A) of the figure, it is visible that while the GFR-MDRD values have an erratic trend, there are no excessively large changes between visits. The biggest jump is around 10 mL/min/1.73m^2, between visits 1 and 2 (counting from 0). The predictive model predicts a slightly too high GFR-MDRD, but the error is less than 2 mL/min/1.73m^2, which indicates a prediction that is easily within the 15% (in this case roughly 6.9 mL/min/1.73m^2) biological margin that is used in clinical settings.

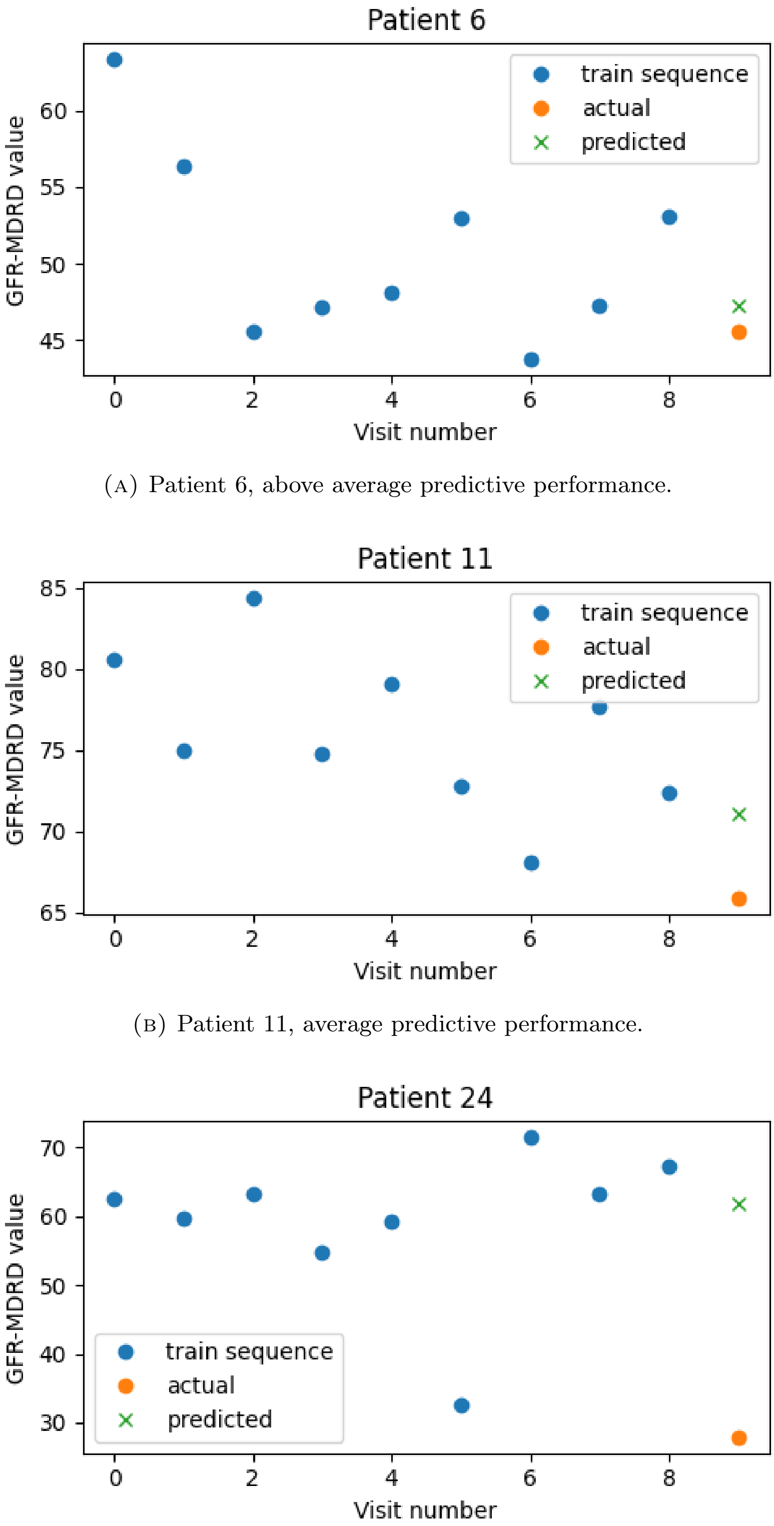
We again use the final phase of our pipeline to explain how our predictive model came to these predictions. The explanations for patients 6, 11, and 24 are shown in Figures 4.8, 4.9, and 4.10 respectively. Once more, the leftmost element of the explanation shows the range in which predictions can fall and the prediction value. The middle element contains a discretised view of the top 20 features and their influence on the prediction value, the rightmost element contains the exact value for each of the included features.

When looking at the explanations for the diabetes case study, we immediately notice that the previous values of the target variable are strongly represented in the explanations. For each explanation shown, many of the most recent eGFR values are highly influential on the final prediction. Mixed between these, however, we also observe values such as the Kreat (creatinine) biomarker and blood glucose exercising influence on the final prediction. Furthermore, we see that for some specific assessments, a feature that is multiple visits ago is included in the top 20 features. This suggests that the specific feature stands out amongst other assessments taken in the same visit, and that the predictive model recognises such data points as characteristic for the patient.

To validate the explanation results for the diabetes case study, the data owner was asked for their expert opinion on the performance of the proposed pipeline. Their first comment was that the degree to which the model can predict kidney function seems adequate, with 84.1% of all predictions being within the commonly used 15% biological margin around the ground truth eGFR. According to them, that level of performance could give the laboratory the possibility to start classifying patients in a low and high risk category for decline in kidney function. Based on such classification, low-risk patients could have a lowered monitoring frequency for kidney function in the future.

When asked specifically about the explanations for the three example patients, shown in Figures 4.8, 4.9, and 4.10, the domain expert mentioned that the explanations make sense in general. According to them, this is as expected because the available assessments are directly or indirectly related to the eGFR. Interestingly, the domain expert also mentioned that they expected the blood glucose to be represented among the most influential features, which they only observed for the above average performing prediction. As such, it is possible that this relation is not captured very well by the prediction model.

Another comment from the domain expert regarding the explanations, was that they would not have known how to interpret the explanations without the guidance provided during our meetings. We took this comment into consideration while identifying further research opportunities and found that this issue has also been raised by Dieber and Kirrane [40], albeit in a different problem context. This is discussed further in Chapter 5.

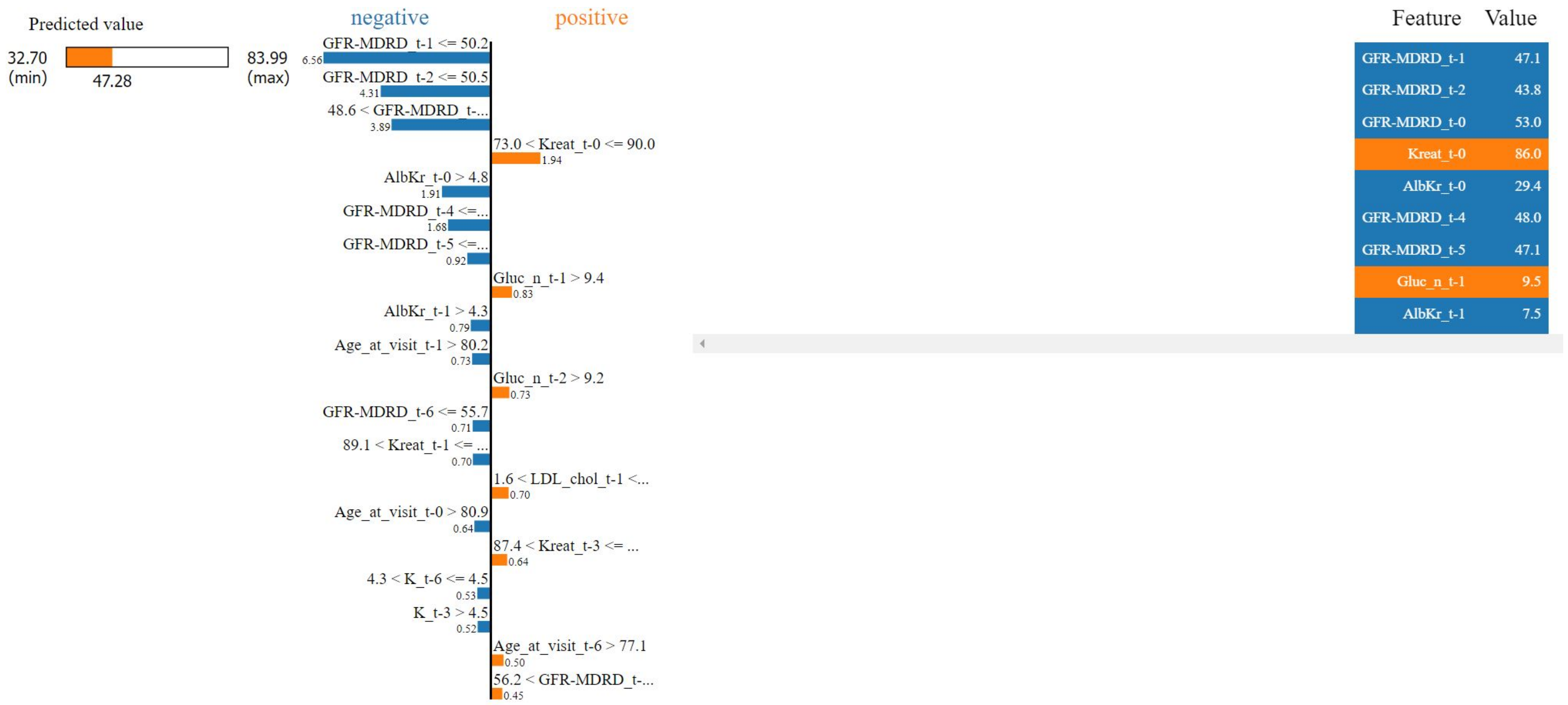


(A) Patient 6, above average predictive performance.

(B) Patient 11, average predictive performance.

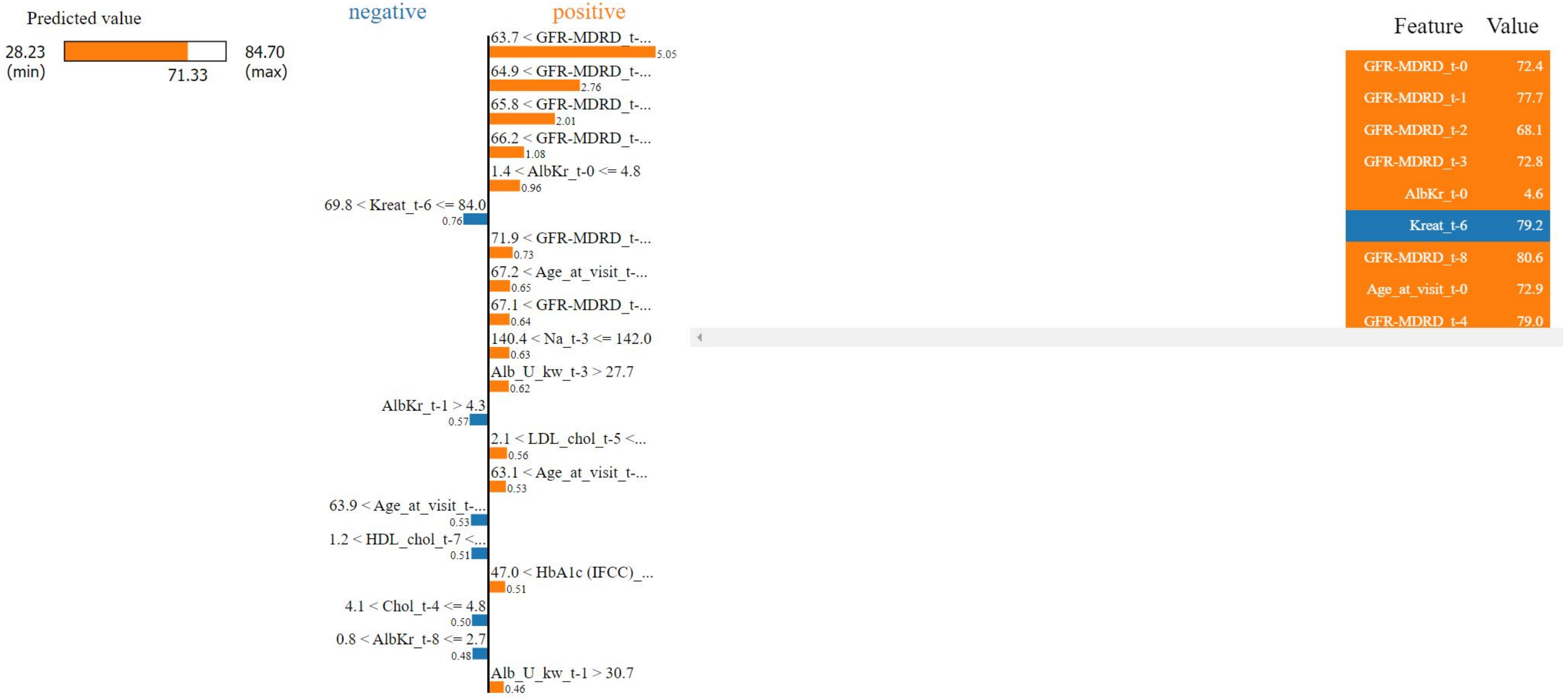
(C) Patient 24, below average predictive performance.

FIGURE 4.7: Three plots containing the GFR-MDRD (in mL/min/1.73m^2) sequence of various patients in the diabetes dataset.



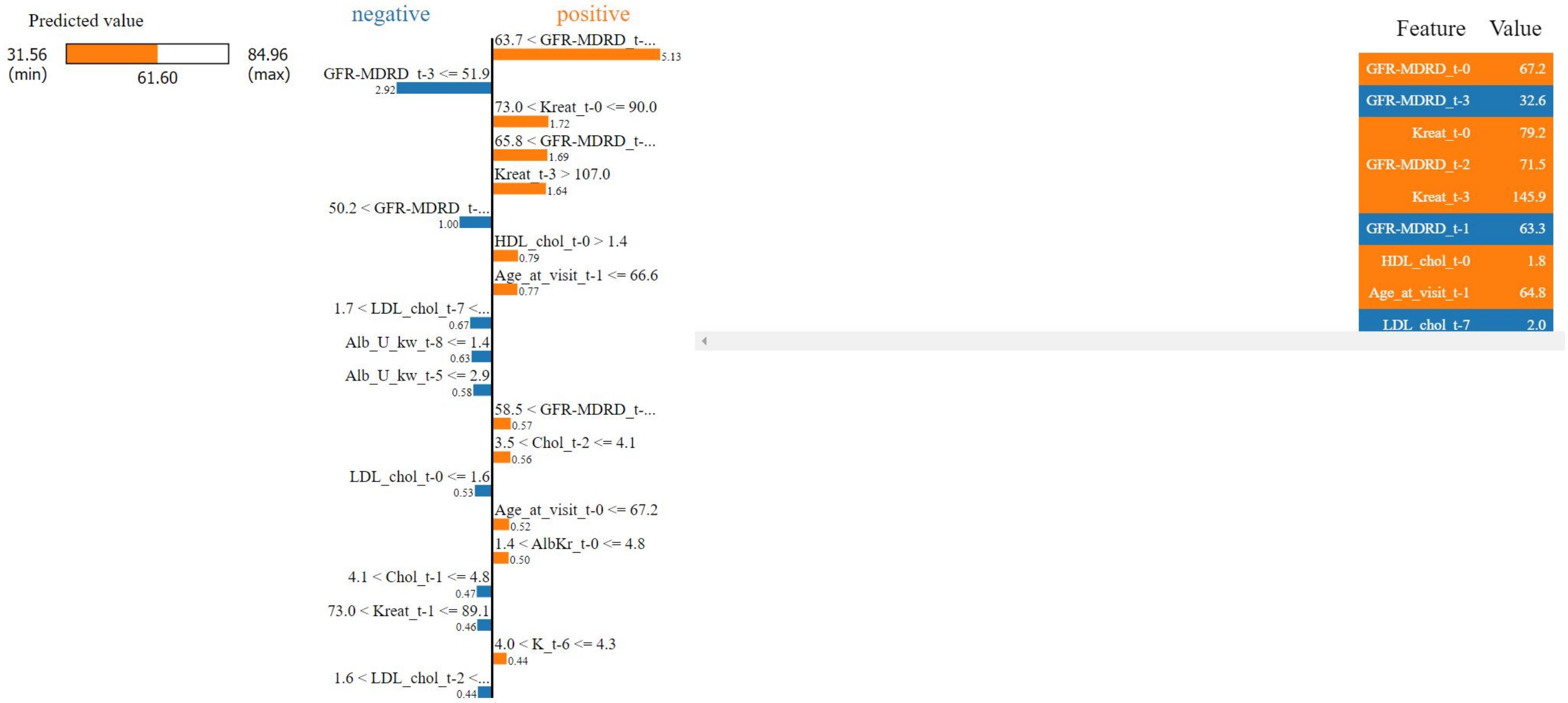
Feature	Value
GFR-MDRD_t-1	47.1
GFR-MDRD_t-2	43.8
GFR-MDRD_t-0	53.0
Kreat_t-0	86.0
AlbKr_t-0	29.4
GFR-MDRD_t-4	48.0
GFR-MDRD_t-5	47.1
Gluc_n_t-1	9.5
AlbKr_t-1	7.5

FIGURE 4.8: The explanation offered by LIME for the progression prediction of patient 6.



