An Investigation of Interpretability Techniques for Deep Learning in Predictive Process Analytics for Healthcare

Abstract

This paper explores interpretability techniques for two of the most successful learning algorithms in medical decision-making literature: deep neural networks and random forests. We applied these algorithms in a real-world medical dataset containing information about patients with cancer, where we learn models that try to predict the type of cancer of the patient, given their set of medical activity records.

We explored different algorithms based on neural network architectures using long short term deep neural networks, and random forests. Since there is a growing need to provide decision-makers understandings about the logic of predictions of black boxes, we also explored different techniques that provide interpretations for these classifiers. In one of the techniques, we intercepted some hidden layers of these neural networks and used autoencoders in order to learn what is the representation of the input in the hidden layers. In another, we investigated an interpretable model locally around the random forest's prediction.

Results show learning an interpretable model locally around the model's prediction leads to a higher understanding of why the algorithm is making some decision. Use of local and linear model helps identify the features used in prediction of a specific instance or data point. We see certain distinct features used for predictions that provide useful insights about the type of cancer, along with features that do not generalize well. In addition, the structured deep learning approach using autoencoders provided meaningful prediction insights, which resulted in the identification of nonlinear clusters correspondent to the patients' different types of cancer.

Keywords: Explainable AI, Deep Neural Networks, Long Short Term Memory, Random Forests, Medical event logs, LIME, Autoencoders

1. Introduction

A recent report by [1] forecasts that "AI augmentation will create $2.9 trillion of business value in 2021" and "decision support and AI augmentation will surpass all other types of AI initiatives to account for 44% of the global AI-derived business value" by 2030. AI augmentation involves the combined use of AI technologies and human decision-making capabilities. A prominent example can be seen in the recent trend in business process analytics, where AI learning techniques, such as machine learning and deep learning, are used to build predictive capabilities into business processes in order to support more timely and accurate decision-making by business stakeholders.

Business processes form a lifeline of business within and across organisations. Executions of business processes involve a wide range of stakeholders and are generally supported by a variety of IT systems, and data generated by process executions are recorded in event logs (of IT systems). Business process analytics focuses on analysing relevant event logs in order to extract insights about process behaviour and performance and thus support decision-making in organizations [2]. More recently, machine learning and deep learning techniques are applied to constructing predictive models from event logs that record historical execution of a business process, and such models can be used for process prediction, i.e., forecasting the future behaviour of running instances of a business process (e.g., which task will be carried out next, when, and who will perform it; when will an ongoing process instance complete, and what will be the outcome upon completion).

Despite the promising benefits that can be delivered by *predictive process analytics* (that is underpinned by AI learning techniques), its adoption by (human) users is impeded by *a lack of trust in the accuracy of the results generated by predictive models*. AI learning techniques, in particular deep learning models, have sophisticated internal representations, and thus are often *applied as a black-box in building predictive capabilities in process analytics*.

Consequently, it is hard for users (e.g., data analysts, process stakeholders, business executives) to gain insights regarding the underlying reasons that have led to certain process predictions. Without understanding the rationale of the black-box machinery, there will be a lack of trust in the accuracy of the predictions, thus a reluctance to use the predictions, and in the worse case, consequences of an incorrect decision based on the predictions (misclassifications).

Consequences of misclassifications, either due to biases in data, or simply because the algorithm could not generalise well enough, can be damaging for decision-makers and can put certain societal groups at risk [3, 4]. An example of such consequences was explored in [5] in which the authors compared three commercial facial recognition systems (including from Microsoft and IBM). The authors tested the systems with a balanced dataset in terms of gender and race. What they found was that all algorithms work better on male faces than female faces. All algorithms performed better on lighter faces. And all classifiers perform worst on darker female faces (ranging from 20.8% error for Microsoft to 34.7% for IBM). This study already shows a warning that decision-makers are interacting with biased commercial systems that require urgent attention in order to make them more fair and transparent in terms of facial recognition. Another well known example from the Information Retrieval domain is concerned with the *word2vec* algorithm where several gender and ethnic bias were found [6, 7, 8]. In the particular study of [8] for instance, the authors showed that the neural network used to train the English language words was already encoding societal and gender biases. The authors used *Word2vec* to generate missing words in analogies. One of the findings in the study show that, for the analogy *Man is to computer programmer as woman is to 'X'*. The missing variable X was replaced by *"homemaker"*, illustrating that the algorithm is already biased towards gender and stereotypes.

In this paper, we focus on medical event logs describing all the processes that patients with cancer went through and in predictive process analytics that can enable us to understand if we can detect patterns, using neural networks, in the event logs that correspond to patients with different types of cancer. And if those patterns exist, and we are able to predict the type of cancer of a patient given his/her processes, then how can we trust the predictions of the neural networks. Are they correct, and why? Are they incorrect, and why?

If the goal of deep learning systems is to provide decision-making systems that can assist decision-makers across different fields, including medical decision-making, then one needs systems that have underlying interpretable and explanatory mechanisms that can help decision-makers trust the system's decisions: understand why they work, why they failed, etc [9]. A misclassification in a patient using a deep learning medical system can have extremely high human costs if one blindly accepts and trusts the system [10]. This trust can be achieved by creating explanatory models that are able to provide interpretations of why deep learning algorithms are making certain choices [11]. In this sense, there is a high demand for interpretable deep learning methods that can make the behaviour and predictions of deep learning decision support systems understandable to humans [12].

Although opaque decisions are more common in medicine than researchers might be aware of [13], doctors are constantly confronted with uncertainty, and with data that is incomplete, imbalanced, heterogeneous, noisy, dirty, erroneous, inaccurate and therefore there should be a moral responsibility to provide decision-makers sufficient evidence of why deep learning algorithms are making some predictions in such complicated decision scenarios [13]. Ultimately, medical decisions should belong to the decision-maker rather than the algorithm. The information of the algorithm should therefore complement and augment the knowledge of the decision-maker in scenarios under uncertainty. This leads to a dilemma in terms of the accuracy vs. interpretability tradeoff: either we have models that achieve very high accuracies, such as deep neural networks, but they do not provide any understandings of how the features interact when it comes to predictions; or we have weaker models that provide a reasonable interpretation of the impact different features in the prediction process, such as decision trees, but with much less predictive capacity [14].

In medical decision support systems, predictive tasks using deep learning approaches are hard, due to the fact that doctors are constantly confronted with uncertainty, and with data that is incomplete, imbalanced, heterogeneous, noisy, dirty, erroneous, inaccurate. This data is also expressed in arbitrarily, and unfixed high-dimensional spaces, which makes it hard to model it and to apply machine learning algorithms [15, 16]. Moreover, datasets are small, which makes the learnt models very likely to overfit [17].

In this paper, we investigate explainability mechanisms in deep neural networks and random forests, since these two models are have been successfully applied in different predictive tasks in medical decision-making [18, 19]. We explore a real world medical decision event log from the Business Process Intelligence (BPI) Challenge[1] that ran in 2011. This event log corresponds to data that was collected in the Gynecology department from a hospital in

[1] https://www.win.tue.nl/bpi/doku.php?id=2011:challenge

the Netherlands. The dataset contains the history of all medical activities undergone by the patient (e.g., blood test, x-rays, medical appointments, etc.), together with information about the treatments and specific information about the patient (e.g., age, number of years spent in treatment, etc). The main challenge with this dataset is that a patient is not defined by an N-dimensional feature vector. Instead, a patient is defined by a set of N features that change throughout T timesteps. This means that patients are represented by an unfixed length of medical activities, depending on the severeness of their disease (e.g., a patient might have gone through a set of 70 different medical activities, together with specific information about other features, and another patient might have gone through a set of 300 medical activities). Figure 1 shows a small example of the event log that we will analyse in this paper.

The extraction and analysis of meaningful processes out of these event logs, has been the core of Process Mining, which is a research field that aims to analyse the main processes that underpin an organizational activities by analyzing the organization's event logs. This analysis is important for different reasons: (1) assess internal performances of the organization, (2) to raise awareness of how people work sand how they interact with each other, and ultimately (3) to identify opportunities for efficiency improvement and better usage of resources [20, 21]. Predictive process mining is the usage of event logs in order to make reliable predictions in the workflow of a given process, such as predicting the next activity in the event log [22], predicting the continuation of a business process or even make predictions about the time of each cycle in a sequence of activities.

Patient ID	Activity ID	Timestamp	Cancer ID	Age
1515	A	25-05-2005T16:15	M13	22
1515	B	25-05-2005T16:45	M13	22
1515	A	27-05-2005T09:30	M13	22
1515	C	25-05-2005T16:15	M13	22
...	...	...	...	...

Medical Event Log

Figure 1: Example of an event log showing different features that are dynamic and change through time and features that are static.

Given that a patient is represented not by a single N-dimensional feature vector, but by a set of $T \times N$ medical activities and features that change throughout time, one can visualise the medical processes that a single patient goes through during treatment. In order to demonstrate the complexity of the medical data that we will cover in this paper, Figure 2 shows all medical activities that a single patient who has been diagnosed with cancer of vulva has been through.

