



**Utrecht
University**

**Explainable AI methods in clinical practice
to obtain satisfactory performance
and doctors' confidence**

Julita Sobiczewska

5112168

j.sobiczewska@students.uu.nl

First Supervisor: Dr. Pablo Mosteiro Romero

Second Supervisor: Dr. Marijn Schraagen

Artificial Intelligence, Utrecht University

The Netherlands

13 July 2022

Acknowledgments

Firstly, I want to express my gratitude to my first supervisor - Dr. Pablo Mosteiro Romero, for his invaluable supervision, continuous advice, and support. Secondly, I would like to offer my special thanks to Dr. Marijn Schraagen for providing additional feedback and valuable suggestions. Additionally, I would like to thank Utrecht University for 2 years of study and intensive learning, thanks to which I am following my dream education path. Finally, I would like to extend my thanks to my family and friends for their encouragement and support all through my studies, keeping me motivated everyday, and my friend Julia for daily sessions in the library and mutual support over the past weeks.

Abstract

Artificial Intelligence (AI) techniques can greatly contribute to many fields of medicine by creating cutting-edge, efficient and effective methods to treat, monitor, and analyze patient records. Unfortunately, the lack of transparency of the models causes limited adoption of AI techniques in the treatment of patients. In order to use AI systems in medical practice, models that interpret the decisions made by these systems are applied. Such models are called Explainable AI (XAI).

This thesis presents a comprehensive analysis on application of XAI models in the medical field through literature review and experimental work. I carry out two clinical classification tasks to gain better understanding on which explainability methods and NLP models should be used in the different classification tasks in clinical practice to obtain satisfactory results and doctors' confidence. Throughout the experiments I follow a framework proposed by Markus et al. with recommendations for choosing between different explainable AI methods.

Obtained results show that the framework with recommendations help with the decisions during study execution, but there are some ambiguities in the graph. This thesis points out the problem of non human-interpretable explanations. In addition, considerations and improvements to the previously proposed framework are presented.

Contents

1	Introduction	8
1.1	Background	8
1.2	Research Question	11
1.3	Thesis Outline	12
2	Literature Study	13
2.1	Clinical Data	14
2.2	NLP Methods in Clinical Practice	15
2.2.1	Word Representation	17
2.2.2	Classification Methods	19
2.2.2.1	Interpretable Models	19
2.2.2.2	Non-interpretable Models	21
2.2.2.3	State-of-the-Art Models	22
2.2.3	Evaluation Metrics	24
2.3	Explainable AI in Clinical Practice	26
2.3.1	Role of Explainability and Formalization	26
2.3.2	XAI Methods	30
2.3.3	Selected Studies	31
2.3.4	Challenges	34
3	Experimental Setup	37
3.1	Clinical Datasets	37
3.1.1	Smoking Status Dataset	38
3.1.2	Acute Ischemic Stroke Dataset	41
3.2	Study Design	43
4	Study Execution & Results	45
4.1	Identifying Patient Smoking Status	45
4.1.1	Replication of the study	45
4.1.2	Interpretable Models	50

4.1.3	Non-interpretable Models	51
4.1.4	XAI Methods	52
4.2	Predicting Acute Ischemic Stroke	59
4.2.1	Replication of the study	59
4.2.2	Interpretable Models	60
4.2.3	Non-interpretable Models	61
4.2.4	XAI Methods	63
5	Discussion	65
5.1	Findings	65
5.2	Limitations	68
6	Conclusions	69
A	Recognizing Obesity: replication of the study	76
B	Decision Tree - whole graph	79

List of Figures

1	A framework for learning paragraph vector. Adapted from ‘Distributed representations of sentences and documents’ by Le and Mikolov [27].	19
2	The logistic function. Reprinted from <i>Interpretable machine learning</i> by Molnar [29].	20
3	Support Vector Machine. Reprinted from <i>Support-vector machine</i> by Wikipedia [30].	21
4	Fully connected (dense) artificial neural network. Adapted from <i>Deep learning collocation method for solid mechanics: Linear elasticity, hyperelasticity, and plasticity as examples</i> by Abueidda, Lu and Koric [31].	22
5	The Transformer - model architecture. Adapted from ‘Attention is all you need’ by Vaswani, Shazeer, Parmar <i>et al.</i> [32].	23
6	Classification of explainable AI methods.	29
7	Word cloud for smoking status dataset.	39
8	Word cloud for stroke dataset (AIS examples).	42
9	Word cloud for stroke dataset (Non-AIS examples).	42
10	Proposed framework with recommendations to choose amongst explainable AI methods. Reprinted from ‘The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies’ by Markus, Kors and Rijnbeek [34].	43
11	Confusion matrix for Cohen’s approach applied to the task of identifying smoking status.	48
12	SHAP explanation for Cohen’s approach applied to the task of identifying smoking status.	53
13	LIME explanation for Cohen’s approach for correct smoking status classification (current smoker class).	55