Feature	Value
GFR-MDRD_t-0	72.4
GFR-MDRD_t-1	77.7
GFR-MDRD_t-2	68.1
GFR-MDRD_t-3	72.8
AlbKr_t-0	4.6
Kreat_t-6	79.2
GFR-MDRD_t-8	80.6
Age_at_visit_t-0	72.9
GFR-MDRD_t-4	79.0

Figure 4.9: The explanation offered by LIME for the progression prediction of patient 11.



Feature	Value
GFR-MDRD_t-0	67.2
GFR-MDRD_t-3	32.6
Kreat_t-0	79.2
GFR-MDRD_t-2	71.5
Kreat_t-3	145.9
GFR-MDRD_t-1	63.3
HDL_chol_t-0	1.8
Age_at_visit_t-1	64.8
LDL_chol_t-7	2.0

FIGURE 4.10: The explanation offered by LIME for the progression prediction of patient 24.

Chapter 5

Discussion

In this chapter, first various limitations of the research are discussed. Subsequently, a number of recommendations for future research are provided.

5.1 Limitations

This section mentions various limitations of this research. The causes of these limitations range from assumptions and design choices that we make to technical feasibility. Furthermore, some limitations are identified based on the results discussed in Chapter 4.

5.1.1 Data imputation

In the literature review in Appendix C, we explore data quality issues in the field of prediction of disease progression. One of the issues encountered in the reviewed literature, which we also encounter in this research, is data sparsity. For both case study datasets, there are missing values across various rows and columns. The conclusion we draw in the literature review is that data imputation should preferably be avoided, as one of the drawbacks of imputation is that possibly non-representative values are injected in the dataset, leading to reduced predictive performance of models trained on the dataset. Alternatives for imputation that we identify in the review are removal of records containing missing values, and using a ML model that is robust to missing values. In this research, we do not use either of these alternatives, instead opting to use data imputation regardless of the known drawbacks associated with it.

The reason why we do not use record removal as an alternative is because we aim to create a pipeline that is applicable for a wide range of cases. To achieve this, we need to account for varying sizes in available training data. The first case study, using the PPMI dataset, reflects a case where a very small amount of data is available. This small dataset, combined with the fact that each patient was missing at least one value across all their visits and that we require patients included in the processed data to have a registration for each visit, means that removing records containing missing values leads to an empty dataset. As such, this is no applicable alternative for this research.

The second alternative, using a ML model that is robust to missing values, was attempted initially. However, only a single ML model presented in the reviewed literature fits the criteria we have set (time-series input data, continuous prediction value, RNN-based model) whilst also being robust to sparse input data: a GRU-based model proposed by

De Brouwer et al. [51]. Furthermore, the highly complex nature of their model means that integration between the predictive model and the Python implementation of the LIME framework is unfeasible.

For these reasons, there are no alternatives for imputing the missing data in this research. Hence, we were forced to use imputation as a solution to the missing data problem in our ML pipeline. Comparing various imputation methods, performing cross-validation, and repeating the experiments for multiple seeds is done because we want to mitigate the drawbacks of the imputation approach as much as possible.

5.1.2 Data imputation effects on explanations

A limitation that we must note, which is directly related to the limitation discussed above, is the influence of data imputation on the pipeline phase after predicting the disease progression. It is likely that negative effects of imputation extend into the generated explanation fidelity when assessed against the ground truth explanation. Since LIME generates explanations based on fitting a locally representative model, this locally fitted model will suffer from the same negative influence of data imputation as the pipeline's main predictive model. As such, features containing imputed values may stand out significantly from their non-imputed counterparts and be flagged as important or interesting because they are not representative of the missing value, rather than because the feature truly is important.

At the same time, we note that this does not necessarily mean such an explanation is incorrect. In fact, if LIME emphasises a feature as a result of an imputed value, the real-world feature importance is not reflected, but the explanation does reflect the feature importance within the model. This distinction is noteworthy, because it highlights the ability of the explanation phase to aid in finding shortcomings of the predictive phase.

5.1.3 Imputation comparison

A very different limitation caused by data imputation in this research is a result of the choice of imputation methods. As mentioned in Chapter 3, we use three imputation methods. Two of these methods rely on "horizontal" data for calculating the value used to impute a missing value with: the population mean and event mean techniques. Both of these can fill a missing value as long as there is at least one other patient record for the column that must be filled, either across the entire dataset population or across the population for a specific event respectively. This is very likely to be the case, as both datasets contain a population in the hundreds or even thousands. However, the third technique, forward and backward filling, relies on "vertical" data. Instead of having hundreds or even thousands of records that can be used as reference for imputation, only a number of records equal to the length of the sequence that we are performing imputation on can be used. This number is orders of magnitude smaller, for the PPMI and diabetes case studies it is 6 and 10 respectively. As such, the likelihood of a sequence having a column for which imputation cannot be performed due to there being no reference values increases drastically.

In such cases, removing the column is not the desired solution. This is because it is likely that a large majority of other sequences do not have this issue, and removing the column would have us actively remove a lot more information than simply removing the problematic sequence. Hence, we remove such sequences to retain as much information as possible while still solving the imputation problem. Even though this methodology solves the problems with the imputation step itself, it can be considered undesirable for our comparison of imputation methods. After all, removing any sequences with empty columns to accommodate forward and backward filling effectively "hobbles" the other imputation techniques. These techniques could include the removed sequences without issue and possibly use this additional information to improve predictive performance. As such, we considered doing an additional comparison between using all sequences and using the sequences suited to forward and backward filling, for each mean-based imputation method. However, we decided against performing this additional comparison.

We justify this decision by pointing out the way data splits and cross-validation folds are constructed. While we can fix the distribution of data across the splits and folds using seeds to compare various techniques if the input data is the exact same, even adding only a single sample results in the distribution being completely different regardless of the seed due to the input data being different. As such, there is no way of fairly comparing only the effect of having additional sequences, because there is also an effect on the performance caused by the different distribution of data across the various splits and folds.

5.1.4 Data dimensionality

As mentioned in Chapter 4, in the context of the PPMI case study the explanation results do not correlate with our expectations, whereas the explanation results for the diabetes case study do. We hypothesise that the observed phenomena in the PPMI case have the same underlying cause. We note the following: Rad et al. [52] state that as the dimensionality of data increases, the likelihood of feature values incidentally being correlated with events or anomalies increases as well. They explicitly add that for time-series data, high dimensionality is even more problematic. As time-series data does not only contain x amount of features, but also y timepoints at which each feature exists, the "effective" dimensionality of the data increases to $x * y$. As such, the likelihood of incidental correlation is high. If this knowledge is applied to the PPMI dataset, we find that the dataset effectively contains 483 features over all timepoints (83 feature columns * 6 visits). That number is almost twice the number of sequences available for training the pipeline, which is 247 after removing the test set.

Considering the extremely high effective dimensionality of the PPMI dataset, it is then logical that the predictive performance in this case study is relatively lower than the predictive performance for the diabetes case study. In turn, we expect the explanations to reflect a poor prediction model. However, explanations reflecting a subpar prediction model may not be the only issue resulting from the dimensionality problem. Because the LIME framework is based on local approximation, as discussed in Chapter 2, the high dimensionality of the data also leads to poor fitment of the local approximation in the LIME framework. Molnar [53] mentions that this dimensionality problem is exacerbated by the calculation of the kernel width in the Python implementation of LIME. The kernel width is what determines how close perturbed instances must be to influence the model. The larger the kernel, the farther away an instance can be while still exercising influence on the approximation model. Molnar explicitly mentions that the Python implementation of LIME calculates the kernel width using Equation 5.1.

$$kernel\_width = 0.75 * sqrt(n_{columns}) \tag{5.1}$$

If we calculate the kernel width for the diabetes dataset and the PPMI dataset using the same equation, we see that the diabetes dataset has a kernel size of 8, as shown in Equation 5.2.

$$kernel\_width = 0.75 * sqrt(120) \approx 8 \tag{5.2}$$

whereas the PPMI dataset has a kernel size of 16, as shown in Equation 5.3.

$$kernel\_width = 0.75 * sqrt(483) \approx 16 \tag{5.3}$$

Thus, even though the PPMI dataset contains 23 times less data sequences, the kernel size for this dataset is twice the kernel size for the diabetes dataset. This oversized kernel likely causes too many perturbed instances to have an impact on the local approximation, leading to the poor explanations observed in the results for the PPMI case study.

5.1.5 Generalisability of the pipeline

A high degree of generalisability is a desirable feature for most systems, and this is no different for our pipeline. A generalisable pipeline means the pipeline is applicable to a wide range of diseases. To assess the degree to which our proposed pipeline is generalisable, we use a train-test split as well as cross-validation. By assessing how variable the predictive performance is across these splits and cross-validations, we can get an indication of whether the pipeline performs well on unseen data. There are two primary reasons why out-of-sample performance may be subpar. The first is that the pipeline is simply not generalisable, and only suits a specific case due to the architecture of the pipeline. The second is that a data distribution problem is present, leading to the model not being trained on data that is representative of the population.

In the PPMI case study, we see extreme variability in the predictive performance of the pipeline as shown in Table A.1. Across different seeds for splitting the data as well as cross-validation model variants, there is a very high standard deviation in the MSE. In the worst case for example, seed 1998, we see a standard deviation of 37.4% of the average MSE when using population mean imputation. For seed 787 on the other hand, the same imputation method results in a much lower standard deviation of 9.0% of the average MSE. This suggests that the data distribution over the splits and folds has a very strong effect on the predictive performance.

One of the most logical explanations for the high variability in this context would be the size of the PPMI dataset. If we look at the difference in size between the diabetes and PPMI datasets, we see that once processed, the diabetes dataset contains nearly 6700 patients, whereas PPMI only contains 290. As the train-test split is 85-15%, this only gives us a test set of 43 patients. If this very small sample contains a disproportionate amount of patients that are more difficult to predict for than average, for example due to them having a lot of imputed (and thus possibly not representative) values or biologically uncommon jumps in disease severity or some biomarker, the performance on this test set will be very poor. The same principle applies to the distribution of patients across cross-validation folds. To further solidify this explanation, the standard deviation across cross-validation within the diabetes dataset is significantly lower than the PPMI dataset, between 3.2% (seed 959143, population mean imputation) and 1.2% of the MSE (seed 1998, population mean imputation). The larger dataset size appears to correlate with a lower degree of variability.

5.1.6 Temporal irregularity assumptions

A limitation that is caused by an assumption we make, is that we do not actively account for temporal irregularity in this research. In this research, we note that the problem of temporal irregularity exists in (healthcare) time-series data, but we do not actively align the time of data points. As briefly discussed in Chapter 3, our assumption is not without reason.

For our first case study, PPMI, we justify the assumption that any temporal irregularity does not require intervention because of the structured planning that is the backbone of the PPMI study. The visits we use in our research are yearly, and each visit is registered with the appropriate event identifier. As such, even if patient A has their second visit exactly a year after their first visit, and patient B has their second visit 13 months after their first visit, the influence of this different interval on the captured trend of the progression should be minimal.

For our second case study, diabetes, there is no such strict planning. However, according to the domain expert for this case, we can base our assumption on the Nederlands Huisartsen Genootschap (NHG) *(English: Dutch General Practitioners Association)* guidelines for monitoring diabetes patients [54]. Because the data in the dataset is generated by primary care providers such as general practitioners, this guideline is followed. Hence, the data does not include (parts of) sequences which are recorded within only a few days as a result of constant monitoring in the hospital. The reason for the request of the assessments is important for justifying our temporal irregularity assumption too, because the dataset mostly contains visits where the primary goal of the visit was monitoring the HbA1c biomarker, as per the NHG guideline mentioned above. As the HbA1c biomarker is an indication of blood glucose levels over the past three months, this is the interval that the guideline suggests between visits. Thus, it is likely that all measurements are taken at such intervals, along with the HbA1c measurement. As a result, the likelihood of semi-consistent intervals is high enough to assume minimal impact of any irregularity across the entire dataset population.

5.1.7 Validating explanations

The final limitation we need to mention is that we do not validate our explanations using a domain expert for our first case study. As the data for our second case study is provided to us directly by Medlon, we have a line of communication with a domain expert on the diabetes dataset. This domain expert could share their expertise on the dataset and their opinion on the predictions and explanations over the course of a number of meetings. We used these meetings as our validation for the explanations for this case study.

For the PPMI case study, we acquired our dataset using a web portal. Thus, there was no contact with a domain expert during this process or after. Due to time constraints, we did not manage to get in contact with a domain expert on the PPMI study. As such, the only validation that we can perform for this case study is for predictive performance. As mentioned earlier in this chapter, however, it is unlikely that the explanations for the PPMI case study are representative due to the dimensionality issue with the dataset.

5.2 Recommendations for future research

Our recommendations for future research focus on improving two parts of the pipeline: the predictive model and the explanations. These are discussed separately.

5.2.1 Predictive model

The predictive model we use in this research is constructed using existing, state-of-the-art components. Because some of these components are - as of yet - not widely used in the context relevant for this research, the flexibility of some implementations is not very high. All of the recommendations below are related to a lack of presence of data, so it is possible that all of them need to be addressed simultaneously from an implementation perspective.

An example of flexibility that would be beneficial for the performance of the pipeline as a whole, is the ability for the predictive model to handle ragged tensors. A set of ragged tensors is essentially a set of data sequences of varying length. For example, three patients that have six, five and six visits respectively. As currently implemented, the model in the pipeline cannot handle this and would require the patients with six visits be pruned back to five visits, or for the patient with five visits to be removed, leaving only two patients. In either case, information is lost as a result of inflexibility of the model. Hence, we recommend that the flexibility of the model's input is to be improved.

While the form of flexibility described above can be seen as revolving around missing visits at the *end* of a sequence, the next flexibility improvement concerns missing visits *within* a sequence. Such flexibility works differently for the two case studies in this research: in the PPMI dataset, visit records are explicitly connected to a planning. As such, a patient's second visit may be labeled "visit 3" rather than "visit 2", because the visit takes place at the intended time for "visit 3". The record "visit 2" remains non-existent for that patient. In the diabetes dataset, there is no such planning. Each visit simply follows up on the last registered visit and will be numbered as the direct followup. However, if the model is capable of handling a missing visit, flexibility regarding such instances is increased dramatically for both case studies. In the PPMI case, a patient visit could simply be considered missing if a later visit is present for the patient. In the diabetes case, it could be possible to use the time interval between visits to determine whether a visit inbetween two visits is "logically" missing. For either case, such flexibility would improve the amount and correctness of information captured in the dataset, subsequently improving predictive performance.

Finally, we note the lack of flexibility regarding missing values within a single record. The current implementation of the pipeline requires the input data to be complete, without a single missing value. As discussed earlier as a limitation in Chapter 5, due to technical feasibility, we were forced to use imputation regardless of the known associated drawbacks. As such, we recommend researching whether the predictive model could be made robust to missing values without imputation, and comparing whether this leads to improved predictive performance.

5.2.2 Explanations

There are two aspects to the improvement of the explanations given by our pipeline. The first is the fidelity or correctness of the explanations when compared to the ground truth, the second is in the presentation of the explanations.

As mentioned in an earlier part of this chapter, we find that for the PPMI dataset, there is most likely a dimensionality problem leading to very poor explanations. While this problem is not explicitly caused by the explanation phase, one of the possible solutions may lie there. In the explanation of the problem, we mentioned the calculation of kernel width for the Python LIME implementation. This calculation might be a solution to the high dimensionality problem for LIME. If the calculation for the kernel width is changed so that it remains lower for high dimensional data, the explanations for this data may better reflect the true inner workings of a prediction model. We recommend that the influence of changing the kernel width calculation is tested in further research.

In our pipeline, we use the LIME framework to explain our predictions. LIME is an established framework, and is used as a benchmark for other explainability frameworks on a number of occasions as mentioned by Dieber and Kirrane [40]. However, they also mention that LIME is not properly benchmarked itself. Dieber and Kirrane perform an evaluation of LIME from a usability perspective, but they do not test the fidelity of the framework. We define fidelity as the degree to which the explanation for a prediction generated by LIME matches the ground truth of why the prediction should be what it is. Because there is very little, if any testing done to confirm the fidelity of LIME, the fidelity of the framework is to some extent uncertain. In this research we do not perform formal validation either, instead only using anecdotal evidence from a domain expert for one of our two case studies. As a suggestion for future research, it would be valuable to assess the performance of LIME including fidelity. A possible method for this would be to use the Co-12 properties and associated evaluation methods introduced by Nauta et al. [55].

As a final recommendation, we would suggest improving the presentation of the explanations provided by LIME. The paper by Dieber and Kirrane [40] mentioned earlier, concludes that the visualisation of explanations offered by LIME are not very intuitive to interpret without background knowledge of what each presented element of the explanation represents. Furthermore, some parts of the explanations are cut off when the explanation is presented. These cut-offs lead to, for example, discretisation boundaries being unreadable. The raw information of the explanation, such as feature influence on the prediction value, is accessible in the Python LIME implementation. As such, it is possible to use this information to create a custom, improved presentation of the explanation. Examples of such an improvement are a clearer guide of what each element of the explanation represents, as well as a different display of the feature influence plot, so that feature names and values are not cut off as they are now.