In this sense, we are interested in analysing whether a set of medical activities is targeted to a patient's specific type of cancer. According to Holzinger [23], health practices should be adjusted to the individual patient and they

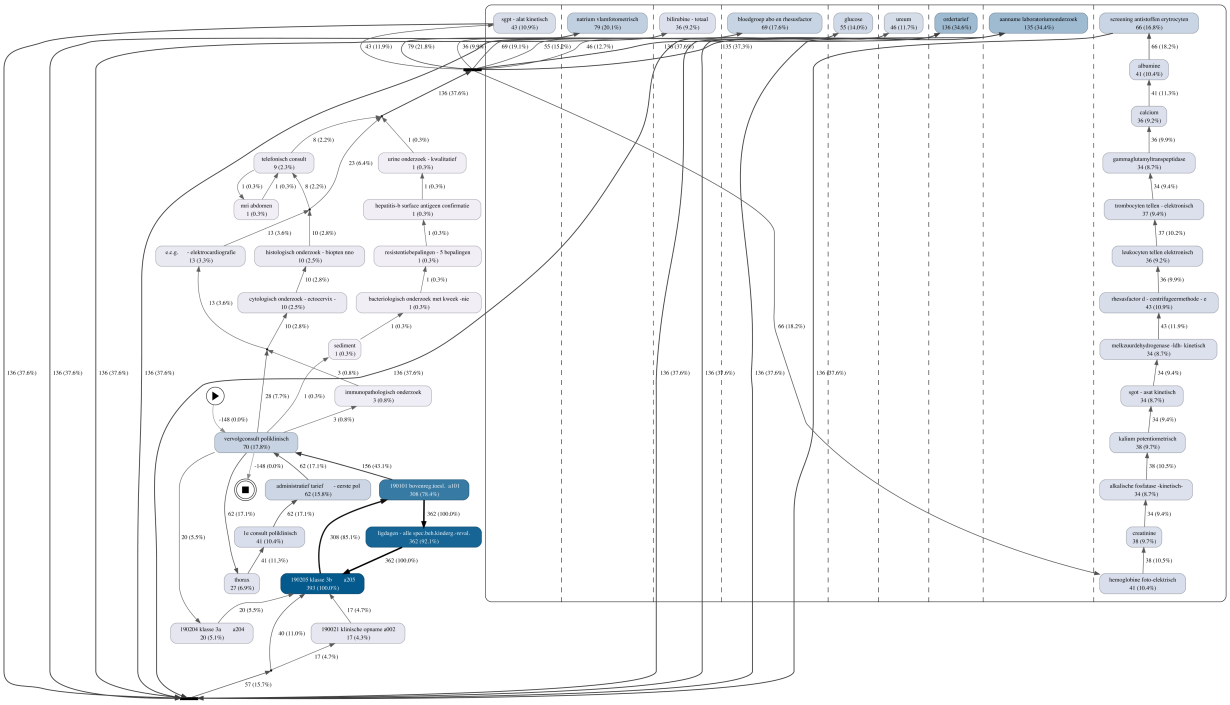


Figure 2: Representing 30% of the most representative medical activities associated to a single patient diagnosed with cancer of vulva.

should be reflected in the hospitals underlying medical decision models.

The two main question addressed by this paper are (1) to understand if patients with specific types of cancer have a targeted and specific set of medical activities associated to them and (2) what type of explainability mechanisms should be involved in this specific task to provide the user (a medical doctors, for instance) the right information that allows the understanding of why the algorithm is making such predictions.

To answer these questions, we use long short term memory (LSTM) neural networks and random forests (RF) to make this prediction. In order to provide explainability mechanisms to the predictions of these models, we explore the usage of autoencoders, where we use these structures to intercept hidden layers of the neural network and try to derive interpretations of clusters that can be found in the data. We also explore a novel explanation technique that explains the predictions of random forests in an interpretable and faithful manner, by learning an interpretable model locally around the prediction [24].

In summary, the paper aims to contribute the following:

1. A deep learning architecture of Long Short Term Neural Networks for Cancer Prediction based on real world event logs of cancer patients
2. Investigate potential interpretations and explanations of the predictions of the Long Short Term Neural Networks using autoencoders.
3. Explore the usage of the LIME framework [24] in the scope of event logs for medical decision making. LIME

consists in a technique that explains the predictions of classifiers in an interpretable manner, by learning an explainable model locally around the prediction.

This paper is organised as follows. In Section 2 we present the main works in the literature that provide interpretable models for black boxes. In Section 3, we present the dataset used, how we cleaned it and some initial understandings about the data. In Section 4, we use deep neural networks to predict the type of cancer of a patient given his track of medical records. In Section 4.4, we present an analysis where we use autoencoders to gain deeper insights about how the predictions are being made in the neural network. In Section 5 we model the same data using random forests in order to predict the type of cancer of the patient given the track record of medical activities. In Section 5.4, we apply a local interpretable model-agnostic explanation technique to extract rule-based insights from the predictions of the data. We conclude this paper in Section 6 where we summarise the main findings in this work.

2. Related Work: From Predictions to Explanations

Over recent years, Deep Learning has demonstrated significant impacts on several predictive tasks in medical decision-making, ranging from advanced decision support systems [25, 26, 27], diagnosis of different types of cancers [28], Alzheimer's disease [29, 30], heart disease prediction [31], diabetes diagnosis [32], etc. However, the high performances that these algorithms achieve in terms of accuracy comes at the cost of low explainability and interpretability of the predicted outcomes. Since these classifiers work by computing correlations between features, and since correlation cannot be confused with causation, a solid understanding is required when making and explaining decisions.

The recent body of literature has emphasised the need to understand and trust the predictions, as having a clear understanding of the behaviour of predictive models is imperative in domains such as finance, healthcare and criminal justice [33, 34]. This has led to an increased focus and interest in the research community on *interpretable* or *explainable* machine learning [35]. To avoid ambiguity, we define the terms interpretability and explainability as discussed in the machine learning literature [35, 36].

"To interpret means to give or provide the meaning or to explain and present in understandable terms some concept" [35]. The characteristics of a machine learning model that make it *Interpretable* have been presented in recent literature [36]. Three key characteristics have been defined: *Stimulatability*, *Decomposability*, and *Algorithmic transarency*.

Simulatability is when a human can arrive at the prediction by using the model input and parameters. An example of such a model is a linear regression model. *Decomposability* is when the model input and its parameters can be understood. Decision tree model is decomposable as it is possible to understand the influence of an input parameter on the prediction. *Algorithmic transparency* refers to the level of transparency supported by the model algorithm. Neural network models are complex and hard to follow and hence have low algorithmic transparency.

Another distinct form of interpretability is post hoc interpretability. Here, explanations, visualisations are extracted from a learned model. The explanations are derived after the model has been trained and hence are model agnostic. Classification models such as Random Forest [37], and XGBoost [38] provide measures of variables/features/input that are important. There are other different techniques of extracting *explanations* from a model as presented in Figure 3 and discussed in detail.

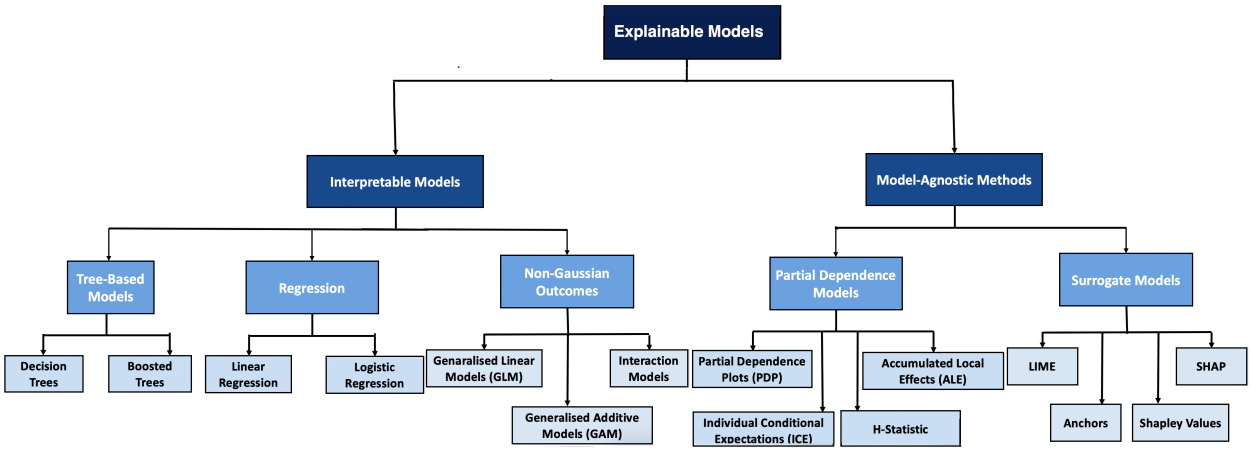


Figure 3: Most relevant model-agnostic methods proposed in the literature [14].

Partial Dependence Models

Partial Dependence Plots (PDP) show the marginal effect of at most two features on the predicted outcome of a machine learning model [39]. A partial dependence plot can depict the relationship between the label and a feature: linear, monotonic or more complex.

Generally speaking, PDP approaches use Monte Carlo methods to estimate partial functions by calculating averages and marginal effects in the training data. This allows one to get information about how the effect that these averages have in the prediction. In Zhao & Hastie [40], the authors extended PDP to incorporate causal relationships between features and predictions. One main disadvantage of this approach is that it plots the average effect of a feature in the global overall average. The approaches suffer from the independence assumption and assume that features are not correlated between each other, which may not always be true. When computing the marginals, this independence assumption can lead to marginalizations that are not representative of the data.