14	LIME explanation for Cohen’s approach for correct smoking status classification (non-smoker class).	55
15	LIME explanation for Cohen’s approach for correct smoking status classification (past smoker class).	56
16	LIME explanation for Cohen’s approach for incorrect smoking status classification (current smoker class).	57
17	LIME explanation for Cohen’s approach for incorrect smoking status classification (non-smoker class).	58
18	LIME explanation for Cohen’s approach for incorrect smoking status classification (past smoker class).	58
19	Comparison of ML and NLP algorithms for classifying the brain MRI reports. Reprinted from ‘Natural language processing and machine learning algorithm to identify brain MRI reports with acute ischemic stroke’ by Kim, Zhu, Obeid <i>et al.</i> [58]	59
20	Neural network model architecture.	62
21	Decision Tree graph for decision tree model applied to the task of predicting acute ischemic stroke (short version).	63
22	SHAP explanation for neural network model with tf-idf applied to the task of predicting acute ischemic stroke.	64
23	Extension of the framework from Figure 10, with one of the main findings of this study, namely, that explanation models do not always provide human-interpretable explanations. Therefore, evaluations of explanations should always be carried out.	67
24	Decision Tree graph for decision tree model applied to the task of predicting acute ischemic stroke (long version).	79

List of Tables

1	Confusion matrix for classification.	25
2	Distribution of classes in the smoking status dataset.	38
3	Number of occurrences of the 20 most frequent sections in the smoking status dataset.	40
4	Distribution of classes in the stroke dataset.	41
5	Performance measures for Cohen’s approach applied to the task of identifying smoking status.	49
6	Performance measures for interpretable models applied to the task of identifying smoking status.	50
7	Parameters for the Longformer model.	51
8	Performance measures for non-interpretable models applied to the task of identifying smoking status.	52
9	Performance measures for interpretable models applied to the task of predicting acute ischemic stroke.	61
10	Performance measures for non-interpretable models applied to the task of predicting acute ischemic stroke.	61
11	Distribution of classes in the obesity dataset.	77
12	Performance measures applied to the task of recognizing obesity. . .	78

1 Introduction

1.1 Background

Artificial Intelligence is the ability of a machine to demonstrate human-like skills such as learning, planning, reasoning, or creativity. AI plays a significant role and brings unlimited research opportunities in many different areas. It creates systems that perceive their environment, manage what they perceive and work towards achieving a specific goal using data specially collected for a given problem. AI systems are able to adapt their behavior to some extent by analyzing the effects of previous actions and acting autonomously.

In AI, there may be simple linear models that can be easily interpreted by humans, but may not lead to perfect predictions of complex problems. The second option are nonlinear, intricate models that provide better performance for most tasks but are too complex for humans to understand. It is not always the rule, because it all depends on the task, dataset, preprocessing, and many other factors.

One of the basic concepts that is associated with AI is Machine Learning (ML). The first definition of the term are the words of Arthur Samuel, who introduced the term ML in the 1950s, describing it as: “field of study that gives computers the ability to learn without being explicitly programmed” [1]. Samuel’s definition does not fully capture how machine learning works today. Thus, it would be more accurate to define ML as providing a program with the ability to automatically learn and self-improve based on previously provided data without exact programming.

Currently most state-of-the-art AI systems use Deep Learning (DL), which is a subset of ML based on artificial neural networks. The learning process is deep because the structure of artificial neural networks consists of multiple input, output, and hidden layers. Each layer contains units that transform input data into information that subsequent layers can use to perform some predictive task.

With this structure, the machine can learn using its own data processing. In this way, human functions are replaced by machines and sometimes even improved and sped up. Machine learning can be used anywhere there is a need to analyze and classify large amounts of data. With ML algorithms, computers can independently analyze data, and create and adjust models to acquire new knowledge needed to solve a given problem. The use of machine learning in data analysis, in most cases, is expected to provide increases in efficiency, productivity and uptime, and to reduce costs [2]. Machine learning will give better results than classical solutions when it is not possible to implement rules that will work well or when the scale of the problem is too large to solve it manually. While the implemented rules may not be sufficient, there are too many data instances, or the rules overlap to some extent, machine learning will be a better solution.