Chapter 6

Conclusion

In this chapter, we will answer each of our subquestions based on the performed systematic literature review and subsequently answer our main research question "How can we improve the interpretability and transparency of machine learning models aimed at prediction of disease progression?". Furthermore, we present our recommendations for future research.

6.1 Subquestions

We start by answering each subquestion as defined in Chapter 1 separately.

6.1.1 Subquestion 1

Our first subquestion is as follows: "What are current trends and state-of-the-art in machine learning for disease detection and prediction of disease progression?". From our systematic literature review included in Appendix C, we conclude that predicting a continuous value is preferred over a classification or discretised value, as a continuous value is the most precise definition of what the disease severity level will be. A drawback of a continuous prediction compared to a classification, for example, is that a good proxy variable for disease severity must be identified for the disease of which one is predicting the progression. Furthermore, we find that time-series patient data leads to the highest performance, due to the complete overview of the progression mechanism that can be captured by looking at feature interactions both at a certain time as well as over time. Finally, we conclude that to accommodate the above requirements for state-of-the-art performance consists of deep, recurrent neural network based models. These models are capable of handling the high complexity of the input data due to their sophisticated deep structure, as well as the time-series interpretation aspect as a result of the recurrent nature of the model.

6.1.2 Subquestion 2

The second subquestion we formulated, concerns data quality challenges: "What are common data quality challenges in machine learning for prediction of disease progression?". We identify various challenges in our systematic literature review, of which we highlight three.

The first challenge identified is a high degree of heterogeneity in patient data. Patients with the same disease may progress at very different rates, and even have biomarker values that are highly uncommon for the condition that they are in. This is something inherent to healthcare data which cannot be avoided by, for example, performing additional data pre-processing.

The second challenge is irregular temporal alignment of patient records. This challenge is caused by patients having different time intervals between visits or assessments, but also by patients missing a visit. Depending on the time scale of the time-series data, interval discrepancies may or may not be significant for modelling purposes.

The final challenge we find in the literature is a high frequency of missing values in medical datasets. This challenge may be the result of physicians not requesting the same assessment for each patient visit, or an assessment failing to be taken or registered properly. Missing data issues can be tackled in various ways, such as imputation or removal of incomplete records.

6.1.3 Subquestion 3

Our third and final subquestion is "What are common techniques for explaining machine learning models in the context of disease progression modelling?". From our literature review, we only learn that for state-of-the-art model types, there is very little, if any, focus on transparency or interpretability. After the systematic literature review, we use knowledge from other domains to identify explainability frameworks that are possibly applicable in the context of ML models for predicting the progression of diseases.

6.2 Main research question

By combining the knowledge gained from answering our three subquestions, we can answer our main research question: "How can we improve the interpretability and transparency of machine learning models aimed at prediction of disease progression?". As this is a DSR, explained in Chapter 3, our main research question is essentially a design problem. Thus, we aim not only to gain knowledge on what needs to be done to achieve the solution to the design problem, but also to create an artefact that solves the design problem.

If we start by answering our main research question as if it were a knowledge question, the answer would be as follows: the interpretability and transparency of machine learning models aimed at prediction of disease progression can be improved by integrating explainability frameworks in the prediction process, such that the result of the entire process is not only a prediction of disease progression, but also an explanation of why the model generates that prediction. However, because we treat our main research question as a design problem, we also create the artefact that does this.

The resulting artefact we design and develop is our pipeline. The pipeline solves the research problem that we identify at the start of this research: state-of-the-art machine learning models for prediction of disease progression are not adequately transparent or interpretable. Our pipeline does offer transparency and interpretability, whilst using a state-of-the-art prediction model architecture. The demonstration of the functionality of the pipeline, through applying it to two separate case studies with very different datasets, shows that the pipeline is applicable to a range of diseases, even if particular components do not perform well in their current form. Furthermore, a domain expert and data owner for one of the case studies mentions that the predictions and explanations provided by our pipeline align with his expectations, and that the majority of explanations offered by the pipeline seem to be logical. For the other case study, we observe poor explanatory performance and hypothesise that this is due to the extremely high dimensionality of the dataset. We also provide possible solutions for the poor performance in case our hypothesis is correct.

6.3 Future research

This research identifies a number of opportunities for future work. These opportunities are primarily focused on two aspects of the pipeline: the predictive model and the explanations.

For the first aspect, the predictive phase, we recommend improving the flexibility of the predictive model with regards to input data. Valuable improvements include the ability to handle data sequences of unequal length within a dataset and robustness to missing records or missing values within a record. These improvements may increase the number of disease contexts that the pipeline can be applied in.

We provide two recommendations related to the explanation aspect of the pipeline. We suggest investigating whether changing the kernel width calculation for LIME leads to improvements in the fidelity of generated explanations. Such changes could prevent fitment problems with the local approximation model that LIME trains as a result of an oversized kernel width. Furthermore, we recommend improving the visualisation of the explanations offered by LIME, as both literature and a domain expert in this research consider the visualisations difficult to interpret.

Bibliography

[1] M. I. Jordan and T. M. Mitchell, “Machine learning: Trends, perspectives, and prospects,” *Science*, vol. 349, no. 6245, pp. 255–260, 2015. [Online]. Available: https://doi.org/10.1126/science.aaa8415

[2] E. Vayena, A. Blasimme, and I. G. Cohen, “Machine learning in medicine: Addressing ethical challenges,” *PLOS Medicine*, vol. 15, no. 11, pp. 1–4, 11 2018. [Online]. Available: https://doi.org/10.1371/journal.pmed.1002689

[3] D. Gunning and D. Aha, “Darpa’s explainable artificial intelligence (xai) program,” *AI Magazine*, vol. 40, no. 2, pp. 44–58, June 2019. [Online]. Available: https://doi.org/10.1609/aimag.v40i2.2850

[4] E. Tjoa and C. Guan, “A survey on explainable artificial intelligence (xai): Toward medical xai,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4793–4813, 2021.

[5] D. R. Mould, “Models for disease progression: new approaches and uses,” *Clinical Pharmacology & Therapeutics*, vol. 92, no. 1, pp. 125–131, 2012. [Online]. Available: https://doi.org/10.1038/clpt.2012.53

[6] K. Ito, S. Ahadieh, B. Corrigan, J. French, T. Fullerton, and T. Tensfeldt, “Disease progression meta-analysis model in alzheimer’s disease,” *Alzheimer’s & Dementia*, vol. 6, no. 1, pp. 39–53, 2010. [Online]. Available: https://doi.org/10.1016/j.jalz.2009.05.665

[7] T. M. Post, J. I. Freijer, J. DeJongh, and M. Danhof, “Disease system analysis: basic disease progression models in degenerative disease,” *Pharmaceutical research*, vol. 22, no. 7, pp. 1038–1049, 2005. [Online]. Available: https://doi.org/10.1007/s11095-005-5641-5

[8] A. Rafiee Zadeh, M. Askari, N. N. Azadani, A. Ataei, K. Ghadimi, N. Tavoosi, and M. Falahatian, “Mechanism and adverse effects of multiple sclerosis drugs: a review article. part 1,” *International Journal of Physiology, Pathophysiology and Pharmacology*, vol. 11, no. 4, pp. 95–104, Aug. 2019.

[9] A. Rafiee Zadeh, K. Ghadimi, A. Ataei, M. Askari, N. Sheikhinia, N. Tavoosi, and M. Falahatian, “Mechanism and adverse effects of multiple sclerosis drugs: a review article. part 2,” *International Journal of Physiology, Pathophysiology and Pharmacology*, vol. 11, no. 4, pp. 105–114, Aug. 2019.

[10] F. Papaiz, J. Dourado, M. E. T., R. A. D. M. Valentim, A. H. F. de Morais, and J. P. Arrais, “Machine learning solutions applied to amyotrophic lateral sclerosis prognosis: A review,” *Frontiers in Computer Science*, vol. 4, 2022. [Online]. Available: https://doi.org/10.3389/fcomp.2022.869140

[11] J. Bhutani and S. Bhutani, “Worldwide burden of diabetes,” *Indian Journal of Endocrinology and Metabolism*, vol. 18, no. 6, pp. 868–870, Nov. 2014. [Online]. Available: http://doi.org/10.4103/2230-8210.141388

[12] GBD 2016 Parkinson’s Disease Collaborators, “Global, regional, and national burden of parkinson’s disease, 1990-2016: a systematic analysis for the global burden of disease study 2016,” *The Lancet Neurology*, vol. 17, no. 11, pp. 939–953, Nov. 2018. [Online]. Available: https://doi.org/10.1016/S1474-4422(18)30295-3

[13] L. V. Kalia and A. E. Lang, “Parkinson’s disease,” *The Lancet*, vol. 386, no. 9996, pp. 896–912, 2015. [Online]. Available: https://doi.org/10.1016/S0140-6736(14)61393-3

[14] B. R. Bloem, M. S. Okun, and C. Klein, “Parkinson’s disease,” *The Lancet*, vol. 397, no. 10291, pp. 2284–2303, Jun. 2021. [Online]. Available: https://doi.org/10.1016/S0140-6736(21)00218-X

[15] C. G. Goetz, B. C. Tilley, S. R. Shaftman, G. T. Stebbins, S. Fahn, P. Martinez-Martin, W. Poewe, C. Sampaio, M. B. Stern, R. Dodel, B. Dubois, R. Holloway, J. Jankovic, J. Kulisevsky, A. E. Lang, A. Lees, S. Leurgans, P. A. LeWitt, D. Nyenhuis, C. W. Olanow, O. Rascol, A. Schrag, J. A. Teresi, J. J. van Hilten, and N. LaPelle, “Movement disorder society-sponsored revision of the unified parkinson’s disease rating scale (mds-updrs): Scale presentation and clinimetric testing results,” *Movement Disorders*, vol. 23, no. 15, pp. 2129–2170, 2008. [Online]. Available: https://doi.org/10.1002/mds.22340

[16] W. H. Organisation, “Diabetes,” World Health Organisation, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes

[17] A. D. Association, “Diagnosis and Classification of Diabetes Mellitus,” *Diabetes Care*, vol. 37, no. Supplement_1, pp. S81–S90, 12 2013. [Online]. Available: https://doi.org/10.2337/dc14-S081

[18] R. Alicic, M. Rooney, and K. Tuttle, “Diabetic kidney disease: Challenges, progress, and possibilities,” *Clinical journal of the American Society of Nephrology*, vol. 12, 05 2017. [Online]. Available: https://doi.org/10.2215/CJN.11491116

[19] C. P. Wen, C. H. Chang, M. K. Tsai, J. H. Lee, P. J. Lu, S. P. Tsai, C. Wen, C. H. Chen, C. W. Kao, C. K. Tsao, and X. Wu, “Diabetes with early kidney involvement may shorten life expectancy by 16 years,” *Kidney International*, vol. 92, no. 2, p. 388 – 396, 2017, cited by: 84; All Open Access, Bronze Open Access.

[20] S. Hussain, A. Habib, and A. K. Najmi, “Anemia prevalence and its impact on health-related quality of life in indian diabetic kidney disease patients: Evidence from a cross-sectional study,” *Journal of Evidence-Based Medicine*, vol. 12, no. 4, p. 243 – 252, 2019, cited by: 10. [Online]. Available: https://doi.org/10.1111/jebm.12367

[21] S. Hussain, M. Chand Jamali, A. Habib, M. S. Hussain, M. Akhtar, and A. K. Najmi, “Diabetic kidney disease: An overview of prevalence, risk factors, and biomarkers,” *Clinical Epidemiology and Global Health*, vol. 9, pp. 2–6, 2021. [Online]. Available: https://doi.org/10.1016/j.cegh.2020.05.016

[22] T. J. Lyons and A. Basu, “Biomarkers in diabetes: hemoglobin a1c, vascular and tissue markers,” *Translational Research*, vol. 159, no. 4, pp. 303–312, 2012, biomarkers: New Tools of Modern Medicine. [Online]. Available: https://doi.org/10.1016/j.trsl.2012.01.009

[23] A. S. Levey, J. P. Bosch, J. B. Lewis, T. Greene, N. Rogers, and D. Roth, “A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation. modification of diet in renal disease study group,” *Annals of Internal Medicine*, vol. 130, no. 6, pp. 461–470, Mar. 1999. [Online]. Available: https://doi.org/10.7326/0003-4819-130-6-199903160-00002

[24] A. S. Levey, L. A. Stevens, C. H. Schmid, Y. L. Zhang, A. F. Castro, 3rd, H. I. Feldman, J. W. Kusek, P. Eggers, F. Van Lente, T. Greene, J. Coresh, and CKD-EPI (Chronic Kidney Disease Epidemiology Collaboration), “A new equation to estimate glomerular filtration rate,” *Annals of Internal Medicine*, vol. 150, no. 9, pp. 604–612, May 2009. [Online]. Available: https://doi.org/10.7326/0003-4819-150-9-200905050-00006

[25] A. Krogh, “What are artificial neural networks?” *Nature Biotechnology*, vol. 26, no. 2, pp. 195–197, Feb. 2008. [Online]. Available: https://doi.org/10.1038/nbt1386

[26] M. A. Nielsen, *Toward deep learning.* Determination Press, 2015, p. 35–37.

[27] C. C. Aggarwal, *An Introduction to Neural Networks.* Springer International PU, 2023, p. 1–48. [Online]. Available: https://doi.org/10.1007/978-3-319-94463-0

[28] A. Mcgovern, R. Lagerquist, D. Gagne, G. Jergensen, K. Elmore, C. Homeyer, and T. Smith, “Making the black box more transparent: Understanding the physical implications of machine learning,” *Bulletin of the American Meteorological Society*, vol. 100, 08 2019. [Online]. Available: https://doi.org/10.1175/BAMS-D-18-0195.1

[29] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *nature*, vol. 323, no. 6088, pp. 533–536, 1986. [Online]. Available: https://doi.org/10.1038/323533a0

[30] G. Petneházi, “Recurrent neural networks for time series forecasting,” *Computing Research Repository*, vol. abs/1901.00069, 2019. [Online]. Available: https://doi.org/10.48550/arXiv.1901.00069

[31] Y. Bengio, P. Simard, and P. Frasconi, “Learning long-term dependencies with gradient descent is difficult,” *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[32] J. Dancker, “A brief introduction to recurrent neural networks,” Dec 2022. [Online]. Available: https://towardsdatascience.com/a-brief-introduction-to-recurrent-neural-networks-638f64a61ff4

[33] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, p. 1735–1780, nov 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[34] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, “On the properties of neural machine translation: Encoder–decoder approaches,” in *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation.* Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 103–111. [Online]. Available: http://doi.org/10.3115/v1/W14-4012

[35] A. Graves, *Long Short-Term Memory.* Springer Berlin, 2014, p. 37–46. [Online]. Available: https://doi.org/10.1007/978-3-642-24797-2

[36] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” in *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014. [Online]. Available: https://doi.org/10.48550/arXiv.1412.3555

[37] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS’17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 4768–4777. [Online]. Available: https://doi.org/10.48550/arXiv.1705.07874

[38] S. Hart, “Shapley value,” in *Game Theory*, J. Eatwell, M. Milgate, and P. Newman, Eds. London: Palgrave Macmillan UK, 1989, pp. 210–216. [Online]. Available: https://doi.org/10.1007/978-1-349-20181-5_25

[39] M. T. Ribeiro, S. Singh, and C. Guestrin, “"why should i trust you?": Explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD ’16. New York, NY, USA: Association for Computing Machinery, 2016, p. 1135–1144. [Online]. Available: https://doi.org/10.1145/2939672.2939778

[40] J. Dieber and S. Kirrane, “Why model why? assessing the strengths and limitations of lime,” 2020. [Online]. Available: https://doi.org/10.48550/arXiv.2012.00093

[41] S. E. Whang and J.-G. Lee, “Data collection and quality challenges for deep learning,” *Proceedings of the VLDB Endowment*, vol. 13, no. 12, p. 3429–3432, aug 2020. [Online]. Available: https://doi.org/10.14778/3415478.3415562

[42] D. Rubin and R. J. A. Little, *Single Imputation Methods.* John Wiley & Sons, Ltd, 2019, ch. 4, pp. 67–84. [Online]. Available: https://doi.org/10.1002/9781119482260.ch4

[43] J. Kendrick, R. Francis, G. M. Hassan, P. Rowshanfarzad, R. Jeraj, C. Kasisi, B. Rusanov, and M. Ebert, "Radiomics for identification and prediction in metastatic prostate cancer: A review of studies," *Frontiers in Oncology*, vol. 11, 2021. [Online]. Available: https://doi.org/10.3389/fonc.2021.771787

[44] R. J. Wieringa, *What is Design Science?* Springer, 2014, p. 3–12. [Online]. Available: https://doi.org/10.1007/978-3-662-43839-8

[45] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," *Journal of Management Information Systems*, vol. 24, no. 3, pp. 45–77, 2007. [Online]. Available: https://doi.org/10.2753/MIS0742-1222240302

[46] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: https://doi.org/10.48550/arXiv.1412.6980

[47] Freepik, "All the assets you need, in one place." [Online]. Available: https://www.freepik.com/

[48] F. Pukelsheim, "The three sigma rule," *The American Statistician*, vol. 48, no. 2, pp. 88–91, 1994.