Two algorithms were proposed to address the limitations of PDP: *Individual Conditional Expectation* (ICE) and *Accumulated Local Effects* (ALE). ICE is a model proposed by [41] is very similar to PDP, but focuses on individual data instances, rather than taking the overall averages. Hence, ICE is an equivalent of PDP for individual data instance. ALE Plot proposed by [42] describes the influence of features on an average, on the prediction.

Surrogate Models

Surrogate models are defined by starting from the input data and the black box model (machine learning

model) by performing several evaluations of the objective functions with the original model [43]. In other words, they are approximation models that use interpretable models (stimulatable, decomposable, and algorithmically transparent) to approximate the predictions of a black box model, enabling a decision-maker to draw conclusions and interpretations about the black box [14]. Interpretable machine learning algorithms, such as linear regression and decision trees are used to learn a function using the predictions of the black box model. This means that this regression or decision tree will learn both well classified examples, as well as misclassified ones. Distance functions are used to assess how close the predictions of the surrogate model approximate the blackbox. The explanations are derived from the surrogate model as it reflects a local and linear representation of the black box model.

3. Dataset Description

The Dutch Academic Hospital Dataset is a publicly dataset made available by the Business Process Intelligence (BPI) challenge in 2011 by a hospital in the Netherlands[2]. The business process intelligence challenge is a competition where organisations make their event logs publicly available, together with specific questions that they would like researchers to address.

The Dutch dataset contains a set of 1142 patients that were diagnosed with a certain type of cancer, together with all the medical activities that they went through in the hospital [44]. These activities are dynamic and specific to the process of the patient and can describe some specific urine test, in order to try to identify potential tumours in the bladder, tests to the heart, as well as general blood tests and specific cancer-related treatments. The dataset not only contains dynamic features that are connected to the workflow of the process, but it also contains static information, like the patient's age, diagnosis, etc. In total, we have some 150 291 activities corresponding to all the 1142 patients. Table 1

Code	Cancer Name	# Cases
M11	Cancer of Vulva	60
M12	**Cancer of Vagina (not representative)**	**13**
M13	Cancer of Cervix	195
M14	Cancer of Corpus Uteri	95
M15	**Cancer of Corpus Uteri of type Sarcoma (related to M14)**	**11**
M16	Cancer of the Ovary	128
106	Mix of cancers: cervix, vulva, corpus uteri and vagina	113
821	**Cancer of the Ovary (related to M16)**	**29**
822	**Cancer of the Cervix (uteri) (related to M13)**	**22**
823	**Mix of cancers: corpus uteri, endometrium and ovary**	**8**
839	**Mix of cancers: ovary, uterine appendages and vulva**	**14**

Table 1: Summary of the different types of cancer that can be found in the dataset. Codes 821, 822, 823, 839 and M12 were ignored, since they were not representative in the data.

[2] http://www.win.tue.nl/bpi/doku.php?id=2011:challenge

The original dataset contains up to 67 features. Many of these features had redundant information. For instance, the diagnosis of the patient was spread across 16 features: *Diagnosis, Diagnosis:1, Diagnosis:2, . . . , Diagnosis:15*. This diagnosis attribute can take values such as *Squamous cell ca cervix st IIb*, which is a squamous cell carcinoma of the cervix at stage IIb of malignancy. Associated to a diagnosis, the dataset contains a set of 16 features with the diagnosis code: *Diagnosis Code, Diagnosis Code:1, Diagnosis Code:2, . . . , Diagnosis Code:15* which can be one of 11 different types of cancer that are specified in Table 1. The original dataset contains the following attributes:

- **Activity**: describes the medical activities that the patient went through;
- **Department**: identifies the department connected to the activity;
- **Timestamp**: record of the time that the activity took place;
- **Number of executions**: number of times the activity was performed;
- **Activity code**: The dataset does not provide information about this feature;
- **Producer code**: The dataset does not provide information about this feature;
- **Section**: The dataset does not provide information about this feature;
- **Age**: age of the patient;
- **Diagnosis, Diagnosis:1, . . . , Diagnosis:15**: specific diagnosis of the patient, referring to tumours, carcinomas, metasteses, sarcomas, etc;
- **Diagnosis code, Diagnosis code:1, . . . , Diagnosis code:15**: general code specific to a type of cancer;
- **Treatment code, Treatment code:1, . . . , Treatment code:10**: code specific to the treatment applied. The dataset does not provide information about these codes;
- **Diagnosis Treatment Combination ID, Diagnosis Treatment Combination ID:1, . . . , Diagnosis Treatment Combination ID:10**: code specific to the combination of the treatment and the diagnosis of the patient. The dataset does not provide information about these codes;
- **Start Date, Start Date:1, . . . , Start Date:15**: start date of the activity of the patient;
- **End Date, End Date:1, . . . , End Date:15**: end date of the activity of the patient;
- **Specialism code, Specialism code:1, Specialism code:2**: code specific to the specialism of the diagnosis of the patient. The dataset does not provide information about these codes;

The data cleaning process was conducted in the following steps:

- **Missing values.** The dataset contained 455 instances of patients who did not have any diagnosis code. The diagnosis code was spread across 16 different features (*Diagnosis, Diagnosis:1, . . . , Diagnosis:15*). In many cases, missing values were found in the remaining 15 features. For the cases where this information was not available across other features, we were able to infer the type of diagnosis based on patients who shared similar activities and treatment codes.

- **Time features.** The dataset contains 33 time related features: tart Date, Start Date:1, . . . , End Date:15, End Date, End Date:1, . . . , End Date:15 and Timestamp. The start and end dates had a huge amount of missing information and it was difficult to make any inferences about the distribution of the missing data. For that reason, we ignored these features, and instead, we created a new feature *years* that corresponds to the total amount of years a patient was under treatment. This information was taken by making the different between the timestamp recorded for the first and last activities.

- **Repeated features.** Features whose information was spread around multiple features (e.g. Treatment code, Treatment code:1, . . . , Treatment code:10) were collapsed into a single feature representing the last event recorded.

After cleaning the dataset, we ended up with 12 features: Activity, Department, Number of executions, Activity code, Producer code, Section, Age, Diagnosis Code, Treatment code, Diagnosis Treatment Combination ID and years. We analysed the distribution and correlation of the features of the dataset. Figure 4 show these relationships.

An initial look at the correlation map of the feature shows that the features do not show many correlations with the diagnosis code. This preliminary analysis suggested that there can be a template set of procedures to apply to patients that show some potential symptoms of cancer, however it does not seem to be any targeted set of procedures that a patient goes through that is specific to a type of cancer. This lack of correlation can already indicate that machine learning approached might not have very high accuracies in this specific dataset for the task of cancer prediction.

4. Experiment I: Explanatory Mechanisms for Predictions Using Deep Neural Networks

In this section, we test the hypothesis that, in theory, patients with a specific type of cancer should be associated to a more targeted set of medical activities. We test this hypothesis by formulating our problem under a deep neural network approach.

4.1. Problem Definition

Contrary to traditional deep learning approaches in the literature, where a patient is defined by a single F-dimensional feature vector, when using event logs, we have a description of daily (or even by hour / minutes / seconds) medical activities associated to a patient. This means that a single patient $X^(i)$ from a set of M patients,