One of the fields where AI turns out to be extremely useful is medicine. The greatest success would be helping healthcare professionals in their daily work, freeing them from some of the paperwork, and speeding up the diagnostic process. Artificial intelligence may not be able to compete with doctors for a long time. Specialists are distinguished primarily by the ability to holistically look at the patient, combining several different seemingly unrelated information. There are currently many applications of Artificial Intelligence in medicine. Specifically distinguished are: applications in assessing disease risk and estimating treatment efficacy; applications in managing or mitigating complications; roles in ongoing patient care, and applications in ongoing pathology and treatment efficacy research [3]. The subfield of AI that deals with processing textual data is called Natural Language Processing (NLP). Research on the use of NLP in medicine has been going on for many years. Automation in this field can develop clinical information to improve patient care and lower costs [4].

However, a major obstacle to their use is the difficulty in explaining the mechanisms behind their operation. Neural networks are usually very complex, so the trouble with understanding their functioning is a problem in many fields especially in those where humans are the subject of research. This is called the black box problem. We know what information goes into the system and what

comes out of it, but we do not know is how the inference itself works. There are divided opinions about explainable models, but many examples show that doctors are more likely to use explainable models that are both effective and speed up and help them in their work [5]. Additionally, there are concerns in medicine of an ethical nature, especially in the care of elderly patients, diminished human interaction, a sense of loss of control by patients, loss of personal freedom [3], fairness, privacy and anonymity, explainability, and interpretability [6].

For fields like medicine, explainable AI is essential for creating AI systems that help to elucidate model inference. The process of training a model is complicated for a human: sorting input variables, assigning weights to each variable, and inputting final output. The AI model is able to deal with it quickly without human control. AI systems learn on their own, based on patterns in the data. Explainable AI systems aim to explain how they make their decisions. AI applications that explain how they arrive at their predictions allow people to have a proper interpretation of those decisions. To achieve the credibility and ethical and moral standards of the field, it is necessary to delve into the decisions made by AI. Such explanations should provide insight inside the thought processes behind AI conclusions. However, many analysts do indeed blindly accept the outcome of the model, whether by necessity or choice [7].

There are issues that often prevent even very advanced systems from being put into practice. The main problem is the issue of accountability. The question of who is responsible for an incorrect diagnosis is very similar to the question of liability in the case of an accident involving an autonomous car. In the traditional case, the diagnosis of a disease must be signed by a doctor. In the case of automated diagnosis, this is not specified. Intelligent systems will continue to support specialists for a long time to come, and for “routine” processes health care may be fully automated in the future. The second problem is more technical: though AI can be trained to recognize specific cases, its spectrum is much narrower than that of a human expert [8].

Another possible problem is access to clinical data. Datasets form the basis of systems based on Artificial Intelligence. The more diverse the data for a specific

task, the greater the possibilities when it comes to implementation - which is associated with better results. Nowadays, many institutes, clinics and hospitals provide data for research. This requires proper preparation of this data as it is sensitive data and has the information about physical and mental condition of the patient. This implies that these data cannot be accessed by people outside the target group, so such data often require permission from the institute for access, which is relatively rare as most of the datasets are not available. It is a huge problem to find data that can be used in a study, due to its private and sensitive nature. Medical notes are very often in their raw state: it can be free text written by the doctor, data in tables with patient results, and other notes that the doctor has written down for later analysis. For text data it is necessary to analyze it in order to understand every aspect and relevance of the information, as well as thorough preprocessing.

1.2 Research Question

Taking all the information and concerns from previous section into consideration, the following research question can be posed:

Which explainability methods and NLP models should be used in the different classification tasks in clinical practice to obtain satisfactory results and doctors' confidence?.

When considering this question, I will also focus on the following sub-questions:

- *Is there a single framework for selecting classification and explainability models?*
- *Can we assume that all explanations are human-interpretable?*
- *Are small datasets (hundreds of examples) of clinical notes are sufficient to implement an effective model and its explainable model?.*

This research work aims to use datasets for different tasks in medicine and create a machine learning model to predict output depending on the medical task.

In addition, an explainable model for each black box model will be implemented to better interpret the machine learning techniques used. A large part of my work will be the discussion about the size and quality of the datasets, the ML methods used for the given tasks, the selection of appropriate XAI models. Manual and automatic analysis of the data will be an important part. It will allow me to understand the data and adjust the appropriate preprocessing. Finally, the discussion about my findings on XAI models in the medical domain based on studied examples of medical tasks will be included. The goal will be to answer the research question and sub-questions completely and accurately.

1.3 Thesis Outline

This thesis consists of six sections. In Section 2 I will discuss the literature review that helped me understand aspects such as clinical data (Section 2.1), NLP methods in clinical practice (Section 2.2) and explainable AI in clinical practice (Section 2.3). In Section 3 I will describe and analyze two datasets used in the project (Section 3.1) and present design of the study (Section 3.2). In Section 4 I will present all methods, graphs and results for both clinical tasks. Section 5 is to focus on discussion, findings and limitations. Finally, in Section 6 I will draw the conclusions.