[49] S. K. Van Den Eeden, C. M. Tanner, A. L. Bernstein, R. D. Fross, A. Leimpeter, D. A. Bloch, and L. M. Nelson, "Incidence of Parkinson's Disease: Variation by Age, Gender, and Race/Ethnicity," *American Journal of Epidemiology*, vol. 157, no. 11, pp. 1015–1022, 06 2003. [Online]. Available: https://doi.org/10.1093/aje/kwg068

[50] S. Wilkerson, "Application of the paired t-test," *XULAneXUS*, vol. 5, no. 1, 2008. [Online]. Available: https://digitalcommons.xula.edu/xulanexus/vol5/iss1/7

[51] E. De Brouwer, T. Becker, Y. Moreau, E. K. Havrdova, M. Trojano, S. Eichau, S. Ozakbas, M. Onofrj, P. Grammond, J. Kuhle, L. Kappos, P. Sola, E. Cartechini, J. Lechner-Scott, R. Alroughani, O. Gerlach, T. Kalincik, F. Granella, F. Grand'Maison, R. Bergamaschi, M. José Sá, B. Van Wijmeersch, A. Soysal, J. L. Sanchez-Menoyo, C. Solaro, C. Boz, G. Iuliano, K. Buzzard, E. Aguera-Morales, M. Terzi, T. C. Trivio, D. Spitaleri, V. Van Pesch, V. Shaygannejad, F. Moore, C. Oreja-Guevara, D. Maimone, R. Gouider, T. Csepany, C. Ramo-Tello, and L. Peeters, "Longitudinal machine learning modeling of ms patient trajectories improves predictions of disability progression," *Computer Methods and Programs in Biomedicine*, vol. 208, 2021. [Online]. Available: https://doi.org/10.1016/j.cmpb.2021.106180

[52] B. Rad, F. Song, V. Jacob, and Y. Diao, "Explainable anomaly detection on high-dimensional time series data," in *Proceedings of the 15th ACM International Conference on Distributed and Event-Based Systems*, ser. DEBS '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 2–14. [Online]. Available: https://doi.org/10.1145/3465480.3468292

[53] C. Molnar, *Local Surrogate (LIME).* Lulu.com, 2020.

[54] E. Barents, H. Bilo, M. Bouma, M. Dankers, A. De Rooij, H. Hart, S. Houweling, R. IJzerman, P. Janssen, A. Kerssen, M. Oud, J. Palmen, A. Van den Brink-Muinen, M. Van den Donk, A. Verburg-Oorthuizen, and T. Wiersma, "Diabetes mellitus type 2." [Online]. Available: https://richtlijnen.nhg.org/standaarden/diabetes-mellitus-type-2#samenvatting-controles

[55] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, and C. Seifert, "From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI," *ACM Computing Surveys*, vol. 55, no. 13s, pp. 1–42, jul 2023. [Online]. Available: https://doi.org/10.1145%2F3583558

[56] B. Kitchenham, "Procedures for performing systematic reviews," *Keele, UK, Keele University*, vol. 33, 08 2004.

[57] A. Nunes, R. Ardau, A. Berghöfer, A. Bocchetta, C. Chillotti, V. Deiana, J. Garnham, E. Grof, T. Hajek, M. Manchia, B. Müller-Oerlinghausen, M. Pinna, C. Pisanu, C. O'Donovan, G. Severino, C. Slaney, A. Suwalska, P. Zvolsky, P. Cervantes, M. del Zompo, P. Grof, J. Rybakowski, L. Tondo, T. Trappenberg, and M. Alda, "Prediction of lithium response using clinical data," *Acta Psychiatrica Scandinavica*, vol. 141, no. 2, pp. 131–141, 2020. [Online]. Available: https://doi.org/10.1111/acps.13122

[58] Y. Muhammad, M. Almoteri, H. Mujlid, A. Alharbi, F. Alqurashi, A. K. Dutta, S. Almotairi, and H. Almohamedh, "An ML-Enabled internet of things framework for early detection of heart disease," *BioMed research international*, vol. 2022, p. 3372296, Sep. 2022. [Online]. Available: https://doi.org/10.1155/2022/3372296

[59] G. Marti-Juan, G. Sanroma-Guell, and G. Piella, "A survey on machine and statistical learning for longitudinal analysis of neuroimaging data in alzheimer's disease," *Computer Methods and Programs in Biomedicine*, vol. 189, 2020. [Online]. Available: https://doi.org/10.1016/j.cmpb.2020.105348

[60] Z. Li, X. Jiang, Y. Wang, and Y. Kim, "Applied machine learning in alzheimer's disease research: Omics, imaging, and clinical data," *Emerging Topics in Life Sciences*, vol. 5, no. 6, pp. 765–777, 2021. [Online]. Available: https://doi.org/10.1042/ETLS20210249

[61] M. Amini, M. M. Pedram, A. Moradi, M. Jamshidi, and M. Ouchani, "Single and combined neuroimaging techniques for alzheimer's disease detection," *Computational Intelligence and Neuroscience*, vol. 2021, 2021. [Online]. Available: https://doi.org/10.1155/2021/9523039

[62] N. Goenka and S. Tiwari, "Deep learning for alzheimer prediction using brain biomarkers," *Artificial Intelligence Review*, vol. 54, no. 7, pp. 4827–4871, 2021. [Online]. Available: https://doi.org/10.1007/s10462-021-10016-0

[63] D. Agarwal, G. Marques, I. de la Torre-Díez, M. A. Franco Martin, B. García Zapiraín, and F. Martín Rodríguez, "Transfer learning for alzheimer's disease through neuroimaging biomarkers: A systematic review," *Sensors*, vol. 21, no. 21, 2021. [Online]. Available: https://doi.org/10.3390/s21217259

[64] C. Su, Z. Xu, J. Pathak, and F. Wang, "Deep learning in mental health outcome research: a scoping review," *Translational Psychiatry*, vol. 10, no. 1, 2020. [Online]. Available: https://doi.org/10.1038/s41398-020-0780-3

[65] M. Z. Hossain, E. Daskalaki, A. Brustle, J. Desborough, C. J. Lueck, and H. Suominen, "The role of machine learning in developing non-magnetic resonance imaging based biomarkers for multiple sclerosis: a systematic review," *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, 2022. [Online]. Available: https://doi.org/10.1186/s12911-022-01985-5

[66] A. M. Westerlund, J. S. Hawe, M. Heinig, and H. Schunkert, "Risk prediction of cardiovascular events by exploration of molecular data with explainable artificial intelligence," *International Journal of Molecular Sciences*, vol. 22, no. 19, 2021. [Online]. Available: https://doi.org/10.3390/ijms221910291

[67] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research," *Computational and Structural Biotechnology Journal*, vol. 15, pp. 104–116, 2017. [Online]. Available: https://doi.org/10.1016/j.csbj.2016.12.005

[68] F. Fernandes, I. Barbalho, D. Barros, R. Valentim, C. Teixeira, J. Henriques, P. Gil, and M. Dourado Júnior, "Biomedical signals and machine learning in amyotrophic lateral sclerosis: a systematic review," *BioMedical Engineering Online*, vol. 20, no. 1, 2021. [Online]. Available: https://doi.org/10.1186/s12938-021-00896-2

[69] L. J. Marcos-Zambrano, K. Karaduzovic-Hadziabdic, T. Loncar Turukalo, P. Przymus, V. Trajkovik, O. Aasmets, M. Berland, A. Gruca, J. Hasic, K. Hron, T. Klammsteiner, M. Kolev, L. Lahti, M. B. Lopes, V. Moreno, I. Naskinova, E. Org, I. Paciência, G. Papoutsoglou, R. Shigdel, B. Stres, B. Vilne, M. Yousef, E. Zdravevski, I. Tsamardinos,

E. Carrillo de Santa Pau, M. J. Claesson, I. Moreno-Indias, J. Truu, and Ml4Microbiome, "Applications of machine learning in human microbiome studies: A review on feature selection, biomarker identification, disease prediction and treatment," *Frontiers in Microbiology*, vol. 12, 2021. [Online]. Available: https://doi.org/10.3389/fmicb.2021.634511

[70] S. Mistry, N. O. Riches, R. Gouripeddi, and J. C. Facelli, "Environmental exposures in machine learning and data mining approaches to diabetes etiology: A scoping review," *Artificial Intelligence in Medicine*, vol. 135, 2023. [Online]. Available: https://doi.org/10.1016/j.artmed.2022.102461

[71] P. Kaur, A. Singh, and I. Chana, "Computational techniques and tools for omics data analysis: State-of-the-art, challenges, and future directions," *Archives of Computational Methods in Engineering*, vol. 28, no. 7, pp. 4595–4631, 2021. [Online]. Available: https://doi.org/10.1007/s11831-021-09547-0

[72] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 34, no. 2, pp. 187–220, 1972. [Online]. Available: https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

[73] A. Nazha, R. Komrokji, M. Meggendorfer, X. Jia, N. Radakovich, J. Shreve, C. Beau Hilton, Y. Nagata, B. K. Hamilton, S. Mukherjee, N. Al Ali, W. Walter, S. Hutter, E. Padron, D. Sallman, T. Kuzmanovic, C. Kerr, V. Adema, D. P. Steensma, A. Dezern, G. Roboz, G. Garcia-Manero, H. Erba, C. Haferlach, J. P. Maciejewski, T. Haferlach, and M. A. Sekeres, "Personalized prediction model to risk stratify patients with myelodysplastic syndromes," *Journal of Clinical Oncology*, vol. 39, no. 33, pp. 3737–3746, 2021. [Online]. Available: https://doi.org/10.1200/JCO.20.02810

[74] X. Chen, W. Gao, J. Li, D. You, Z. Yu, M. Zhang, F. Shao, Y. Wei, R. Zhang, T. Lange, Q. Wang, F. Chen, X. Lu, and Y. Zhao, "A predictive paradigm for covid-19 prognosis based on the longitudinal measure of biomarkers," *Briefings in Bioinformatics*, vol. 22, no. 6, 2021. [Online]. Available: https://doi.org/10.1093/bib/bbab206

[75] U. Schmidt-Erfurth, S. M. Waldstein, S. Klimscha, A. Sadeghipour, X. Hu, B. S. Gerendas, A. Osborne, and H. Bogunović, "Prediction of individual disease conversion in early amd using artificial intelligence," *Investigative Ophthalmology and Visual Science*, vol. 59, no. 8, pp. 3199–3208, 2018. [Online]. Available: https://doi.org/10.1167/iovs.18-24106

[76] M. Taghavi, F. Staal, F. Gomez Munoz, F. Imani, D. B. Meek, R. Simões, L. G. Klompenhouwer, U. A. van der Heide, R. G. H. Beets-Tan, and M. Maas, "Ct-based radiomics analysis before thermal ablation to predict local tumor progression for colorectal liver metastases," *CardioVascular and Interventional Radiology*, vol. 44, no. 6, pp. 913–920, 2021. [Online]. Available: https://doi.org/10.1007/s00270-020-02735-8

[77] S. W. M. Eng, F. A. Aeschlimann, M. van Veenendaal, R. A. Berard, A. M. Rosenberg, Q. Morris, and R. S. M. Yeung, "Patterns of joint involvement in juvenile idiopathic arthritis and prediction of disease course: A prospective study with multilayer non-negative matrix factorization," *PLoS Medicine*, vol. 16, no. 2, 2019. [Online]. Available: https://doi.org/10.1371/journal.pmed.1002750

[78] Z. Yılmaz Acar, F. Başçiftçi, and A. H. Ekmekci, "Future activity prediction of multiple sclerosis with 3d mri using 3d discrete wavelet transform," *Biomedical Signal Processing and Control*, vol. 78, 2022. [Online]. Available: https://doi.org/10.1016/j.bspc.2022.103940

[79] R. Seccia, S. Romano, M. Salvetti, A. Crisanti, L. Palagi, and F. Grassi, "Machine learning use for prognostic purposes in multiple sclerosis," *Life*, vol. 11, no. 2, pp. 1–18, 2021. [Online]. Available: https://doi.org/10.3390/life11020122

[80] S. El-Sappagh, H. Saleh, F. Ali, E. Amer, and T. Abuhmed, "Two-stage deep learning model for alzheimer's disease detection and prediction of the mild cognitive impairment time," *Neural Computing and Applications*, vol. 34, no. 17, pp. 14 487–14 509, 2022. [Online]. Available: https://doi.org/10.1007/s00521-022-07263-9

[81] C. Sun, S. Hong, M. Song, H. Li, and Z. Wang, “Predicting covid-19 disease progression and patient outcomes based on temporal deep learning,” *BMC Medical Informatics and Decision Making*, vol. 21, no. 1, 2021. [Online]. Available: https://doi.org/10.1186/s12911-020-01359-9

[82] N. H. Ho, H. J. Yang, J. Kim, D. P. Dao, H. R. Park, and S. Pant, “Predicting progression of alzheimer’s disease using forward-to-backward bi-directional network with integrative imputation,” *Neural Networks*, vol. 150, pp. 422–439, 2022. [Online]. Available: https://doi.org/10.1016/j.neunet.2022.03.016

[83] L. Chen, A. J. Saykin, B. Yao, F. Zhao, and I. Alzheimer’s Disease Neuroimaging, “Multi-task deep autoencoder to predict alzheimer’s disease progression using temporal dna methylation data in peripheral blood,” *Computational and Structural Biotechnology Journal*, vol. 20, pp. 5761–5774, 2022. [Online]. Available: https://doi.org/10.1016/j.csbj.2022.10.016

[84] Y. An, K. Tang, and J. Wang, “Time-aware multi-type data fusion representation learning framework for risk prediction of cardiovascular diseases,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2021. [Online]. Available: https://doi.org/10.1109/TCBB.2021.3118418

[85] K. He, S. Huang, and X. N. Qian, “Early detection and risk assessment for chronic disease with irregular longitudinal data analysis,” *Journal of Biomedical Informatics*, vol. 96, 2019. [Online]. Available: https://doi.org/10.1016/j.jbi.2019.103231

[86] I. D. Dinov, B. Heavner, M. Tang, G. Glusman, K. Chard, M. Darcy, R. Madduri, J. Pa, C. Spino, C. Kesselman, I. Foster, E. W. Deutsch, N. D. Price, J. D. Van Horn, J. Ames, K. Clark, L. Hood, B. M. Hampstead, W. Dauer, and A. W. Toga, “Predictive big data analytics: A study of parkinson’s disease using large, complex, heterogeneous, incongruent, multi-source and incomplete observations,” *PLoS ONE*, vol. 11, no. 8, 2016. [Online]. Available: https://doi.org/10.1371/journal.pone.0157077

[87] A. Dadu, V. Satone, R. Kaur, S. H. Hashemi, H. Leonard, H. Iwaki, M. B. Makarious, K. J. Billingsley, S. Bandres-Ciga, L. J. Sargent, A. J. Noyce, A. Daneshmand, C. Blauwendraat, K. Marek, S. W. Scholz, A. B. Singleton, M. A. Nalls, R. H. Campbell, and F. Faghri, “Identification and prediction of parkinson’s disease subtypes andprogression using machine learning in two cohorts,” *NPJ Parkinson’s Disease*, vol. 8, no. 1, p. 172, 2022. [Online]. Available: https://doi.org/10.1038/s41531-022-00439-z

[88] L. Y. Ma, Y. Tian, C. R. Pan, Z. L. Chen, Y. Ling, K. Ren, J. S. Li, and T. Feng, “Motor progression in early-stage parkinson’s disease: A clinical prediction model and the role of cerebrospinal fluid biomarkers,” *Frontiers in Aging Neuroscience*, vol. 12, 2021. [Online]. Available: https://doi.org/10.3389/fnagi.2020.627199

[89] N. Bhagwat, J. D. Viviano, A. N. Voineskos, M. M. Chakravarty, and I. Alzheimer’s Disease Neuroimaging, “Modeling and prediction of clinical symptom trajectories in alzheimer’s disease using longitudinal data,” *PLoS Computational Biology*, vol. 14, no. 9, 2018. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1006376

[90] M. U. Sadiq, K. Kwak, E. Dayan, and I. for the Alzheimer’s Disease Neuroimaging, “Model-based stratification of progression along the alzheimer disease continuum highlights the centrality of biomarker synergies,” *Alzheimer’s Research and Therapy*, vol. 14, no. 1, 2022. [Online]. Available: https://doi.org/10.1186/s13195-021-00941-1

[91] P. Schulam and S. Saria, “Integrative analysis using coupled latent variable models for individualizing prognoses,” *The Journal of Machine Learning Research*, vol. 17, no. 1, p. 8244–8278, jan 2016.