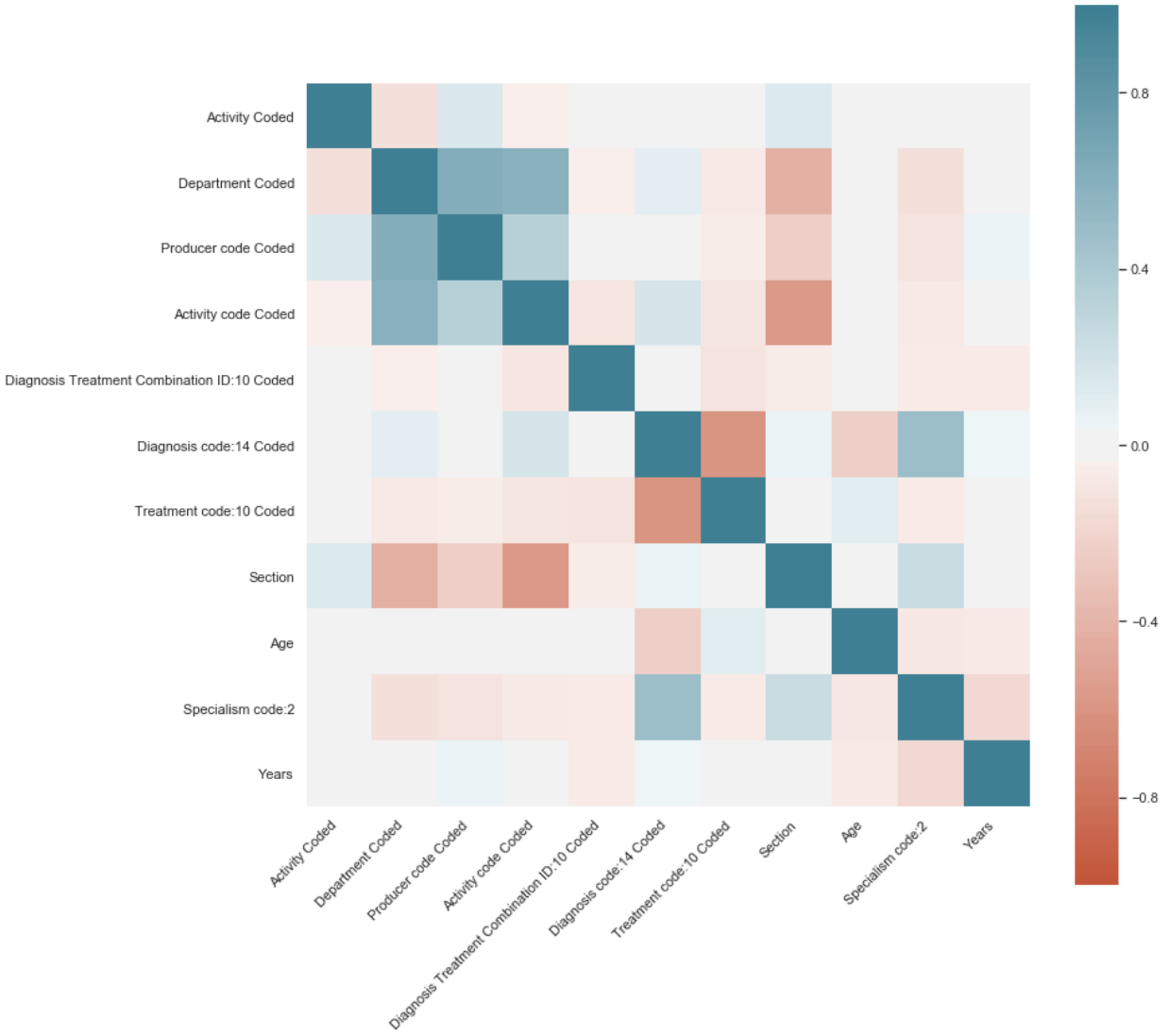


Figure 4: Correlation between features in the medical event log, after balancing and cleaning the data.

$X^(i) \in \{X^(1), X^(2), \ldots, X^(M)\}$, is defined by a set of F features that are both dynamic (showing the evolution of medical activities throughout time) and static (features concerned with the number of years the patient stays in the hospital). The length, T, of these features is also dynamic, meaning that a patient that stays 2 years in the hospital can have records of more than 1000 medical activities associated to him, while another patient that spends 1 month in the hospital can only have 20 activities in his records, for instance. Therefore, a set of patients is represented by a tensor with dimensions $(M \times L \times F)$ where M corresponds to the total number of patients, L corresponds to the length of the patient's medical records and D is the set of features associated with the patients. Each patient is also associated to a label that corresponds to the specific type of cancer that he has been diagnosed with, $Y^{(1)}$, where $Y^(i) \in \{Y^(1), Y^(2), \ldots, Y^(M)\}$. This is the class that we are interested in predicting.

$$X^{(1)} = \begin{pmatrix} f_{1,1}^{(1)} & f_{1,2}^{(1)} & \cdots & f_{1,F}^{(1)} \\ f_{2,1}^{(1)} & f_{2,2}^{(1)} & \cdots & f_{2,F}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ f_{T,1}^{(1)} & f_{T,2}^{(1)} & \cdots & f_{1,F}^{(1)}) \end{pmatrix}_{T \times F} \cdots$$

$$X^{(M)} = \begin{pmatrix} f_{1,1}^{(M)} & f_{1,2}^{(1)} & \cdots & f_{1,F}^{(M)} \\ f_{2,1}^{(M)} & f_{2,2}^{(1)} & \cdots & f_{2,F}^{(M)} \\ \vdots\vdots & \ddots & \vdots & \\ f_{T,1}^{(M)} & f_{T,2}^{(1)} & \cdots & f_{1,F}^{(M)}) \end{pmatrix}_{T \times F}$$

$$Y = \begin{pmatrix} class^{(1)} \\ class^{(2)} \\ \vdots \\ class^{(M)} \end{pmatrix}_{M \times 1}$$

4.2. Exploring Deep Learning Architectures for Cancer Prediction

In the scope of this work, we analyse a trail of medical activities and appointments associated to a patient. This set of medical activities is recorded in a given order, which suggests dependence between them.

Since the nature of the data analysed in this work is dynamic, one needs a supervised learning mechanism that is able to cope with data that has a strong and meaningful dependency between features and that is also able to keep in memory all the information from previous time steps. For these reasons, we opted for a Recurrent Neural network (RNN). RNNs were originally proposed by [45] and consist in a neural network with hidden units capable of analysing streams of data and that has reveled to be effective in many different applications which require a dependency in previous computations during the learning process, such as text classification [46], speech [47], or even DNA sequences [48]. One important characteristic of RNNs is that they share the same weights across all training steps, which is something that does not occur in traditional deep neural network models.

In this work, we explored two different types of Recurrent Neural Networks:

- **Long Short Term Memory (LSTM) Neural Networks:** are a type of recurrent neural networks that are particularly suitable for applications where there are very long time lags of unknown sizes between important events. They provide a solution for the vanishing and exploding gradient problems by using memory cells [49]. These memory cells, C_t are composed of a self recurrent neuron together with three gates: an input gate, i_t, an output gate, o_t, and a forget gate, f_t. These gates are used to regulate the amount of information that goes in / out of the cell. Information on a new input will be accumulated to the memory cell if i_t is activated.

Additionally, the past memory cell status, C_{t-1} can be *forgotten* if f_t is activated. The information on C_t will be propagated to h_t based on the activation of output gate o_t. Based in the activation functions, new candidates for the memory cell, $\tilde{C}$, are created.

- **Bidirectional Long Short Term Neural Networks (BiLSTM):** are also a type of recurrent neural network that connect two hidden layers of opposite directions to the same output, which was originally proposed by [50]. The motivation of bidirectional neural networks is due to certain contexts specific to datasets. It is not enough to learn from the past to predict the future activities, but also it should be possible to look at the future activities in order to fix the current predictions.

4.3. Predicting Patient's Type of Cancer

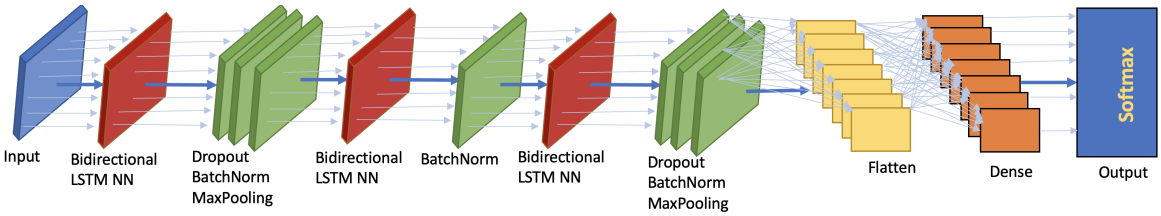


Figure 5: Deep neural network architecture used in our experimental setup.

In this section, we test the hypothesis that, in theory, patients with a specific type of cancer should be associated to a more targeted set of medical activities. To validate this, we performed a cross validation setting with a train/test set split of 80% / 20% over the network architecture illustrated in Figure 5. Table 2 illustrates the results obtained.

	Nodes	**Epochs**	**Accuracy**	**Loss**
Deep NN	25	30	0.468	1.297
LSTM NN	**20**	**200** .	**0.552**	**1.216**
BiLSTM NN	20	150	0.517	1.230

Table 2: Results obtained after conducting a cross validation grid search method over the distribution of neurons and epochs using the architecture illustrated in Figure 5. Best results were found when using a deep Long Short Term Memory recurrent neural network during 200 epochs and 20 neurons in the hidden layers.

One major challenge with deep neural networks is that they require a significant amount of training data. Given that the medical dataset is small (only 1142 patients). The best results obtained were with a Long Short Term neural network that keeps memory of previous past activities in order to predict the type of cancer of the patients. However, due to the lack amounts of training data, the algorithm could not generalise well and all models found using a grid search approach showed some levels of overfitting as it can be seen in Figures 6 that show the evolution of the accuracy and loss scores for the best performing algorithm, the LSTM neural network with 20 neurons in the hidden layers during 200 training epochs.

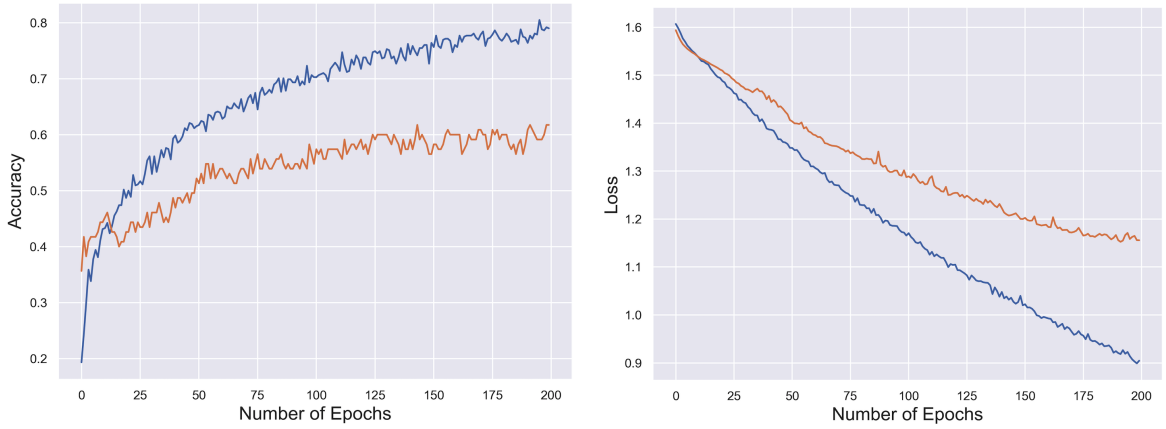


Figure 6: Accuracy and loss obtained over the training and validation sets in the best performing algorithm, showing a high degree of overfitting.