[92] D. Larie, G. An, and R. C. Cockrell, “The use of artificial neural networks to forecast the behavior of agent-based models of pathophysiology: An example utilizing an agent-based model of sepsis,” *Frontiers in Physiology*, vol. 12, 2021. [Online]. Available: https://doi.org/10.3389/fphys.2021.716434

[93] T. Dang, J. Han, T. Xia, D. Spathis, E. Bondareva, C. Siegele-Brown, J. Chauhan, A. Grammenos, A. Hasthanasombat, R. A. Floto, P. Cicuta, and C. Mascolo, "Exploring longitudinal cough, breath, and voice data for covid-19 progression prediction via sequential deep learning: Model development and validation," *Journal of Medical Internet Research*, vol. 24, no. 6, 2022. [Online]. Available: https://doi.org/10.2196/37004

[94] W. Jung, E. Jun, H. I. Suk, and I. Alzheimer's Disease Neuroimaging, "Deep recurrent model for individualized prediction of alzheimer's disease progression," *NeuroImage*, vol. 237, 2021. [Online]. Available: https://doi.org/10.1016/j.neuroimage.2021.118143

[95] C. Sun, H. Li, R. E. Mills, and Y. Guan, "Prognostic model for multiple myeloma progression integrating gene expression and clinical features," *GigaScience*, vol. 8, no. 12, 2019. [Online]. Available: https://doi.org/10.1093/gigascience/giz153

[96] L. Chan, G. N. Nadkarni, F. Fleming, J. R. McCullough, P. Connolly, G. Mosoyan, F. El Salem, M. W. Kattan, J. A. Vassalotti, B. Murphy, M. J. Donovan, S. G. Coca, and S. M. Damrauer, "Derivation and validation of a machine learning risk score using biomarker and electronic patient data to predict progression of diabetic kidney disease," *Diabetologia*, vol. 64, no. 7, pp. 1504–1515, 2021. [Online]. Available: https://doi.org/10.1007/s00125-021-05444-0

[97] R. Sharma, H. Anand, Y. Badr, and R. G. Qiu, "Time-to-event prediction using survival analysis methods for alzheimer's disease progression," *Alzheimer's and Dementia: Translational Research and Clinical Interventions*, vol. 7, no. 1, 2021. [Online]. Available: https://doi.org/10.1002/trc2.12229

[98] J. Zhang, J. Ning, X. Huang, and R. Li, "On the time-varying predictive performance of longitudinal biomarkers:measure and estimation," *Stat Med*, vol. 40, no. 23, pp. 5065–5077, 2021. [Online]. Available: https://doi.org/10.1002/sim.9111

[99] X. Liu, J. Wang, F. Ren, and J. Kong, "Group guided fused laplacian sparse group lasso for modeling alzheimer's disease progression," *Computational and Mathematical Methods in Medicine*, vol. 2020, 2020. [Online]. Available: https://doi.org/10.1155/2020/4036560

[100] Y. Zheng and X. Hu, "Healthcare predictive analytics for disease progression: a longitudinal data fusion approach," *Journal of Intelligent Information Systems*, vol. 55, no. 2, pp. 351–369, 2020. [Online]. Available: https://doi.org/10.1007/s10844-020-00606-9

[101] M. Velazquez, Y. Lee, and I. for the Alzheimer's Disease Neuroimaging, "Random forest model for feature-based alzheimer's disease conversion prediction from early mild cognitive impairment subjects," *PLoS ONE*, vol. 16, no. 4 April, 2021. [Online]. Available: https://doi.org/10.1371/journal.pone.0244773

[102] N. Shafiee, M. Dadar, S. Ducharme, and D. L. Collins, "Automatic prediction of cognitive and functional decline can significantly decrease the number of subjects required for clinical trials in early alzheimer's disease," *Journal of Alzheimer's Disease*, vol. 84, no. 3, pp. 1071–1078, 2021. [Online]. Available: https://doi.org/10.3233/JAD-210664

[103] N. P. Singh, R. S. Bapi, and P. K. Vinod, "Machine learning models to predict the progression from early to late stages of papillary renal cell carcinoma," *Computers in Biology and Medicine*, vol. 100, pp. 92–99, 2018. [Online]. Available: https://doi.org/10.1016/j.compbiomed.2018.06.030

[104] W. Su, X. Wang, and R. D. Szczesniak, "Flexible link functions in a joint hierarchical gaussian process model," *Biometrics*, vol. 77, no. 2, pp. 754–764, 2020. [Online]. Available: https://doi.org/10.1111/biom.13291

[105] P. Jiang, X. Wang, Q. Li, L. Jin, and S. Li, "Correlation-aware sparse and low-rank constrained multi-task learning for longitudinal analysis of alzheimer's disease," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 4, pp. 1450–1456, 2019. [Online]. Available: https://doi.org/10.1109/JBHI.2018.2885331

[106] J. Giorgio, S. M. Landau, W. J. Jagust, P. Tino, and Z. Kourtzi, "Modelling prognostic trajectories of cognitive decline due to alzheimer's disease," *NeuroImage: Clinical*, vol. 26, 2020. [Online]. Available: https://doi.org/10.1016/j.nicl.2020.102199

[107] T. Uphaus, F. Steffen, M. Muthuraman, N. Ripfel, V. Fleischer, S. Groppa, T. Ruck, S. G. Meuth, R. Pul, C. Kleinschnitz, E. Ellwardt, J. Loos, S. Engel, F. Zipp, and S. Bittner, "Nfl predicts relapse-free progression in a longitudinal multiple sclerosis cohort study: Serum nfl predicts relapse-free progression," *EBioMedicine*, vol. 72, 2021. [Online]. Available: https://doi.org/10.1016/j.ebiom.2021.103590

[108] J. F. Silva and S. Matos, "Modelling patient trajectories using multimodal information," *Journal of Biomedical Informatics*, vol. 134, 2022. [Online]. Available: https://doi.org/10.1016/j.jbi.2022.104195

[109] L. M. Aksman, M. A. Scelsi, A. F. Marquand, D. C. Alexander, S. Ourselin, A. Altmann, and A. for, "Modeling longitudinal imaging biomarkers with parametric bayesian multi-task learning," *Human Brain Mapping*, vol. 40, no. 13, pp. 3982–4000, 2019. [Online]. Available: https://doi.org/10.1002/hbm.24682

[110] M. Wang, D. Zhang, D. Shen, and M. Liu, "Multi-task exclusive relationship learning for alzheimer's disease progression prediction with longitudinal data," *Medical Image Analysis*, vol. 53, pp. 111–122, 2019. [Online]. Available: https://doi.org/10.1016/j.media.2019.01.007

[111] Z. Y. Shu, S. J. Cui, X. Wu, Y. Xu, P. Huang, P. P. Pang, and M. Zhang, "Predicting the progression of parkinson's disease using conventional mri and machine learning: An application of radiomic biomarkers in whole-brain white matter," *Magnetic Resonance in Medicine*, vol. 85, no. 3, pp. 1611–1624, 2021. [Online]. Available: https://doi.org/10.1002/mrm.28522

[112] M. Sindelar, E. Stancliffe, M. Schwaiger-Haber, D. S. Anbukumar, K. Adkins-Travis, C. W. Goss, J. A. O'Halloran, P. A. Mudd, W. C. Liu, R. A. Albrecht, A. García-Sastre, L. P. Shriver, and G. J. Patti, "Longitudinal metabolomics of human plasma reveals prognostic markers of covid-19 disease severity," *Cell Reports Medicine*, vol. 2, no. 8, 2021. [Online]. Available: https://doi.org/10.1016/j.xcrm.2021.100369

[113] M. J. Donovan, F. M. Khan, G. Fernandez, R. Mesa-Tejada, M. Sapir, V. B. Zubek, D. Powell, S. Fogarasi, Y. Vengrenyuk, M. Teverovskiy, M. R. Segal, R. J. Karnes, T. A. Gaffey, C. Busch, M. Haggman, P. Hlavcak, S. J. Freedland, R. T. Vollmer, P. Albertsen, J. Costa, and C. Cordon-Cardo, "Personalized prediction of tumor response and cancer progression on prostate needle biopsy," *Journal of Urology*, vol. 182, no. 1, pp. 125–132, 2009. [Online]. Available: https://doi.org/10.1016/j.juro.2009.02.135

[114] K. Morino, Y. Hirata, R. Tomioka, H. Kashima, K. Yamanishi, N. Hayashi, S. Egawa, and K. Aihara, "Predicting disease progression from short biomarker series using expert advice algorithm," *Scientific Reports*, vol. 5, 2015. [Online]. Available: https://doi.org/10.1038/srep08953

[115] A. Shiino, Y. Shirakashi, M. Ishida, K. Tanigaki, and I. Japanese Alzheimer's Disease Neuroimaging, "Machine learning of brain structural biomarkers for alzheimer's disease (ad) diagnosis, prediction of disease progression, and amyloid beta deposition in the japanese population," *Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring*, vol. 13, no. 1, 2021. [Online]. Available: https://doi.org/10.1002/dad2.12246

[116] G. Lee, K. Nho, B. Kang, K. A. Sohn, D. Kim, M. W. Weiner, P. Aisen, R. Petersen, J. Jack, C. R., W. Jagust, J. Q. Trojanowki, A. W. Toga, L. Beckett, R. C. Green, A. J. Saykin, J. Morris, L. M. Shaw, Z. Khachaturian, G. Sorensen, M. Carrillo, L. Kuller, M. Raichle, S. Paul, P. Davies, H. Fillit, F. Hefti, D. Holtzman, M. M. Mesulam, W. Potter, P. Snyder, T. Montine, R. G. Thomas, M. Donohue, S. Walter, T. Sather, G. Jiminez, A. B. Balasubramanian, J. Mason, I. Sim, D. Harvey, M. Bernstein, N. Fox, P. Thompson, N. Schuff, C. DeCarli, B. Borowski, J. Gunter, M. Senjem, P. Vemuri, D. Jones, K. Kantarci, C. Ward, R. A. Koeppe, N. Foster, E. M. Reiman, K. Chen, C. Mathis, S. Landau, N. J.

Cairns, E. Householder, L. Taylor-Reinwald, V. Lee, M. Korecka, M. Figurski, K. Crawford, S. Neu, T. M. Foroud, S. Potkin, L. Shen, K. Faber, S. Kim, L. Tha, R. Frank, J. Hsiao, J. Kaye, J. Quinn, L. Silbert, B. Lind, R. Carter, S. Dolen, B. Ances, M. Carroll, M. L. Creech, E. Franklin, M. A. Mintun, S. Schneider, A. Oliver, L. S. Schneider, S. Pawluczyk, M. Beccera, L. Teodoro, B. M. Spann, J. Brewer, H. Vanderswag, A. Fleisher, D. Marson, R. Griffith, D. Clark, D. Geldmacher, J. Brockington *et al.*, "Predicting alzheimer's disease progression using multi-modal deep learning approach," *Scientific Reports*, vol. 9, no. 1, 2019. [Online]. Available: https://doi.org/10.1038/s41598-018-37769-z

[117] N. Franzmeier, N. Koutsouleris, T. Benzinger, A. Goate, C. M. Karch, A. M. Fagan, E. McDade, M. Duering, M. Dichgans, J. Levin, B. A. Gordon, Y. Y. Lim, C. L. Masters, M. Rossor, N. C. Fox, A. O'Connor, J. Chhatwal, S. Salloway, A. Danek, J. Hassenstab, P. R. Schofield, J. C. Morris, R. J. Bateman, M. Ewers, i. the Alzheimer's disease neuroimaging, and N. the Dominantly Inherited Alzheimer, "Predicting sporadic alzheimer's disease progression via inherited alzheimer's disease-informed machine-learning," *Alzheimer's and Dementia*, vol. 16, no. 3, pp. 501–511, 2020. [Online]. Available: https://doi.org/10.1002/alz.12032

[118] Y. Zhou, Z. Song, X. Han, H. Li, and X. Tang, "Prediction of alzheimer's disease progression based on magnetic resonance imaging," *ACS Chemical Neuroscience*, vol. 12, no. 22, pp. 4209–4223, 2021. [Online]. Available: https://doi.org/10.1021/acschemneuro.1c00472

[119] J. K. Weaver, K. Milford, M. Rickard, J. Logan, L. Erdman, B. Viteri, N. D'Souza, A. Cucchiara, M. Skreta, D. Keefe, S. Shah, A. Selman, K. Fischer, D. A. Weiss, C. J. Long, A. Lorenzo, Y. Fan, and G. E. Tasian, "Deep learning imaging features derived from kidney ultrasounds predict chronic kidney disease progression in children with posterior urethral valves," *Pediatric Nephrology*, 2022. [Online]. Available: https://doi.org/10.1007/s00467-022-05677-0

[120] Z. Hu, Z. Wang, Y. Jin, and W. Hou, "Vgg-tswinformer: Transformer-based deep learning model for early alzheimer's disease prediction," *Computer Methods and Programs in Biomedicine*, vol. 229, 2023. [Online]. Available: https://doi.org/10.1016/j.cmpb.2022.107291

Appendix A

Imputation method comparison results

Seed	**Imputation type**	**Fold 1**	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**	**Average**	**Std. dev.**
787	Population mean	187.13	188.88	212.74	221.23	235.60	209.12	18.73
787	Event mean	199.37	216.98	217.98	290.36	235.04	231.95	31.31
787	Forward/backward filling	180.94	213.84	232.54	*308.84*	290.86	245.40	47.77
1998	Population mean	173.82	263.44	209.38	*448.41*	212.91	261.59	97.68
1998	Event mean	179.43	347.52	215.52	165.40	172.64	216.10	67.93
1998	Forward/backward filling	222.31	253.87	250.84	237.19	194.40	231.72	21.76
959143	Population mean	185.23	197.35	**139.20**	163.05	212.67	179.50	25.87
959143	Event mean	169.94	164.26	**137.44**	170.36	252.94	178.99	38.91
959143	Forward/backward filling	174.85	157.14	143.11	**130.08**	181.54	157.34	19.17
25	Population mean	274.09	265.23	304.78	372.65	301.4	303.63	37.73
25	Event mean	312.20	289.64	330.55	325.54	*370.75*	325.74	26.59
25	Forward/backward filling	255.71	249.09	237.76	287.00	265.50	259.01	16.65

Table A.1: The predictive performance results for the PPMI dataset, reported in MSE. Bold indicates the best score for the imputation type across all folds and seeds, italic indicates the worst score.

Seed	**Imputation type**	**Fold 1**	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**	**Average**	**Std. dev.**
787	Population mean	42.71	**41.34**	44.12	42.55	41.98	42.54	0.93
787	Event mean	42.81	**41.15**	44.35	42.07	42.48	42.57	1.05
787	Forward/backward filling	42.52	41.83	44.41	**41.63**	42.19	42.52	0.99
1998	Population mean	46.56	45.74	46.35	45.02	45.41	45.82	0.57
1998	Event mean	46.57	45.13	45.97	45.10	44.83	45.52	0.65
1998	Forward/backward filling	45.98	44.60	44.80	43.70	44.65	44.75	0.73
959143	Population mean	41.89	42.24	42.17	42.00	45.48	42.76	1.37
959143	Event mean	41.59	42.40	41.49	42.02	44.96	42.49	1.28
959143	Forward/backward filling	42.39	43.23	42.38	42.64	41.77	42.48	0.47
25	Population mean	45.07	45.14	45.42	*47.08*	43.94	45.33	1.01
25	Event mean	46.12	45.28	45.56	*47.55*	44.58	45.82	1.00
25	Forward/backward filling	*47.09*	45.05	46.74	47.08	44.11	46.01	1.21

Table A.2: The predictive performance results for the diabetes dataset, reported in MSE. Bold indicates the best score for the imputation type across all folds and seeds, italic indicates the worst score.

Appendix B

PPMI case study feature correlations

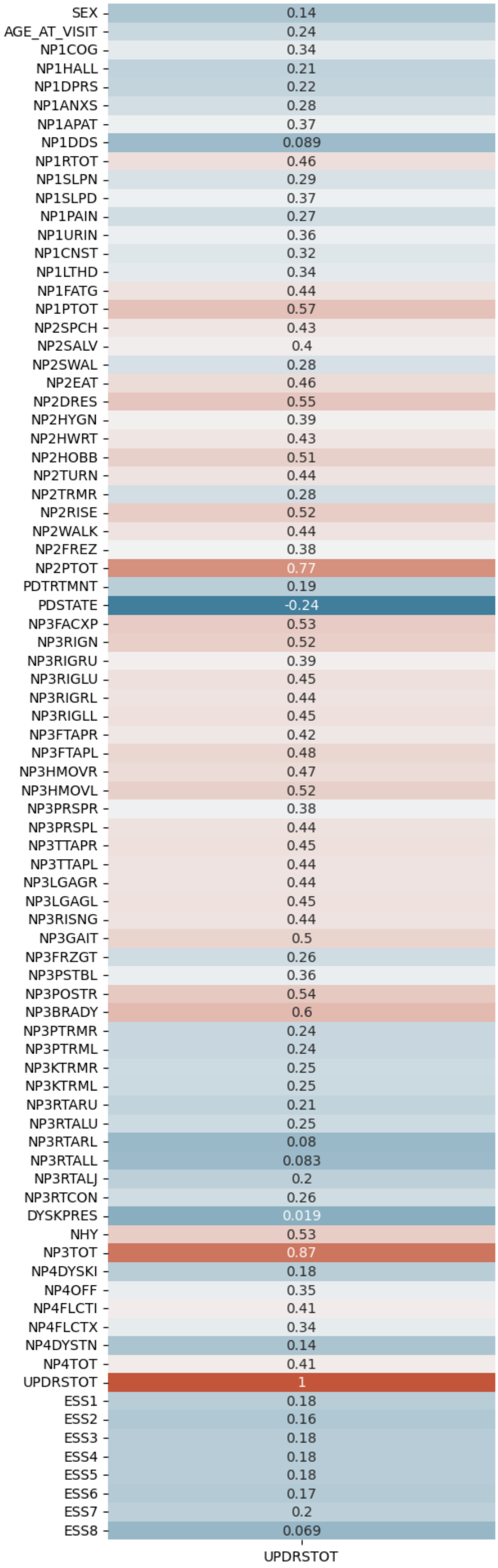


FIGURE B.1: A plot of the correlation between the PPMI features and the target UPDRS total score.

Appendix C

Literature review

This review is aimed at identifying the state-of-the-art in machine learning for disease detection and prediction of disease progression, thus answering SRQ1 as defined in Chapter 1. It identifies various types of predictions and models, as well as various challenges related to them. The review is performed following the Kitchenham methodology [56]. Following this methodology, a review must specify a review protocol. This contains the research questions and review methods. On top of this, a defined search strategy must be used and documented, to allow readers to reproduce and assess the search. To select studies to be reviewed, inclusion and exclusion criteria must be formulated. Finally, what information is to be extracted from each paper must be defined.

The software used to perform this review is Covidence[1], a web-based collaboration software platform that streamlines the production of systematic and other literature reviews.

C.1 Research questions and search strategy

The research questions formulated for this literature review are the following:

RQ1 What are current trends and state-of-the-art in machine learning for disease detection and prediction of disease progression?

SRQ1 What machine learning techniques are used in the field of disease detection?

SRQ2 What machine learning techniques are used in the field of prediction of disease progression over time?

SRQ3 What techniques are used to mitigate common issues in healthcare data quality?

We choose to use Scopus, PubMed, and WebOfScience as source libraries for this literature review. Scopus is the largest indexer of global research content that is mostly focused on technical fields, PubMed is a large knowledge base for medical articles, and Web of Science is a middle ground between the two aforementioned libraries. We perform the searches on the University of Twente network, meaning that institutional access can be used to access all sources available to the university. The search term combinations that we formulate for this are the following:
("Machine learning" OR "deep learning" OR "data science") AND ("disease prediction" OR "disease detection") AND ("biomarker" OR "clinical data")
("Machine learning" OR "deep learning" OR "data science") AND ("disease progression")

[1] Covidence systematic review software, Veritas Health Innovation, Melbourne, Australia. Available at www.covidence.org

AND "prediction" AND ("biomarker" OR "clinical data")
By using this combination, both literature focusing on the detection of diseases, as well as literature concerning the prediction of disease progression over time is included in the results. These queries return a total of 802 results across all three libraries.

C.2 Study selection and data extraction

To be able to filter for literature that is relevant to the aim of this review, we define a number of inclusion and exclusion criteria. Inclusion criteria serve to ensure that papers with high relevance to the research questions are included in the resulting set of literature. Exclusion criteria serve to remove papers that do not add value to answering the research questions.

C.2.1 Inclusion criteria

IC1: The paper directly addresses one or more research questions for this review
IC2: The paper concerns the use of machine learning to analyse biomarker or clinical information to detect diseases or predict the progression of diseases in patients

C.2.2 Exclusion criteria

EC1: The paper cannot be accessed via the internet
EC2: The paper is only published in a language other than English or Dutch
EC3: The paper length (excluding appendices) is 5 pages or less
EC4: The paper is not a peer reviewed paper
EC5: The paper does not directly address the concepts of machine learning and disease (progression) prediction, instead only referring to them as side topics or knowledge domains
EC6: The paper addresses SRQ1, but is not a secondary study

C.2.3 Study selection

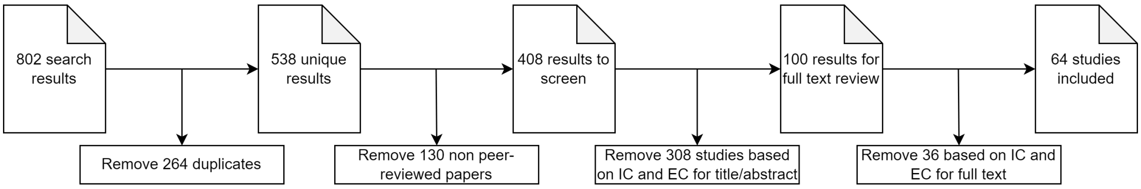

802 search results
538 unique results
408 results to screen
100 results for full text review
64 studies included
Remove 264 duplicates
Remove 130 non peer-reviewed papers
Remove 308 studies based on IC and EC for title/abstract
Remove 36 based on IC and EC for full text

FIGURE C.1: An overview of the literature selection process.

A visualisation of the selection process is shown in Figure C.1. The first step in selecting the studies to be included in the review is removal of duplicates that result from multiple searches across various source libraries. This leads to the removal of 264 studies, leaving 538 to be screened. This means approximately 33% of the results are duplicates.

The next step is removing results that were not peer reviewed papers. This brings the number of results down from 538 by 130, to a total of 408. As such, of the initial 538 non-duplicate results, circa 24% is not a peer reviewed paper.

Following removal of non-peer reviewed paper results, we screen the remaining studies based on title and abstract, using the inclusion and exclusion criteria mentioned above. This leads to the removal of 308 irrelevant studies, leaving exactly 100 studies to be assessed for eligibility based on the full text of the study. At this stage, we have excluded over 87% of the original 802 search results. Examples of irrelevant studies include prediction of patient response to medication [57] and proposals for Internet-of-Things based architectures to improve healthcare [58].

In the final selection step, we apply the inclusion and exclusion criteria to the full text of the 100 papers that remain. This process removes another 36 papers, meaning we used 64 papers for data extraction. This equates to approximately 8% of the original size of the results.

An overview of all papers involved in the last three steps of the selection process can be found on Github[2].

C.2.4 Data extraction

During the data extraction stage, we extract the following information from each paper: title, author, year of publication, and what topic it is relevant to. Aside from this information, we extract various fields regarding the dataset used, data quality and preparation, machine learning techniques used, performance metrics and achieved scores if applicable.

C.3 Results

A full overview of the results of the search and selection process can be found in Appendix D. This includes the basic information for all papers for which data was extracted to complete this review. Insights in distribution of metadata of the papers, as well as findings from the contents will be discussed in this subsection. For an overview of the answers each paper provides to the various research questions, we refer to Appendix E.

C.3.1 Literature details

As can be seen in Figure C.2, by far the largest number of papers used in this review were published in 2021, with 28 papers out of the 64 papers included being published in that year. Aside from a single publication in 2009, all publications are from 2015 or newer. This is in line with the trend of increasing interest in the topic of using machine learning in disease detection and prediction of disease progression in recent years.

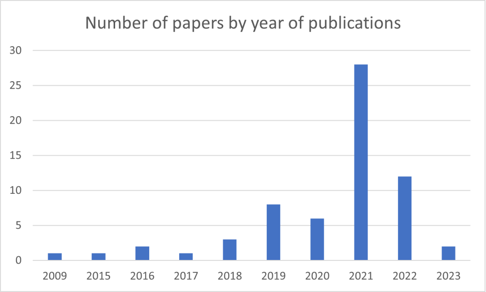


FIGURE C.2: The number of papers published per year.

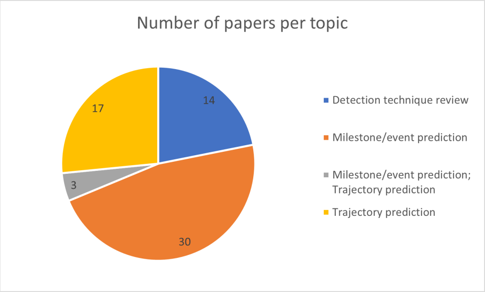


FIGURE C.3: The number of papers per topic.

[2] https://github.com/StijnBerendse/Literature-review

As our research regarding machine learning techniques is split between two research questions, papers were classified as being relevant to one of these. Papers relevant for SRQ1 were classified as "detection technique review". For SRQ2, a subdivision was made between papers on milestone/event prediction (time to conversion from current disease stage to a higher severity class), and papers on trajectory prediction. As visible in Figure C.3, most papers focus on milestone/event prediction. Three papers relevant to SRQ2 are classified as both milestone/event prediction and trajectory prediction, as multiple models were built, covering both topics.

C.3.2 Disease detection

We reviewed various secondary studies to find the most common and best performing machine learning techniques for disease detection. The area of application includes a wide variety of diseases, with the majority being aimed at detection of AD [59, 60, 61, 62, 63].

Interestingly, there seems to be a large overlap between techniques used, regardless of the disease that the model is being applied to. For example, every single study except the one by Su et al. [64] uses Support Vector Machine (SVM) or an extension of SVM capable of handling image data. The SVM (based) models are reported as the best-performing model by many studies [65, 66, 67, 63], while Marti-Juan et al. [59] and Kendrick et al. [43] mention SVM as the best alternative to deep learning methods in case some degree of interpretability is desired.

Another highly popular technique for disease detection is the use of random forest-based models. Out of the 14 total studies on disease detection, 9 mentioned the use of random forest models [65, 43, 66, 68, 67, 69, 61, 70, 71]. This makes the technique slightly less popular than SVM. Aside from the lower popularity, the performance of random forest based models is reported to be lower by all studies comparing SVM and random forest [65, 66, 67, 63, 43].

C.3.3 Prediction of disease progression

In this literature review, we identified two main types of progression prediction: trajectory prediction and milestone prediction. Trajectory prediction is based on predicting biomarkers, disease severity or other disease-related features over time. Milestone prediction is focused on predicting the occurrence of a single event, such as conversion from a lower disease severity level to a higher severity level, or the death of a patient. We first discuss milestone prediction, followed by trajectory prediction.

For milestone prediction, we found that there are two main types of predictions: binary predictions (for example whether a patient will or will not progress to a more severe disease state) and predictions of event probabilities.

We found that the latter type of prediction is often performed with a Cox Proportional Hazards model, often combined with Kaplan-Meier curves for visualisation. The Cox Proportional Hazards model is based on a regression that takes into account the effect of covariates on the survival chance of a subject, first introduced by Cox [72]. Nazha et al. [73] and Chen et al. [74] used the model for its most basic purpose: to predict survival of patients in multiple sclerosis and COVID-19 respectively. As mentioned however, milestone prediction may also substitute survival for some disease event. In most of these cases, this means predicting probability of conversion from a healthy state to a disease state [75, 76], but interestingly we also found a single case in which this was the other way around. Eng et al. [77] used a Cox Proportional Hazards model to predict the probability over time of disease becoming dormant.

Binary predictions are often made using techniques more closely related to disease detection. Out of 27 papers, we found that random forest based techniques are used in 10 papers and SVM based techniques in 8. In three papers, SVM and random forest models are compared [78, 79, 80], with SVM outperforming random forests in both papers that mention performance scores.

Another interesting finding is that most of the aforementioned papers use single data timepoints. Papers using longitudinal data are quite rare for this type of prediction, and tend to use more complicated models, more specifically models with some form of internal recurrence. Examples of such models include Recurrent Neural Network (RNN) based models and its various subtypes. A popular RNN subtype is Long Short Term Memory, or LSTM for short. We found that four papers used LSTM to perform their prediction task [81, 79, 82, 80]. Two papers used LSTM in conjunction with another model. Chen et al. [83] used LSTM in conjunction with an autoencoder model to handle the longitudinal nature of their dataset. An et al. [84] combine LSTM with a convolutional neural network (CNN) to create an LSTM-CNN model. This model first includes an LSTM layer to handle longitudinal data, followed by a convolutional layer for further classification based on the LSTM output. A different subtype of RNN, is a Gated Recurrent Unit (GRU) based model. De Brouwer et al. [51] use this architecture to propose a time-aware model for predicting whether multiple sclerosis patients will progress.

The second type of prediction that we identified is trajectory prediction. This prediction type is based upon predicting the exact condition of patients at a certain timepoint or over time. In the trajectory prediction domain, we find that simpler models such as SVM and random forests are much less popular. Out of all 20 papers in which trajectory prediction is performed, only 3 used SVM or SVM based models [85, 86, 10], the same number as random forest based models [87, 88, 10]. However, a new type of architecture does surface: two-step systems. Dadu et al. [87], Bhagwat et al. [89] and Sadiq et al. [90] propose systems that are based upon the concept of using an initial step to explicitly create patient subgroups that form trajectory clusters or phenotypes, followed by a second step that assigns patients to one of these subgroups. Bhagwat et al. [89] explicitly mention a limitation of these systems, stating that computational complexity is the main reason for the decision to limit the number of phenotypes to only two. Another interesting approach to this is taken by Schulam and Saria [91], who use population baselines to generate a very general trajectory and then manipulate it using patient specific data to create a personalised trajectory prediction.

Similar to the milestone prediction domain, papers using longitudinal data generally move towards more complex models such as the aforementioned RNNs and its subtypes [92, 93, 94, 80].

C.3.4 Data quality issues

In this review, we also identified several problems concerning the data quality in the healthcare domain.

Many papers mention data sparsity, or missing data, as a problem that must be solved: (biomarker) measurements taken are not necessarily consistent throughout various visits, even for the same patient. A large number of these papers mention the use of a form of imputation to mitigate the problem [59, 95, 96, 86, 10, 81, 97, 74, 98, 94, 82, 80]. Imputation techniques vary from longitudinal imputation [87, 95, 94] to cluster-based techniques [80].

However, there are also papers that explicitly avoid imputation to prevent a large drawback: bias resulting from filled values that are non-representative of the truth [83]. Instead, the authors of these papers opt for signaling missing values in a separate vector [99], or performing their experiments both with and without imputation to compare results [88]. A completely different approach to the sparsity problem is taken by Larie et al. [92] and Schmidt-Erfurth et al. [75]. They opted to avoid the issue, using a synthetic dataset and a dataset recorded for their research respectively. Another unique approach was used by Schulam and Saria [91], who used the population average-based nature of their model to mitigate data sparsity to some extent.

Finally, we found various papers that used exclusion of features or data points as an approach to mitigate data sparsity. For example, Dadu et al. [87] excluded patients that did not have complete followup data (not reaching the required number of recorded visits), as well as features that had $> 5\%$ missingness across the dataset population. A similar approach was taken by El-Sappagh et al. [80] and Sun et al. [95]. While this approach has no risk of incorrectly imputed data skewing results, exclusion based mitigation does potentially remove relevant data from a dataset, which may reduce predictive performance.

Another challenge encountered in papers included in this review is temporal irregularity. By this, we mean irregularity in when samples are taken from patients. For example, patient A may have bimonthly samples between t_0 and t_8, and a final sample from t_{11}, while patient B has samples taken at regular three month intervals between t_0 and t_{12}, where t_m denotes the time into the trial in months (see figure C.4). This challenge is exclusive to papers that use longitudinal data rather than single time-points to predict disease progression, but papers comparing these two approaches show an advantage in using longitudinal data over using a single time-point in prediction performance [89]. As such, it is important that this challenge can be solved to achieve optimal performance.

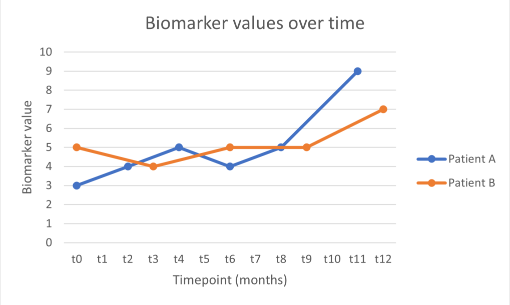


FIGURE C.4: An example of temporal irregularity in biomarker sampling.

Various handling methods are proposed, but they can be divided in two main categories. The first is handling by creating a robust model and the second is handling by pre-processing. This is supported by Marti-Juan et al. [59], who state that "temporal alignment" can be approached in two ways: pre-processing for a specific progression model and approaching this as a standalone problem. The first of these generally uses temporal characteristics to aid in integrating some form of decay function [84, 82, 81, 51, 94] in the proposed model. In these cases, the pre-processing is done for the specific model that is being developed. The latter often concerns some form of weighting time-points based on temporal characteristics [95, 100] or including temporal characteristics as a (normalized) feature to be interpreted by the model [85, 97]. These approaches are more model-agnostic, which fits the standalone problem concept.

A different approach for solving a problem related to temporal irregularity was taken by Chen et al. [83], who discarded "extra" visits between the first ("baseline") and last visit. This ensured each patient had the same number of visits to train with, but comes at the cost of losing the additional information that including all visits may offer.

Similar to data sparsity issues, Larie et al. [92] and Schmidt-Erfurth et al. [75] circumvent temporal irregularity issues altogether due to their choice of dataset.

The last data quality problem we identified is class imbalance. Various papers mention this issue, with the large majority choosing to under- or oversample data to synthetically balance the data, such as Velazques and Lee [101] and Papaiz et al. [10]. Both Dinov et al. [86] and El-Sappagh et al. [80] use SMOTE for this same purpose. A related, but not pre-processing based technique is used by Shafiee et al. [102], who use a balanced random forest technique. This technique performs undersampling internally to balance input data.

A different approach to this problem is taken by Singh et al. [103], who include a precision-recall area under curve metric in their evaluation, to highlight possible resulting bias from the data imbalance on the predictions.

C.4 Discussion

C.4.1 Disease detection

From the literature, we can see that the disease detection field is quite mature: there is a wide variety of techniques that perform well, at a relatively low complexity and thus with a decent level of transparency in how the model comes to its prediction. SVMs are the main model type for this field, offering near deep-learning level performance at a fraction of the required input data and computational cost. However, we note that studies which review (deep learning) techniques for longitudinal data such as the ones by [64] and [62] do not include SVM or do not report on the performance of SVM respectively. Thus, a direct comparison between SVM and deep learning techniques cannot be made for longitudinal data interpretation.