4.4. From Predictability to Explainability using Autoencoders

Understanding the reasons why deep learning algorithms make certain predictions, play an important and fundamental role to assess the effectiveness of the model and as well as providing new insights of how to transform a system or a prediction that is untrustworthy to a trustworthy one.

In this section, we investigate how the different algorithms in Table 2 are classifying the patients' cancers by using autoencoders. Autoencoders were originally proposed by [51] and are unsupervised learning techniques which use neural networks for the task of representation learning. The network architecture enables a compression of knowledge representation of the original input. This implies that correlated features provide a structure that can be learned by the network and consequently one can obtain visualisations of neurons that are being activated in the hidden layer. This compression of knowledge is crucial for the network architecture, since without its presence, the network could simply learn to copy the input values and propagate them throughout the network [52].

The structured deep learning network that was learnt using different layers fuses different modalities of information, based not only on the patients' track of medical activities, but also other features such as age, time spent in treatment, etc. This fusion of information is non-linear and leads to the representation of one single state of knowledge.

To gain understandings about the network's structured representation of this state of knowledge, we intercepted the first hidden layer of both the LSTM and BiLSTM neural networks in Figure 5 and applied an autoencoder to learn the input that led to the projections in this hidden layer. To be more specific, we used an autoencoder with two dense layers to learn the generalized latent space that better approximates to the training data. From the structured deep learning network, the autoenconders apply a non-linear transformation in the data that leads to a non-linear representation of clusters that can be helpful to provide additional insights and that can enable the investigation of misclassifications in the dataset. This provides better insights of why the algorithm is classifying

the data correctly or incorrectly, and new understandings to the decision-maker of

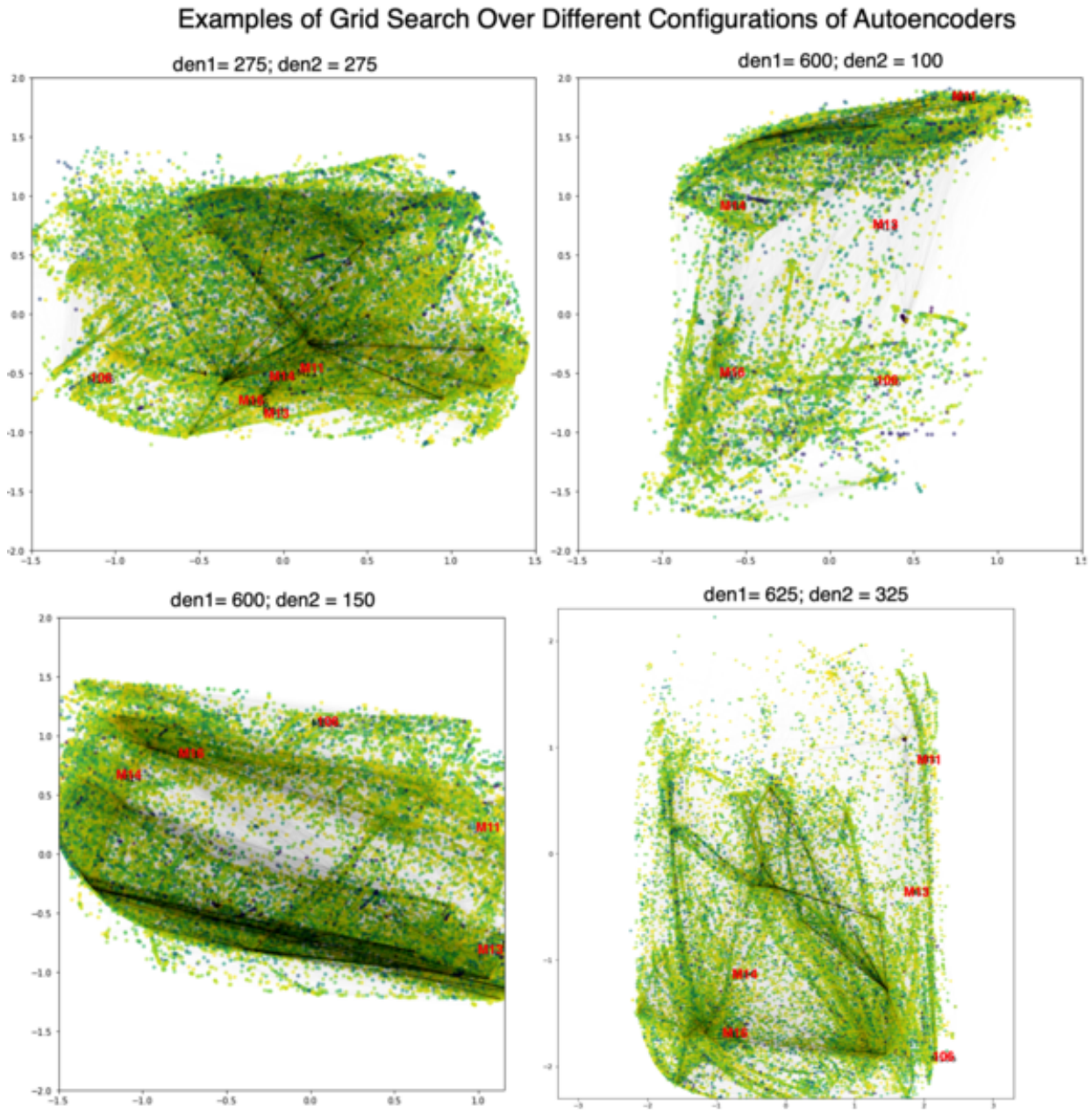


Figure 7: Example of grid searches over the first hidden layer of the LSTM network for different configurations neurons in each of the two dense layers of the autoencoder.

A grid search approach was used in order to find autoencoders that could provide meaningful results to the decision-maker regarding the relationships between the patients features and their types of cancer. Figures 8 and 7 show examples of projections that were obtained using an autoencoder with two dense layers and different number of neurons for the BiLSTM and LSTM layers, respectively.

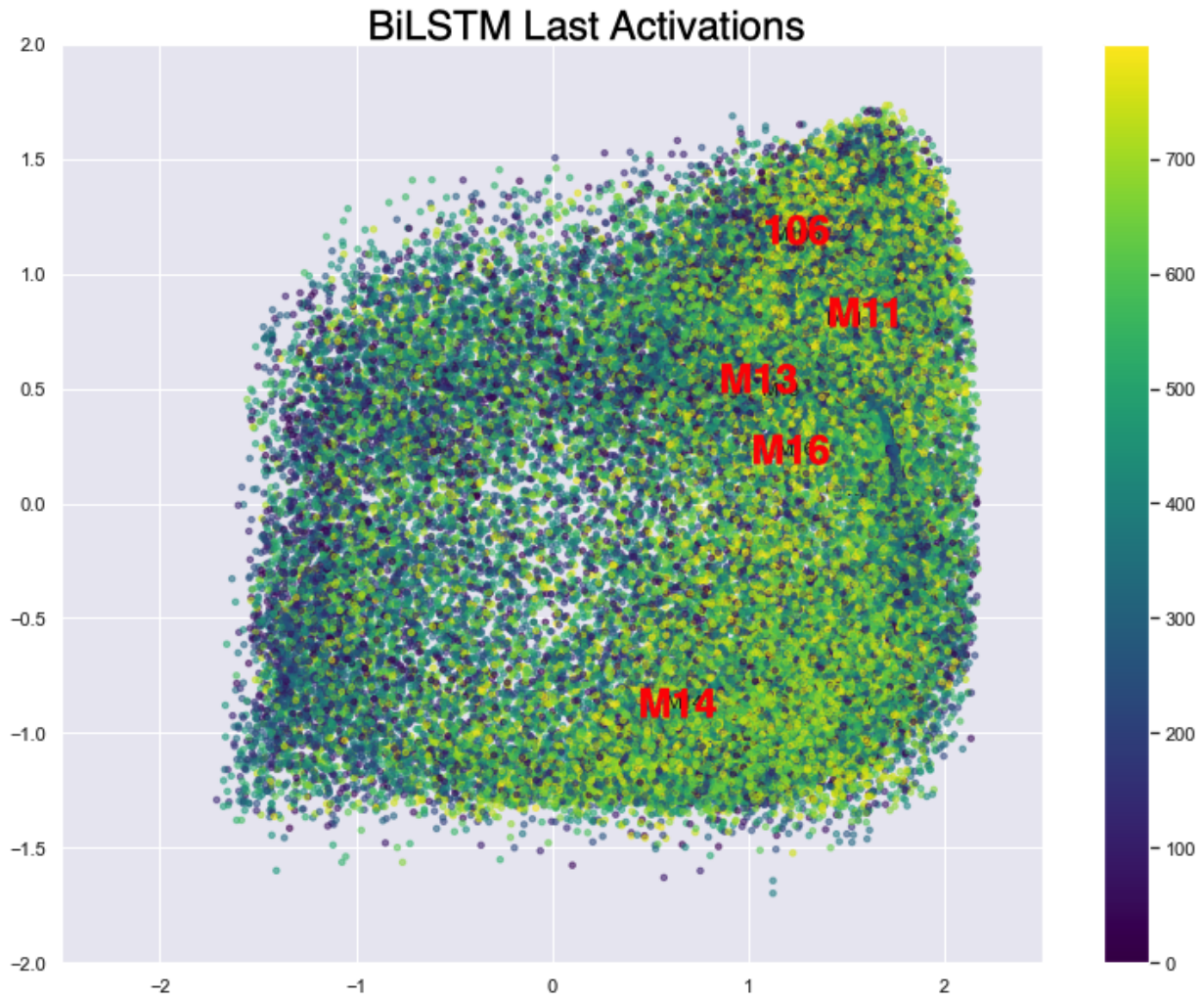


Figure 8: Example of an output image from a grid search approach that intercepts the first BiLSTM layer of the proposed deep neural network architecture. Some clusters identifying the projections of the types of cancer can be found

After performing a grid search, we extracted the most meaningful representations from the non-linear projections of the autoencoders, both for LSTM and BiLSTM network architectures, in order to analyse the misclassifications in each model. Figures 7 and 8, show the general latent spaces that were extracted for the LSTM model and BiLSTM model, respectively.