C.4.2 Prediction of disease progression

For progression prediction, we see two data categories, and two prediction categories. The data categories are single timepoints and longitudinal data. Longitudinal data offers higher performance over longer timespans due to the possibility of using temporal relations between data points, but requires more input data. This is because a single patient creates a single longitudinal training datapoint. Single timepoint based systems on the other hand, can use each patient visit as a separate training datapoint. To achieve the best possible performance, models inherently capable of handling such longitudinal data series should be developed further. This is because chronic disease progression often spans a period up to several years, which is better suited to longitudinal data based systems. Well-performing examples of such models are Recurrent Neural Network based models, which inherently perform well on sequential data due to the ability to feed outputs from a layer back into its input. This allows the model to discover patterns in sequential data.

The two prediction categories are milestone predictions versus trajectory predictions. While milestone predictions largely share their pool of techniques with disease detection techniques, they may offer less value for supporting clinical decision making. The predictions are mostly binary, only stating whether a certain event will take place, but not when or to what extent. Trajectory predictions may offer this additional information, especially being more specific regarding the severity over time. One of the drawbacks of trajectory prediction, is that a good proxy for disease severity must be found. For some diseases this can be a measurable datapoint, such as the forced expiratory volume functioning as a proxy for pulmonary exacerbation in cystic fibrosis patients [104]. In other cases, such as AD, a disease severity is assigned based on tests that result in a cognitive score [105].

C.4.3 Data quality

Data quality issues found in this review are similar to the common problems identified in Chapter 1. From the literature included in this review, we find that data quality is a significant issue that ranges from class imbalance to temporal alignment. Furthermore, it is approached very differently across proposed systems, even if the same dataset is used. This is partially problematic because not all methods are equally performant, but also because it makes comparison between proposed models more difficult. Two models trained using the same dataset may have a performance difference in part due to the variation in pre-processing performed on the data, rather than the performance of the models themselves. The development of models that are robust to some of these data quality issues is promising, but further research in the direction of this would be advantageous to aid in the development of and comparison between prediction models.

C.5 Conclusion

From this review we have learned what techniques are used for disease detection and prediction of disease progression, as well as mitigation of data quality issues in the healthcare domain. We find that SVMs are the main model type for disease detection as well as milestone/event prediction. For prediction of disease trajectory, we find that more complex models such as RNNs are popular. This type of prediction and the associated models would be most useful for supporting clinical decision making at an early stage in (chronic) diseases, due to the added value of using longitudinal data and more fine-grained predictions. The development of models robust to common data quality issues in healthcare is promising, but further research in the direction of this would be advantageous to aid in the development of and comparison between prediction models. To conclude, opportunities for further research aimed at predictive performance lie in predicting disease trajectories using longitudinal data, and improving mitigation of data quality issues.

Appendix D

Literature overview

Table A.1: An overview of all papers used for data extraction.

ID	Title	Author(s)	Ref	Year	Relevance
1	Integrative Analysis using Coupled Latent Variable Models for Individualizing Prognoses	Schulam and Saria	[91]	2016	Trajectory prediction
2	Early detection and risk assessment for chronic disease with irregular longitudinal data analysis	He et al.	[85]	2019	Trajectory prediction
3	A survey on machine and statistical learning for longitudinal analysis of neuroimaging data in Alzheimer's disease	Marti-Juan et al.	[59]	2020	Detection technique review
4	The role of machine learning in developing non-magnetic resonance imaging based biomarkers for multiple sclerosis: a systematic review	Hossain et al.	[65]	2022	Detection technique review
5	Flexible link functions in a joint hierarchical Gaussian process model	Su et al.	[104]	2020	Trajectory prediction
6	Identification and prediction of Parkinson's disease subtypes and progression using machine learning in two cohorts	Dadu et al.	[87]	2022	Trajectory prediction
7	Modelling prognostic trajectories of cognitive decline due to Alzheimer's disease	Giorgio et al.	[106]	2020	Milestone/event prediction
8	Patterns of joint involvement in juvenile idiopathic arthritis and prediction of disease course: A prospective study with multilayer non-negative matrix factorization	Eng et al.	[77]	2019	Milestone/event prediction

Table A.1 – continued from previous page

ID	Title	Author(s)	Ref	Year	Relevance
9	Automatic Prediction of Cognitive and Functional Decline Can Significantly Decrease the Number of Subjects Required for Clinical Trials in Early Alzheimer's Disease	Shafiee et al	[102]	2021	Milestone/event prediction
10	NfL predicts relapse-free progression in a longitudinal multiple sclerosis cohort study	Uphaus et al.	[107]	2021	Milestone/event prediction
11	Personalized Prediction Model to Risk Stratify Patients With Myelodysplastic Syndromes	Nazha et al.	[73]	2021	Milestone/event prediction
12	Prognostic model for multiple myeloma progression integrating gene expression and clinical features	Sun et al.	[95]	2019	Milestone/event prediction
13	The use of Artificial Neural Networks to Forecast the Behavior of Agent-Based Models of Pathophysiology: An Example Utilizing an Agent-Based Model of Sepsis	Larie et al.	[92]	2021	Trajectory prediction
14	Prediction of Individual Disease Conversion in Early AMD Using Artificial Intelligence	Schmidt-Erfurth et al.	[75]	2018	Milestone/event prediction
15	Correlation-Aware Sparse and Low-Rank onstrained Multi-Task Learning for Longitudinal Analysis of Alzheimer's Disease	Jiang et al.	[105]	2019	Trajectory prediction
16	Derivation and validation of a machine learning risk score using biomarker and electronic patient data to predict progression of diabetic kidney disease	Chan et al.	[96]	2021	Milestone/event prediction
17	Predictive Big Data Analytics: A Study of Parkinson's Disease Using Large, Complex, Heterogeneous, Incongruent, Multi-Source and Incomplete Observations	Dinov et al.	[86]	2016	Trajectory prediction
18	Modelling Patient Trajectories Using Multimodal Information	Silva and Matos	[108]	2022	Milestone/event prediction
19	Modeling longitudinal imaging biomarkers with parametric Bayesian multi-task learning	Aksman et al.	[109]	2019	Trajectory prediction
20	Motor Progression in Early-Stage Parkinson's Disease: A Clinical Prediction Model and the Role of Cerebrospinal Fluid Biomarkers	Ma et al.	[88]	2021	Trajectory prediction

Table A.1 – continued from previous page

ID	Title	Author(s)	Ref	Year	Relevance
21	CT-Based Radiomics Analysis Before Thermal Ablation to Predict Local Tumor Progression for Colorectal Liver Metastases	Taghavi et al.	[76]	2021	Milestone/event prediction
22	Radiomic for Identification and Prediction in Metastatic Prostate Cancer: A Review of Studies	Kendrick et al.	[43]	2021	Detection technique review
23	Modeling and prediction of clinical symptom trajectories in Alzheimer's disease using longitudinal data	Bhagwat et al.	[89]	2018	Trajectory prediction
24	Multi-task exclusive relationship learning for alzheimer's disease progression prediction with longitudinal data	Wang et al.	[110]	2019	Milestone/event prediction; Trajectory prediction
25	Predicting the progression of Parkinson's disease using conventional MRI and machine learning: An application of radiomic biomarkers in whole-brain white matter	Shu et al.	[111]	2021	Milestone/event prediction
26	Group Guided Fused Laplacian Sparse Group Lasso for Modeling Alzheimer's Disease Progression	Liu et al	[99]	2020	Milestone/event prediction
27	Machine learning models to predict the progression from early to late stages of papillary renal cell carcinoma	Singh et al	[103]	2018	Milestone/event prediction
28	Risk Prediction of Cardiovascular Events by Exploration of Molecular Data with Explainable Artificial Intelligence	Westerlund et al.	[66]	2021	Detection technique review
29	Model-based stratification of progression along the Alzheimer disease continuum highlights the centrality of biomarker synergies	Sadiq et al.	[90]	2022	Trajectory prediction
30	Biomedical signals and machine learning in amyotrophic lateral sclerosis: a systematic review	Fernandes et al.	[68]	2021	Detection technique review
31	Longitudinal metabolomics of human plasma reveals prognostic markers of COVID-19 disease severity	Sindelar et al.	[112]	2021	Milestone/event prediction
32	Random forest model for feature-based Alzheimer's disease conversion prediction from early mild cognitive impairment subjects	Velazquez and Lee	[101]	2021	Milestone/event prediction
33	Machine Learning Solutions Applied to Amyotrophic Lateral Sclerosis Prognosis: A Review	Papaiz et al.	[10]	2022	Milestone/event prediction; Trajectory prediction

Table A.1 – continued from previous page

ID	Title	Author(s)	Ref	Year	Relevance
34	Personalized Prediction of Tumor Response and Cancer Progression on Prostate Needle Biopsy	Donovan et al.	[113]	2009	Milestone/event prediction
35	Multi-task deep autoencoder to predict Alzheimer's disease progression using temporal DNA methylation data in peripheral blood	Chen et al.	[83]	2022	Milestone/event prediction
36	Future activity prediction of multiple sclerosis with 3D MRI using 3D discrete wavelet transform	Acar et al.	[78]	2022	Milestone/event prediction
37	Predicting COVID-19 disease progression and patient outcomes based on temporal deep learning	Sun et al.	[81]	2021	Milestone/event prediction
38	Predicting disease progression from short biomarker series using expert advice algorithm	Morino et al.	[114]	2015	Trajectory prediction
39	Longitudinal machine learning modeling of MS patient trajectories improves predictions of disability progression	De Brouwer et al.	[51]	2021	Milestone/event prediction
40	Time-to-event prediction using survival analysis methods for Alzheimer's disease progression	Sharma et al.	[97]	2021	Milestone/event prediction
41	Exploring Longitudinal Cough, Breath, and Voice Data for COVID-19 Progression Prediction via Sequential Deep Learning: Model Development and Validation	Dang et al.	[93]	2022	Trajectory prediction
42	Machine learning of brain structural biomarkers for Alzheimer's disease (AD) diagnosis, prediction of disease progression, and amyloid beta deposition in the Japanese population	Shiino et al.	[115]	2021	Milestone/event prediction
43	Predicting Alzheimer's disease progression using multi-modal deep learning approach	Lee et al.	[116]	2019	Milestone/event prediction
44	A predictive paradigm for COVID-19 prognosis based on the longitudinal measure of biomarkers	Chen et al.	[74]	2021	Milestone/event prediction
45	Predicting sporadic Alzheimer's disease progression via inherited Alzheimer's disease-informed machine-learning	Franzmeier et al.	[117]	2020	Trajectory prediction
46	Machine Learning Use for Prognostic Purposes in Multiple Sclerosis	Seccia et al.	[79]	2021	Milestone/event prediction

Table A.1 – continued from previous page

ID	Title	Author(s)	Ref	Year	Relevance
47	Prediction of Alzheimer's Disease Progression Based on Magnetic Resonance Imaging	Zhou et al.	[118]	2021	Milestone/event prediction
48	Healthcare predictive analytics for disease progression: a longitudinal data fusion approach	Zheng and Hu	[100]	2020	Trajectory prediction
49	Applied machine learning in Alzheimer's disease research: omics, imaging, and clinical data	Li et al.	[60]	2021	Detection technique review
50	Time-Aware Multi-Type Data Fusion Representation Learning Framework for Risk Prediction of Cardiovascular Diseases	An et al.	[84]	2022	Milestone/event prediction
51	Deep learning imaging features derived from kidney ultrasounds predict chronic kidney disease progression in children with posterior urethral valves	Weaver et al.	[119]	2022	Milestone/event prediction
52	On the time-varying predictive performance of longitudinal biomarkers: Measure and estimation	Zhang et al.	[98]	2021	Trajectory prediction
53	Deep recurrent model for individualized prediction of Alzheimer's disease progression	Jung et al.	[94]	2021	Trajectory prediction
54	Predicting progression of Alzheimer's disease using forward-to-backward bi-directional network with integrative imputation	Ho et al.	[82]	2022	Milestone/event prediction
55	Machine Learning and Data Mining Methods in Diabetes Research	Kavakiotis et al.	[67]	2017	Detection technique review
56	VGG-TSwinformer: Transformer-based deep learning model for early Alzheimer's disease prediction	Hu et al.	[120]	2023	Milestone/event prediction
57	Applications of Machine Learning in Human Microbiome Studies: A Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment	Marcos-Zambrano et al.	[69]	2021	Detection technique review
58	Single and Combined Neuroimaging Techniques for Alzheimer's Disease Detection	Amini et al.	[61]	2021	Detection technique review
59	Environmental exposures in machine learning and data mining approaches to diabetes etiology: A scoping review	Mistry et al.	[70]	2023	Detection technique review
60	Two-stage deep learning model for Alzheimer's disease detection and prediction of the mild cognitive impairment time	El-Sappagh et al.	[80]	2022	Milestone/event prediction; Trajectory prediction

Table A.1 – continued from previous page

ID	Title	Author(s)	Ref	Year	Relevance
61	Computational Techniques and Tools for Omics Data Analysis: State-of-the-Art, Challenges, and Future Directions	Kaur et al.	[71]	2021	Detection technique review
62	Deep learning in mental health outcome research: a scoping review	Su et al.	[64]	2020	Detection technique review
63	Deep learning for Alzheimer prediction using brain biomarkers	Goenka and Tiwari	[62]	2021	Detection technique review
64	Transfer Learning for Alzheimer's Disease through Neuroimaging Biomarkers: A Systematic Review	Agarwal et al.	[63]	2021	Detection technique review

Appendix E

Answers to literature review research questions

Table A.1: An overview of the answers to research questions provided per paper

ID	SRQ1 answer	SRQ2 answer	SRQ3 answer
1		Various models proposed and tested, such as the latent trajectory model and various B-spline variations. The final proposed method is a coupled latent trajectory model, which is an extension of the latent trajectory model and is based on statistical analysis of multiple biomarkers and correlations between the individual biomarker trajectories. This predicts a continuous trajectory, rather than values at set points in time.	Time irregularity and data sparsity is tackled by developing a model that updates posterior probabilities based on new data. The "baseline" trajectory is based on various averages (such as a population component), which is then fitted to the individual based on the available individual data.

Table A.1 – continued from previous page

ID	SRQ1 answer	SRQ2 answer	SRQ3 answer
2		This paper proposes EDRA, a flexible mixed-kernel method, to predict onset of type 1 diabetes. This is based on Structured Output SVM, which it extends to be usable with longitudinal data. EDRA is compared to regular Structured Output SVM and Max-Margin Early Event Detectors. EDRA performed best.	To handle time irregularity, normalisation is performed so temporal features can be included in a feature matrix.
3	Techniques mentioned include SVM, least squares regression, LR, deep learning, multi-task learning, multiple kernel learning, and LDA. Best performing techniques are SVM and deep learning, with deep learning being more scalable and SVM being more light-weight and interpretable.		The term "temporal alignment" between subjects is mentioned, and split in two variations for approaching this problem: performing preprocessing for a specific progression model, and approaching it as a standalone problem. Data sparsity/missing data is also mentioned, along with various handling options. These include imputing missing data, removing incomplete records, ignoring the issue (strongly advised against), and using a machine learning method robust to the problem.

Table A.1 – continued from previous page

ID	SRQ1 answer	SRQ2 answer	SRQ3 answer
4	Techniques mentioned: Online Sequential Extreme Learning Machine, LR, RF, KNN, SVM, DT, NN, Naive Bayes, Natural Language Processing, Lasso, Generalized Linear Model, AdaBoost, Self-Organizing Map. SVM performed best across various mentioned papers.		
5		This paper proposes the Joint Hierarchical Gaussian Process Model with Flexible Link Functions. The model is aimed at predicting the Forced Expiratory Volume as a direct proxy for future pulmonary exacerbation.	
6		The proposed system consists of two steps: clustering to obtain progression subtypes, followed by prediction what subtype a patient belongs to. Step one is done using non-negative matrix factorization and gaussian mixture models, step two using a stacking ensemble of RF, LightGBM and XGBoost.	To handle data sparsity/missing data: patients without complete followup data (missing timepoints) were excluded. Features with high missingnes were excluded. Longitudinal imputation was performed to tackle missing data points (provided <5% missingness).

Table A.1 – continued from previous page

ID	SRQ1 answer	SRQ2 answer	SRQ3 answer
7		The paper is focused on predicting whether a patient with mild cognitive impairment will progress to a more severe disease stage by determining whether their condition is stable (sMCI) or progressive (pMCI). The proposed technique is Generalised Metric Learning Vector Quantization, which extends the Learning Vector Quantization concept by including a full metric tensor to improve the distance measure.	
8		Only a small section of the paper on progression of the disease. This section mentions the use of Cox Proportional Hazards to model survival time. The survival time is a "reverse" approach: the time to zero is not time to disease stage progression, but rather time to disease inactivity.	
9		Balanced RF method is used to predict whether patients belong to a stable or progressive patient group. Image data is used indirectly by first extracting to nominal data to be usable with the chosen technique.	To combat class imbalance between ground truth stable and progressive patients, a balanced random forest method is used, rather than a regular random forest.
10		This study is aimed at predicting whether a patient progresses to next stage MS based on disability severity score. Binary predictions on this are made using SVM.	