Sparser results were obtained in the LSTM model, which enabled the identification of non-linear cluster representation in this latent space representation of the state of knowledge of the network. As one can see in Figure 9, one is able to find three different clusters of data: (1) cluster 1, M16 (cancer of ovary), (2) cluster 2, M11 and M14 (cancer of vulva and cancer of corpus uteri), and (3) cluster 3, M13 (cancer of cervix).

In all three non-linear clusters that were identified, one can see that patients with different types of cancer were projected to the wrong clusters. For instance, in cluster 1, that is identified as the cluster with patients with cancer

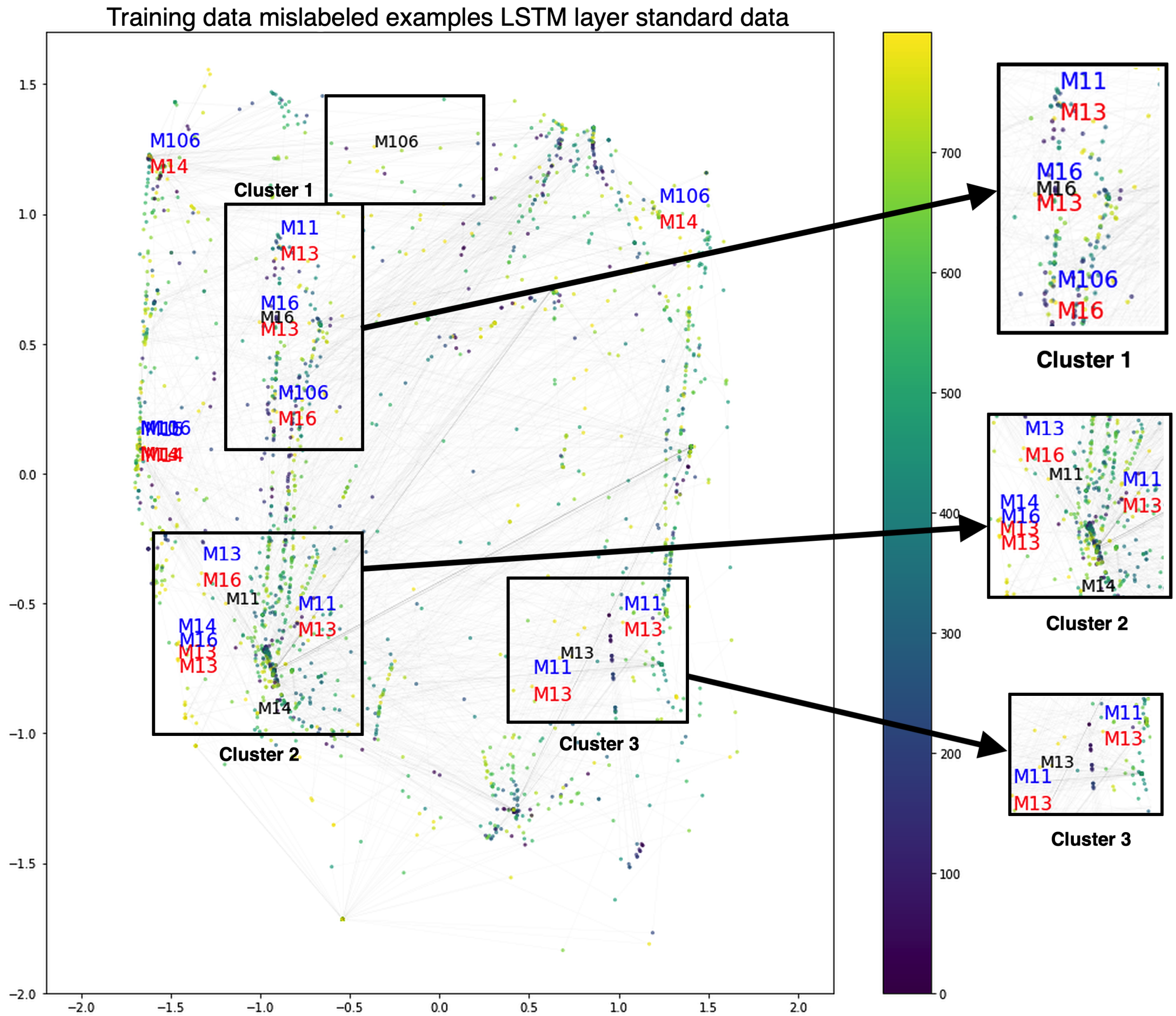


Figure 9: Misclassifications found in the projections of the Long Short Term Memory network with autoencoders.

of ovary (M16), there are patients that have the label cancer of vulva (M11) and yet the model classifies not as M16, but as M13 (cancer of cervix). This can either mean two things, (1) this specific patient diagnosed with M11 shares a very similar track of medical activities as a patient diagnosed with M13 or (2) the non-linearity nature of the projections in the generalised latent representation into a lower dimension space, distorted the distances between these patients, and as a consequence they were assigned to the wrong cluster.

When it comes to diagnosis code 106, which pertains to patients with a mix of cancers (cervix, vulva, corpus uteri and vagina), projections from the general latent representation into lower dimensions did not show specific misclassifications around this code in that specific region of the lower dimensional space. However, since code 106 pertains with patients with different types of cancer, which according to the figure mainly intersects M14 (cancer of corpus uteri), then one can understand the different misclassifications around patients with code 106

throughout the space.

Figure 10: Misclassifications found in the projections of the Bidirectional Long Short Term Memory network with autoencoders.

On the other hand, the non-linear projections found in the BiLSTM (Figure 10) did not show a clear understanding when compared with the projections in the LSTM network. This is due to the fact that during the grid search process, no sparse representations were found, which makes the representation of the space very compact. One can, however, still gain insights about the misclassifications. Like it was found in the LSTM layer, patients with cancer of vulva (M11) were wrongly projected into the M13 cluster (cancer of cervix). Once again, the non-linearity of the projections, together with the mapping into lower dimensions, disturbs the space and the distances between the patients, leading to misclassifications.

5. Experiment II: Explanatory Mechanisms for Predictions Using Random Forests

In this section, we explore alternative sub-symbolic representations and understandings of data using random forests and by learning an interpretable model locally around the model's predictions.

5.1. Problem Definition

The problem is converted to a classical supervised learning problem to compare and contrast traditional approaches while using event logs to predict cancer. Here, for each patient $X^{(i)}$, the set of features F (both dynamic and static) are mapped to the window of length T. The window represents the daily (or hourly) medical activities associated to a patient. A patient $X^{(i)}$ is represented by the vector: $\langle f_{1,1}^{(i)} f_{1,2}^{(i)} \cdots f_{1,F}^{(i)} \cdots f_{T,1}^{(i)} f_{T,2}^{(i)} \cdots f_{T,F}^{(i)} \rangle$. Hence, M patients are represented by a matrix with dimensions $(M \times (F * L))$. The length L is the number of patient's medical records (or activities recorded for each patient). The cancer associated to each patient is the class we predict. The advantage of this approach is that it allows any classical supervised machine learning algorithm to be applied.

5.2. Random Forests for Cancer Prediction in Event Logs

Random forests are an ensemble method that combine several individual classification trees [**?**]. A Random forest classifier uses multiple decision tree classifiers where each decision tree classifier is fit to a random sample, or a bootstrap sample drawn from the original data sample. The feature selected for each split in the classification tree is only from a small random subset of features in each tree. Thus, a random forest classifier consists of a number of classification trees, the value of which is set when identifying the model parameters. From the forest, the class or label is predicted as an average or majority vote of the predictions of all trees.

Random forests are known to have high prediction accuracy as compared to individual classification trees, because the ensemble adjusts for the over-fitting caused by individual trees. However, the interpretability of a random forest is not as straightforward as that of an individual tree classifier, where the influence of a feature variable corresponds to its position in the tree.

5.3. Predicting Patient's Type of Cancer

To validate the Random forest classifier, we performed a cross validation setting with a train/test set split of 80% / 20%. The optimal parameters for the classifier were found using grid search with k-fold cross validation. Table 3 presents the accuracy for two different parameters used during the grid search parameter tuning. Figure 11 presents the top five important predictors. This plot shows the features such as the *Age* of the patient, the type of Treatment, and initial set of *Activity* performed in a given sequence during the treatment (Activity Coded_0, Activity Coded_1, Activity Coded_2 representing `Activity_(sequence_number)`) are among the most important features for predicting the cancer.

Estimators	Maximum features	Accuracy
1000	100	0.556
1500	200	0.572

Table 3: Results obtained while conducting a cross validation grid search over the the number of estimators and size of the random subsets of features used for splitting a node in the tree.