Table A.1 – continued from previous page

ID	SRQ1 answer	SRQ2 answer	SRQ3 answer
11		A Cox Proportional Hazards model is fitted to predict survival time. Probability over time is provided as well.	
12		This study compares the performance of GuanRank Gaussian Process Regression Model, Cox Proportional Hazards, Survival RF, and Gaussian Process Regression for predicting survival in multiple myeloma patients. The proposed technique is the GuanRank Gaussian Process Regression Model. The proposed system outperformed the other techniques.	Various problems were mentioned: time irregularity, data sparsity, and data heterogeneity. To combat sparsity, missing values were imputed using the patient mean. Data heterogeneity was mitigated by normalizing the data as part of preprocessing. Finally, the proposed GuanRank concept was used to mitigate time irregularity by weighting datapoints based on their temporal characteristics.
13		Predicting trajectory of cytokine values over time. This research used a Long Short Term Memory based Recurrent Neural Network to be able to use longitudinal data for predicting the trajectory of biomarkers. LSTM based systems are well suited for capturing information from time series data.	The challenge of heterogeneity in healthcare data was mentioned in this research, but rather than solving this issue for a real world dataset, the challenge was avoided altogether by using synthetic data for training and testing.
14		Prediction whether conversion to disease will take place, as well as the probability of the event over time. Predictions performed using a sparse Cox Proportional Hazards model with LASSO regularisation.	No challenges were encountered, because the dataset used in this paper was specifically for this purpose. Time intervals between check-ups were set regularly, as were the biomarkers measured at each visit.

Table A.1 – continued from previous page

ID	SRQ1 answer	SRQ2 answer	SRQ3 answer
15		Predicting cognitive scores over time using ordinal features extracted from imaging features. Various techniques tested: LASSO, Group LASSO, Low Rank, Temporal Group LASSO, Convex Fused Sparse Group LASSO, Matrix Similarity, and the proposed model: Correlation-aware sparse and Low rank Constrained multi-task learning (CSL). CSL outperformed all other models in this paper.	
16		Use of an RF model to predict progression of diabetic kidney disease based on an internally developed risk score. Probability of progression at t+5 years was provided.	Data sparsity problems were handled using statistical imputation.
17		Forecasting whether patients will develop Parkinson's disease, comparing the performances of SVM and AdaBoost. Best performance was achieved using an AdaBoost model.	Imputation (MICE) done to mitigate sparsity, but also heterogeneity by replacing outliers. R SMOTE was used to mitigate imbalance of classes.

Table A.1 – continued from previous page

ID	SRQ1 answer	SRQ2 answer	SRQ3 answer
18		This paper is not truly focused on progression of a patient, but instead on predicting whether a patient will be readmitted and what their diagnosis will be. This is done using a Neural Network adaptation of the existing SAPBert and ClinicalBERT systems. One of the only papers that use natural text as a data modality.	
19		Modeling and visualising longitudinal trajectories (using annualized rate of change) of various AD-related biomarkers in various regions of the brain. Model based on Parametric Bayesian multi-task learning.	
20		Prediction of motor function in t+1 year, comparing performance of various techniques. These include Linear Regression, Ridge Regression, Bayesian Regression, RF, and Gradient Boosting DT. No model outperformed the others with statistical significance.	Samples where disease severity score decreased over time were discarded due to assumption of increasing severity. Biomarker sparsity influence was investigated by splitting the research in two versions, one with only motor scores (which were complete), and one with both biomarkers and motor scores (containing the sparse features).
21		Predicting probability of the return of cancer tumors over time using a Cox Proportional Hazards model.	

Table A.1 – continued from previous page

ID	SRQ1 answer	SRQ2 answer	SRQ3 answer
22	This paper is on radiomics, so the data is mostly image based. Detection techniques mentioned are SVM, LR, kNN, DT, RF, and CNN. CNNs have the best performance of these, but offer very low interpretability and require a large amount of data to train. Alternatively if this is not desirable, SVM offers slightly worse performance, but without these drawbacks.		
23		The aim of this paper is to address issues regarding longitudinal data use for trajectory prediction in AD. The added value of using longitudinal data as input is tested as well. The methodology is essentially modeling a set of trajectory patterns based on ground truths and classifying new patients as part of one of these patterns. This initial work includes only 2 and 3 trajectory patterns for two different cognitive scores respectively, due to computational complexity. The proposed technique for this is LSN (Longitudinal Siamese Network).	

Table A.1 – continued from previous page

ID	SRQ1 answer	SRQ2 answer	SRQ3 answer
24		This paper covers both prediction by classifying as one of two progression categories, but also trajectory prediction of separate biomarkers for Alzheimer's Disease. Various techniques were tested against the proposed system: LASSO, Multi-Task Feature Learning, Multi-Task Exclusive LASSO, Multi-Task Relationship Learning, Multi-Task Exclusive Relationship Learning. The last of these was the proposed system, which outperformed all other systems.	
25		This paper is aimed at detection whether a patient is at high risk of progression of Parkinson's disease. SVM, NB, kNN, and DT were used to determine predictors from MRI data.	
26		Various methods are used to predict cognitive scores (MMSE, ADAS) in 6 to 12 month time steps. Methods used are Ridge Regression, LASSO, Temporal Group LASSO, Convex Fused Sparse Group LASSO, and the proposed system: Group Guided Fused Laplacian Sparse Group LASSO. The best performing model was the proposed system.	To combat sparsity bias, a separate signal is included to indicate incomplete target vectors.

Table A.1 – continued from previous page

ID	SRQ1 answer	SRQ2 answer	SRQ3 answer
27		The proposed system consists of varSelRF feature selection combined with a Shrunken Centroid classifier to predict the progression of papillary renal cell carcinoma.	To mitigate the influence of class imbalance, the precision-recall AUC metric is included for evaluation of the model.
28	Paper aimed at detecting cardiovascular disease in patients based on clinical and demographical data. The techniques mentioned for this are CNN, LR, SVM, RF, AdaBoost, and MLP. SVM narrowly outperforms CNN for detection of cardiovascular disease.		
29		Predicting whether AD patients will have fast or moderate decline in the future. A two step approach is proposed, with the first step being clustering to define progression phenotypes based on existing data using Dynamic Time Warping. The second step is prediction what phenotype a patient belongs to using Parameter-Efficient Network model (PENet).	
30	Detecting amyotrophic lateral sclerosis. Mentioned techniques are SVM, LDA, Quadratic Classifier, ANN, KNN, RF, MLP, and Factorial Hidden Markov Model. The best performance was achieved using Quadratic Classifier.		

Table A.1 – continued from previous page

ID	SRQ1 answer	SRQ2 answer	SRQ3 answer
31		Using Linear ElasticNet to predict what disease severity COVID patients will experience based on longitudinal biomarker data.	
32		An RF based system was proposed to predict whether conversion of Mild Cognitive Impairment to Alzheimer's Disease will occur.	Oversampling was used to balance the target classes.
33		This concerns a secondary study, which is split in prediction categories: disease progression (predicting changes in ALS functional scale rating), survival time (classifying in survival time group), and need for support (estimating need for ventilation support at various time points). For each category, various techniques were mentioned. For progression: Ordinal DTs, RF, boosting models, Bayesian Regression Tree, SVM, XGBoost, and SPADE. For survival: DNN, Gaussian Regression, and Uniform Manifold Approximation and Projection. For support: RF.	Imputation was mentioned to handle data sparsity. Aside from this, under/oversampling was mentioned to mitigate class imbalance.
34		A Censored Data Support Vector Regression model was proposed to predict "probability of clinical freedom" over time in patients following prostatectomy.	

Table A.1 – continued from previous page

ID	SRQ1 answer	SRQ2 answer	SRQ3 answer
35		This papers proposes two methods: a multi-task LSTM auto-encoder, and a multi-task convolutional auto-encoder. Both were used to predict conversion of disease stage in a binary way, without skipping disease stage steps (so stage 1 to 2, 2 to 3, not 1 to 3).	Issues with varying numbers of visits are handled by only using two: a "first" and "last" visit. Imputation of data was consciously avoided to prevent resulting bias.
36		Predicting whether new lesions will develop in Multiple Sclerosis patients based on MRI data. Methods used for this binary prediction are kNN, DT, SVM, RF, NB, and LR. The best performance was achieved using SVM.	
37		Predicting COVID-19 patient outcome based on longitudinal data. Longitudinal data is short term (circa 7 days). Various models were compared: Cox Proportional Hazards, kNN, SVM, DT, Back Propagation NN, Probabilistic NN, RNN, LSTM, and Time-aware LSTM. The last model outperformed all others at predicting patient outcome.	Data sparsity is handled by cherry picking features without missing data. Time irregularity is handled by adapting the LSTM model to be time aware. This is done by implementing a memory discount based on time gaps between measurements.

Table A.1 – continued from previous page

ID	SRQ1 answer	SRQ2 answer	SRQ3 answer
38		This paper is focused on predicting the trajectory inclusion of a biomarker. It is not exactly a prediction of the trajectory, but of a curve that the algorithm is X% certain (confidence interval) that the trajectory will fall under. Predictions are made using a proposed Temporal Expert Advice algorithm	
39		Several techniques are compared to a proposed system to predict whether disability score in Multiple Sclerosis patients increases to the point that a new disease severity stage is reached. The techniques used are: static RF, dynamic RF, Bayesian Probabilistic Tensor Factorisation (BPTF), BPTF with RF, Time-aware NN (GRU variant of RNNcell), and GRU-ODE-Bayes NN. The proposed system performed best.	The proposed system (GRU-ODE-Bayes) is time-aware to handle time irregularity.
40		A Cox Proportional Hazards model and a proposed Neural Multi-Task LR model were compared for their performance in predicting time-to-event for disease stage progression.	Imputation was used to handle data sparsity. To handle time irregularity, a duration defined as the time between two visits in which disease stage progressed was added as a feature.
41		This paper attempts to predict the trajectory of COVID-19 infections using longitudinal audio data. The proposed model is GRU based.	To handle data heterogeneity, longitudinal features are used to look at relative changes within patients rather than absolute values across patients.

Table A.1 – continued from previous page

ID	SRQ1 answer	SRQ2 answer	SRQ3 answer
42		Predicting probability of conversion to Alzheimer's Disease over a 3 year timespan using SVM with Voxel-Based Morphometry kernel.	
43		Use of GRU and LR to predict conversion from Mild Cognitive Impairment to Alzheimer's Disease. GRU is used to interpret longitudinal data from various modalities, the output of which is used for classification using LR.	
44		Two-step system for predicting COVID-19 survival chance. HTREE used for feature selection, followed by Cox Proportional Hazards model to determine survival chance over time.	Missing data imputed using MICE.
45		Use of Support Vector Regression to predict rate of cognitive decline in patients that are genetically at risk of developing Alzheimer's Disease.	Data normalized to combat heterogeneity.

Table A.1 – continued from previous page

ID	SRQ1 answer	SRQ2 answer	SRQ3 answer
46		A secondary study on predicting progression in Multiple Sclerosis. Various techniques are mentioned to be used for this use case: SVM, kNN, LR, DT, RF, and CNN. A conclusion on best performing models is not drawn. Advantages of LSTM based models in performance are mentioned, but a drawback in available data by concatenating data points into time series rather than using the separate data points is also highlighted.	
47		A secondary study that mostly includes papers that propose systems for predicting Mild Cognitive Impairment to Alzheimer's Disease conversion using MRI data. Various SVM based methods, as well as various CNN based methods are included in the review.	
48		Predicting cognitive scores to forecast Alzheimer's Disease progression. Two techniques used: the proposed system named Disease Progression via Longitudinal Data Fusion with Accelerated Gradient Descent (DPLDF) and LR-Lasso. DPLDF performed best.	Time decay mitigation is performed by doing temporal regularization to weight older vs more recent data points, but this is done using order of points rather than the actual time difference.

Table A.1 – continued from previous page

ID	SRQ1 answer	SRQ2 answer	SRQ3 answer
49	Various techniques mentioned: SVM, MLP, Auto-Encoder, CNN, Bayesian Gaussian Process LR, Elastic Net Regularized LR, deep learning methods (Inception, others). The performance of deep learning methods performed best at detecting Alzheimer's Disease in patients.		
50		This paper proposes Time-Aware Multi-Type Data Fusion Representation (TAMDUR), which is based on combining bidirectional LSTM, CNN, and a self-attention mechanism, to predict development of cardiovascular disease in patients.	Time irregularity is mitigated using a time decay function, which weights visit data based on patient age and time intervals between visits.
51		In this study, ResNet-50 is used to extract image features from kidney ultrasounds. These features are then used in random survival forests to predict chronic kidney disease progression. Three model versions were tested: a random survival forest using only clinical data, one using only the image feature data, and an ensemble model. The latter outperformed the other two.	
52		Predicting the degree of gene expression using a proposed pseudo partial-likelihood model.	Imputation was used to fill missing time points and handle time irregularity to some degree. As such, a lot of data consisted of imputed values.

Table A.1 – continued from previous page

ID	SRQ1 answer	SRQ2 answer	SRQ3 answer
53		This paper proposes various LSTM variants to predict personalised trajectories of Alzheimer's Disease patients. The used techniques were the proposed LSTM with integrated time decay, LSTM-M (mean value imputation), LSTM-F (forward value imputation), Multi-directional RNN, Peephole LSTM-Z (zero imputation), MinimalRNN. The proposed system performed best.	Missing values were imputed based on surrounding longitudinal data and other features. Data heterogeneity was mitigated using normalisation. Temporal data is encoded using an integrated time decay factor in the proposed LSTM system.
54		In this paper, a multifeature aggregated LSTM model with progressive score is proposed to predict whether a patient is progressive or not by predicting future biomarker values.	Imputation of missing values is performed, along with a temporal decay weighting to handle time irregularity.
55	Review on detecting diabetes. Mention of SVM, RF, LR, LDA, NB, ANN, kNN, Multifactor Dimensionality Reduction, Classification and Regression Trees, and fuzzy c-mean. Only the first two had their performance reported, of these SVM performed best.		
56		This paper proses VGG-TSwinformer, a transformer-based deep learning model that uses a temporal attention mechanism to handle longitudinal data.	

Table A.1 – continued from previous page

ID	SRQ1 answer	SRQ2 answer	SRQ3 answer
57	This review is very high level. Some generally relevant remarks about machine learning practices are made. Various ML methods are mentioned to be usable for disease detection, but no performance metrics are mentioned. Mentioned techniques are LR, LDA, kNN, NB, SVM, ANN, CNN, RF, DT, and gradient boosting.		
58	A review on Alzheimer's Disease detection using neuroimaging data. Mentioned techniques: 3D-CNN, CNN, (semi-supervised) Generative Adversarial Networks, RNN, ANN, RF, SVM, Extreme Learning Machine, and kNN combined with Principal Component Analysis.		
59	Detection technique review for diabetes. Mentions DT, ensemble methods such as RF, LR, SVM, NN, Bayesian methods, kNN, dimensionality reduction, clustering, and regularized regression as techniques to detect diabetes.		

Table A.1 – continued from previous page

ID	SRQ1 answer	SRQ2 answer	SRQ3 answer
60		This paper used a two-prong approach to predict Alzheimer's Disease progression. For stage one, predicting whether a patient will progress or not, DT, RF, SVM, linear regression, kNN, and LSTM are used. For stage two, predicting the conversion time, DT, Ridge, LASSO, RF, SVM, feed-forward NN, and LSTM are used. For both stages, LSTM performs best due to its ability to use longitudinal data effectively.	Various pre-processing techniques are used. Any features with >30% missing data are removed. kNN-based imputation is performed for other missing values. Forward filling is used for time series data imputation. SMOTE is used to mitigate imbalanced data issues.
61	This paper mentions a wide variety of techniques that can be used for detection of various different diseases. The techniques mentioned are Sequential Minimal Optimization Multiple Kernel Learning, SVM, KNN, RF, Bayesian Hidden Markov Model, Cox Regression, LASSO, Gradient Boosting, Adaboost, ANN, DNN, Auto-encoder, Principal Component Analysis, XGBoost. Performance scores are not reported.		

Table A.1 – continued from previous page

ID	SRQ1 answer	SRQ2 answer	SRQ3 answer
62	A detection review for mental health disease such as schizophrenia and ADHD based on various biomarkers, imaging, and behaviour data. Techniques mentioned are RNN, CNN, GRU, LSTM, Auto-encoder, Deep Feed-forward NN. No performance scores reported.		
63	Detection review for Alzheimer's Disease using biomarkers. Techniques used: 3D CNN combined with LSTM, Convolutional Auto-encoder, CNN, Stacked Auto-encoder Multi Kernel SVM (SAE-MKSVM), Sparse Auto-encoder, SVM. No performance metrics mentioned.		
64	Detection review on classifying Alzheimer's Disease. Techniques mentioned are SVM (Radial Basis Function), SVM (Linear), Group LASSO SVM, 3D CNN, Deep Belief Network, DNN, 3DCNN, Stacked Auto-encoder combined with SVM, Deep Boltzmann Machine combined with SVM, and Restricted Boltzmann Machine combined with SVM. Best performance reported was achieved with a Deep Boltzmann Machine combined with SVM.		

LSTM= Long Short Term Memory, LR = Logistic Regression, SVM = Support Vector

Machine, RF = Random Forest, DT = Decision Tree, MLP = Multi-Layer Perceptron, kNN = k-Nearest Neighbours, CNN = Convolutional Neural Network, ANN = Artificial Neural Network, NB = Naive Bayes, NN = Neural Network, RNN = Recurrent Neural Network, GRU = Gated Recurrent Unit, LDA = Linear Discriminant Analysis