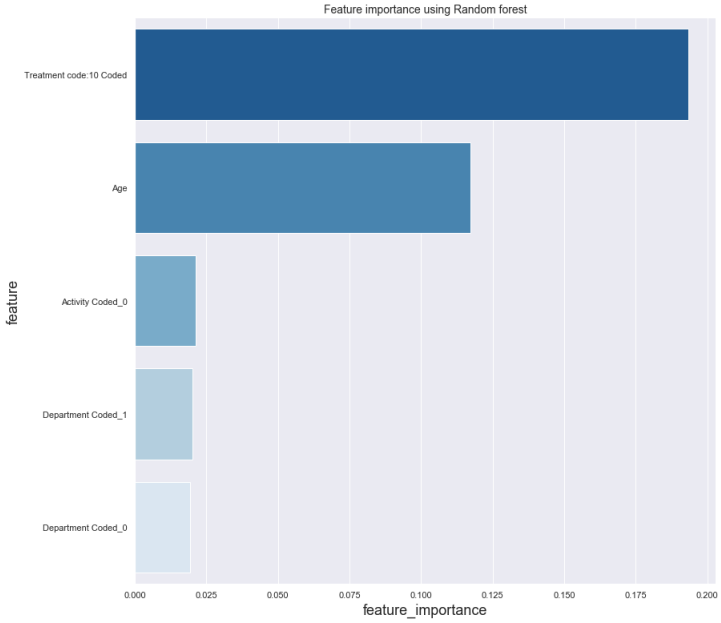


Figure 11: Top 5 important features used by Random Forest classifier.

The importance of a feature when using a Random forest classifier is computed using the 'gini impurity' measure that indicates the effectiveness of a feature in reducing uncertainty when creating decision trees. However, this method tends to inflate the importance of continuous or high-cardinality categorical variables [53]. Hence, while feature importance using 'gini impurity' measure has been consistently used, it is provides interpretability of the entire model and does not provide explanation of a specific instance.

5.4. From Predictability to Explanability using LIME

LIME [24] is used to explain a single prediction as well a global explanation of the model using a subset of individual data points or instances. LIME approximates the underlying model by an interpretable model such as a linear model that is learned on small perturbations of the original data point. This is done by weighting the perturbed instance by their similarity to the instance to be explained. Hence, the explanations are based on a linear model in the neighborhood of the instance and the explanations for an instance does not represent how the model behaves for all data points or cancer patients. Figure 12 illustrates the local explanations of predicting the cancer class '106' which is associated to cervix, vulva, corpus uteri and vagina. The explanations are based on the features $Age > 70$, and specific activities performed at a given step or sequence during the treatment (

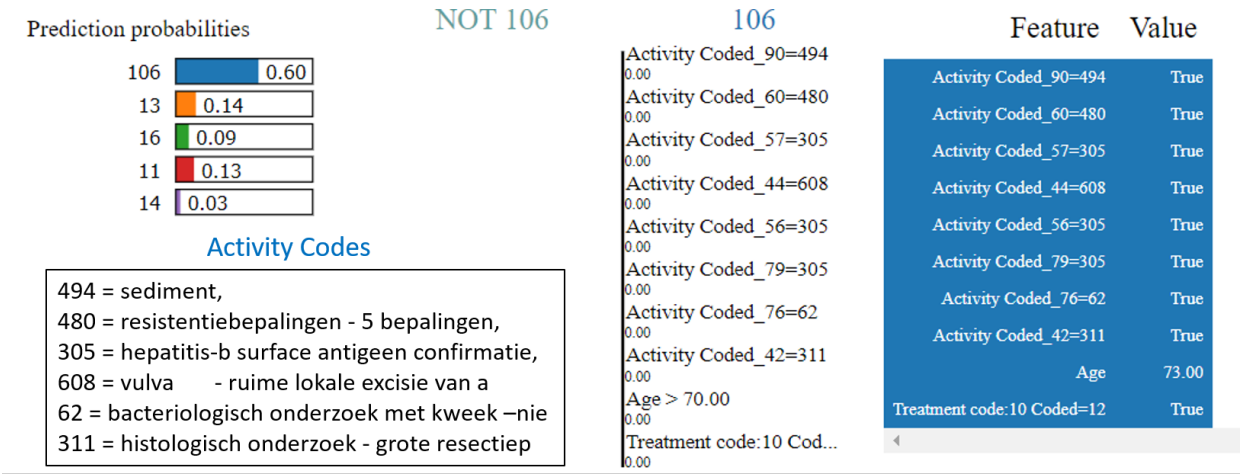
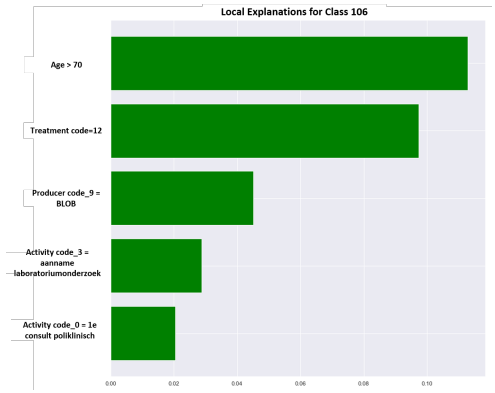


Figure 12: Local interpretation of 106 cancer class for a patient.

`Activity_(sequence_number)`).

The global understanding of the model is provided by explaining a set of individual instances. The global explanations of the model are constructed by picking a subset of instances and their explanations. The importance of a feature in an explanation and the coverage of all features defines a coverage function that is maximized to pick a subset of instances and generate global explanations. Figure 13 presents the global explanation for the cancer class '106'. Here the age, the treatment and activities performed initially provide explanation of the predictions.

Figure 13: Global interpretation of class 106 cancer.



Global explanations for two cancer classes (M11, M14) are presented in Figure 14 and Figure 15 respectively. While some of the features used by the model are relevant such as Age and the treatment undertaken, many features such as the activity 'Consultation', or being associated to the 'Obstetrics & Gynaecology clinic' are not significantly distinct features and cannot be generalized in predicting the type of cancer. However, use of such explanations provides good insight into the model and improves the trust in the prediction, and the features used for the prediction. In the context of traditional machine learning algorithms, use of local explanations provide

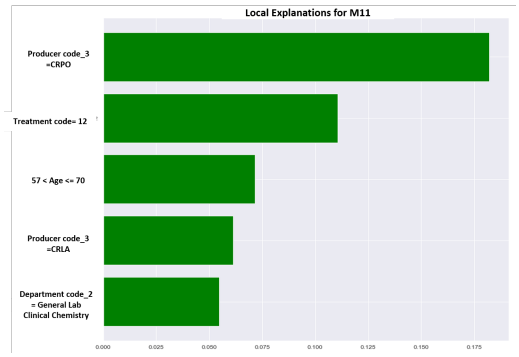


Figure 14: Global interpretation of class M11 cancer.

insights on the design of features.

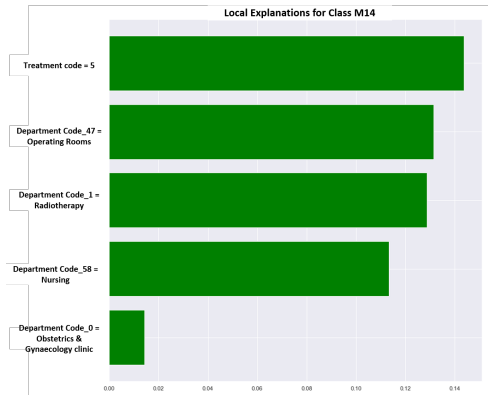


Figure 15: Global interpretation of class M14 cancer.

6. Conclusions

In this work, we explored the usage of deep learning techniques and random forests in a real world medical event log from a hospital in the Netherlands containing the track of medical records undertaken by patients with cancer. Our hypothesis was that, in theory, patients with a specific type of cancer should be associated to a more targeted set of medical activities that are particular to their type of cancer. Results showed significant results and that one could actually predict the type of cancer given past medical records of patients. The structured learning models that we explored learnt to fuse different modalities of information, based not only on the patients' track of medical activities, but also other features such as age, time spent in treatment, etc. This fusion of information is non-linear and leads to the representation of one single non-linear state of knowledge. However, this analysis in terms of accuracies can be misleading since we do not have any understandings of how the learning algorithms were making the classification.

In this sense, this paper also explored explanability and interpretability techniques in the scope of medical event

logs. In order to gain more insights about the model's black box, we intercepted the hidden layers of deep neural networks with autoencoders in order to learn a generalized latent space that better approximates to the training data. From the structured deep learning network, the autoenconders apply a non-linear transformation in the data that leads to a non-linear representation of clusters that can be helpful to provide additional insights and that can enable the investigation of misclassifications in the dataset. This method provided better insights of why the algorithm is classifying the data correctly or incorrectly, and provided new understandings to the decision-maker.

For random forests, we explored local surrogate models, more specifically the local interpretable model-agnostic explanations (LIME) framework. LIME is a metamodel that instead of interpreting directly the black box, it uses the metamodel to draw conclusions and interpretations about the black box. The individual predictions were computed by applying perturbations of the points in the original dataset. This allows one to see how the features change around these points and how they affect the predictions. Results indicate that learning an interpretable model locally around the model's prediction leads to a higher understanding about why the algorithm is making some decision. The use of local and linear model helped to identify the features used during the cancer prediction of an individual patient. We were able to identify distinct features used in different predictions, along with features that do not generalize or are not relevant.

In summary, both methods provided different sub-symbolic interpretation insights, one based on non-linear cluster representations (autoecoders) and the other based on the local impact of features in individual points in the data (LIME).

7. Acknowledgements

Dr. Andreas Wichert was supported by funds through Fundação para a Ciência e Tecnologia (FCT) with reference UID/CEC/50021/2019. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

[1] Gartner, Gartner says AI augmentation will create $2.9 trillion of business value in 2021, accessible via https://www.gartner.com/en/newsroom/press-releases/2019-08-05-gartner-says-ai-augmentation-will-create-2point9-trillion-of-business-value-in-2021 (2019).

[2] W. M. P. van der Aalst, Process Mining: Data Science in Action, Springer, 2016.

[3] M. Kosinski, Y. Wang, Deep neural networks are more accurate than humans at detecting sexual orientation from facial images, Journal of Personality and Social Psychology 114 (2018) 246–257.

[4] C. O'Neil, Weapons of math destruction: How big data increases inequality and threatens democracy, Broadway Books, 2017.

[5] J. Buolamwini, T. Gebru, Gender shades: Intersectional accuracy disparities in commercial gender classification, in: Proceedings of the 1st Conference on Fairness, Accountability and Transparency, 2018, pp. 77–91.

[6] A. Caliskan, J. J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases, Science 356 (2017) 183–186.

[7] N. Garg, L. Schiebinger, D. Jurafsky, J. Zou, Word embeddings quantify 100 years of gender and ethnic stereotypes, Proceedings of the National Academies of Science of the United States of America 115 (2018) 3635–3644.

[8] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, A. Kalai, Man is to computer programmer as woman is to homemaker? debiasing word embeddings, in: Proceedings of the 30th Conference on Neural Information Processing Systems, 2016.

[9] T. Miller, Explanation in artificial intelligence: Insights from the social sciences (2017). arXiv:1706.07269.

[10] A. Shah, S. Lynch, M. Niemeijer, R. Amelon, W. Clarida, J. Folk, S. Russell, X. Wu, M. D. Abràmoff, Susceptibility to misdiagnosis of adversarial images by deep learning based retinal image analysis algorithms, in: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), 2018.

[11] B. Kim, R. Khanna, O. O. Koyejo, Examples are not enough, learn to criticize! criticism for interpretability, in: Proceedings of the 30th Conference on Advances in Neural Information Processing Systems (NIPS), 2016.

[12] A. Holzinger, Introduction to machine learning and knowledge extraction (make), Machine Learning and Knowledge Extraction 1 (2017) 1–20.

[13] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Mller, Causability and explainability of artificial intelligence in medicine, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9 (2019) e1312.

[14] C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, Leanpub, 2018.

[15] A. Holzinger, M. Dehmer, I. Jurisica, Knowledge discovery and interactive data mining in bioinformatics: State of the art, future challenges and research directions, BMC Bioinformatics 15.

[16] S. Lee, A. Holzinger, Knowledge Discovery from Complex High Dimensional Data, Springer, 2016, pp. 148–167.

[17] A. Holzinger, Interactive machine learning for health informatics: When do we need the human-in-the-loop?, Brain Informatics 3 (2016) 119131.

[18] R. Rahman, K. Matlock, S. Ghosh, R. Pal, Heterogeneity aware random forest for drug sensitivity prediction, Scientific Reports 7 (2017) 11347.

[19] Z. Han, B. Wei, Y. Zheng, Y. Yin, K. Li, S. Li, Breast cancer multi-classification from histopathological images with structured deep learning model, Scientific reports 7 (1) (2017) 4172.

[20] D. Ferreira, Enterprise Systems Integration: A Process-Oriented Approach, Springer, 2013.

[21] D. Ferreira, A Primer on Process Mining: Practical Skills with Python and Graphviz, Springer, 2017.

[22] N. Tax, I. Verenich, M. La Rosa, M. Dumas, Predictive business process monitoring with lstm neural networks, in: International Conference on Advanced Information Systems Engineering, 2017.

[23] A. Holzinger, Trends in interactive knowledge discovery for personalized medicine: Cognitive science meets machine learning, IEEE Intelligent Informatics Bulletin 15 (2014) 6–14.

[24] M. T. Ribeiro, S. Singh, 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, 2016.

[25] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker, G. D. Tourassi, Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance, Neural networks 21 (2-3) (2008) 427–436.

[26] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, J. Dean, A guide to deep learning in healthcare, Nature medicine 25 (1) (2019) 24.

[27] D. Mantzaris, G. Anastassopoulos, A. Adamopoulos, Genetic algorithm pruning of probabilistic neural networks in medical disease estimation, Neural Networks 24 (8) (2011) 831–835.

[28] P. J. Lisboa, A. F. Taktak, The use of artificial neural networks in decision support in cancer: a systematic review, Neural networks 19 (4) (2006) 408–415.

[29] Z. Tang, K. V. Chuang, C. DeCarli, L.-W. Jin, L. Beckett, M. J. Keiser, B. N. Dugger, Interpretable classification of alzheimers disease pathologies with a convolutional neural network pipeline, Nature communications 10 (1) (2019) 2173.

[30] K. G. Ranasinghe, H. Kothare, N. Kort, L. B. Hinkley, A. J. Beagle, D. Mizuiri, S. M. Honma, R. Lee, B. L. Miller, M. L. Gorno-Tempini, et al., Neural correlates of abnormal auditory feedback processing during speech production in alzheimers disease, Scientific reports 9 (1) (2019) 5686.

[31] L. Ali, A. Rahman, A. Khan, M. Zhou, A. Javeed, J. A. Khan, An automated diagnostic system for heart disease prediction based on χ^2 statistical model and optimally configured deep neural network, IEEE Access 7 (2019) 34938–34945.

[32] B. Liu, Y. Li, S. Ghosh, Z. Sun, K. Ng, J. Hu, Complication risk profiling in diabetes care: A bayesian multi-task and feature relationship learning approach, IEEE Transactions on Knowledge and Data Engineering.

[33] H. Lakkaraju, E. Kamar, R. Caruana, J. Leskovec, Faithful and customizable explanations of black box models, in: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES, 2019, pp. 131–138.

[34] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1 (5) (2019) 206–215.

[35] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Comput. Surv. 51 (5) (2018) 93:1–93:42.

[36] Z. C. Lipton, The mythos of model interpretability, Commun. ACM 61 (10) (2018) 36–43. doi:10.1145/3233231.

[37] L. Breiman, Random forests, Machine Learning 45 (1) (2001) 5–32.

[38] J. H. Friedman, Greedy function approximation: A gradient boosting machine, The Annals of Statistics 29 (5) (2001) 1189–1232.
URL http://www.jstor.org/stable/2699986

[39] J. H. Friedman, Greedy function approximation: A gradient boosting machine, Annals of Statistics 29 (2000) 1189–1232.

[40] Q. Zhao, T. Hastie, Causal interpretations of black-box models, Journal of Business & Economic Statistics (2019) 1–10.

[41] A. Goldstein, A. Kapelner, J. Bleich, E. Pitkin, Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation, Journal of Computational and Graphical Statistics 24 (1) (2015) 44–65.

[42] D. Apley, J. Zhu, Visualizing the effects of predictor variables in black box supervised learning models (2016). arXiv:1612.08468.

[43] F. Ascione, N. Bianco, R. D. Masi, C. D. Stasio, G. Mauro, G. Vanoli, Artificial neural networks for predicting the energy behavior of a building category: A powerful tool for cost-optimal analysis, in: F. Pacheco-Torgal, C.-G. Granqvist, B. P. Jelle, G. P. Vanoli, N. Bianco, J. Kurnitski (Eds.), Cost-Effective Energy Efficient Building Retrofitting, Woodhead Publishing, 2017, pp. 305–340.

[44] R. P. J. C. Bose, W. van der Aalst, Analysis of patient treatment procedures, in: K. Daniel, Florianand Barkaoui, S. Dustdar (Eds.), Business Process Management Workshops, 2012, pp. 165–166.

[45] R. Williams, D. Zipser, A learning algorithm for continually running fully recurrent neural networks, Neural Computation 1 (1989) 270–280.

[46] P. Liu, X. Qiu, X. Huang, Recurrent neural network for text classification with multi-task learning, in: Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI), 2016.

[47] A. Graves, A. rahman Mohamed, G. Hinton, Speech recognition with deep recurrent neural networks, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013.

[48] Z. Shen, W. Bao, D.-S. Huang, Recurrent neural network for predicting transcription factor binding sites, Recurrent Neural Network for Predicting Transcription Factor Binding Sites.

[49] J. S. Sepp Hochreiter, Long short-term memory, Neural Computation 9 (1997) 1735–80.

[50] M. Schuster, K. Paliwa, Bidirectional recurrent neural networks, IEEE Transactions on Signal Processing 45 (1997) 2673–2681.

[51] M. Kramer, Nonlinear principal component analysis using autoassociative neural networks, AIChE Journal 37 (1991) 233–243.

[52] M. Harradon, J. Druce, B. Ruttenberg, Causal learning and explanation of deep neural networks via autoencoded activations (2018). arXiv:1802.00541.

[53] C. Strobl, A.-L. Boulesteix, A. Zeileis, T. Hothorn, Bias in random forest variable importance measures: Illustrations, sources and a solution, BMC Bioinformatics 8 (1) (2007) 25.