2 Literature Study

To understand better the problem of Artificial Intelligence in medicine, I read and analysed many articles in this field. Some of them were focused more on traditional machine learning methods, others on deep learning methods and a few of them were aimed at implementing explainable AI models. During the analysis of the studies from the articles I focused mainly on the size of the datasets, the preprocessing of the data, the methods used - both ML and explainability methods (if applicable) and the measures they used for evaluation.

Another important task was to find datasets which contain clinical notes (textual data). Medical data are particularly sensitive and they are very rarely shared with other researchers. In the present literature search, I identified three datasets of clinical notes which can be used to tackle the research questions.

This section is divided into 3 subsections: clinical data, NLP methods and explainability methods. I will discuss each of them and present the approaches of the authors of the analyzed articles. In the first part, I will discuss some examples of clinical datasets used for the study. In addition to size and quality, I will examine the methods used to represent the text data in the study and the preprocessing and postprocessing steps used. In the second part I will look at the NLP methods used, the models, the parameters and their effectiveness for the given task. Additionally, I will explain word representation methods, as well as NLP models and evaluation metrics. In the third and last part I will describe the ways in which the authors of the two articles used explainable AI methods in medicine and discuss the conclusions drawn from the articles describing the advantages, disadvantages and challenges that may arise in using explainability models in medicine, and describe two of the XAI methods.

2.1 Clinical Data

In this section I focus on quality and quantity of datasets used for implementing AI in clinical practice. I describe the datasets to which I have access and all relevant information about the datasets that I found during the literature review.

The sizes of datasets used in previous medical AI studies range from hundreds to hundreds of thousands of records. In fact, it is possible to use any size of dataset, but the differences in their studies can be significant. Unfortunately, researchers are not able to choose the size of a dataset. They use the set that will be made available to them which forces them to adjust to the size, making many experiments and analysis. Large datasets allow for more experiments using more complex neural network architectures. However, training such models can be computationally intensive and, in some cases, require many days to complete [9]. Furthermore, these models may be more difficult to analyze. Smaller datasets may be insufficient especially for more complex tasks and predictions. For smaller datasets, rule-based methods or less complex machine learning algorithms such as Support Vector Machine, Naive Bayes algorithm, Logistic Regression and similar are usually used [10].

As for data access, many universities or university hospitals have their own data available. Therefore, it is not easy to find a dataset that is public or that it is possible to access by applying for data as a researcher. Medical data is sensitive, revealing information about past, present or future physical or mental health.

In this study I identified three open datasets of clinical notes labeled for specific tasks:

- Smoking Challenge Data (described in section 3.1.1)
- Acute Ischemic Stroke Data (described in section 3.1.2)
- Obesity Challenge Data (described in appendix A)

Another extremely interesting public dataset is the SEER Cancer Incidence

Public-Use Database for the years 1973—2000. This data helped with research for predicting breast cancer survivability [11]. The SEER data files can be requested through the Surveillance, Epidemiology, and End Results (SEER) website: <http://www.seer.cancer.gov>. The SEER Program is a part of the Surveillance Research Program (SRP) at the National Cancer Institute (NCI). It is responsible for collecting incidence and survival data from the participating nine registries, and disseminating these datasets to institutions and laboratories for the purpose of conducting analytical research [12]. The SEER Public Use Data consists of nine text files, each containing data related to cancer for a specific anatomical sites. The SEER database is considered to be the “most comprehensive source of information on cancer incidence and survival in the USA” [11]. The breast cancer dataset used herein was a single file in a text format and contains over 200,000 cases. An application for access to this data for research purposes is possible. Unfortunately, this dataset can only be accessed with a specific piece of software, which limits the amount of analyses that can be done with the data. Therefore, I do not discuss this dataset further.

2.2 NLP Methods in Clinical Practice

Most research on clinical NLP use traditional machine learning models or rule-based methods [13]. This is related to the size or quality of the sets. There are also recent papers that use deep learning approaches, even state-of-the-art models that have been developed for clinical tasks [14].

There are two typical steps in evaluating a patient’s condition from medical data: information extraction, which uses NLP techniques to extract words, phrases or sentences from texts, and classification, which uses machine learning models to get results from the information extraction [15]. For the first step, the list of NLP toolkits that are used in many studies is as follows: Bag of Words (BOW), TF-IDF, Word Embeddings, Medical Language Extraction and Encoding System (MedLEE), clinical Text Analysis Knowledge Extraction System (cTAKES), Health Information Text Extraction (HITEx), Clinical Language Annotation,

Modeling, and Processing Toolkit (CLAMP), REgenstrief data eXtraction tool (REX), Topaz, KnowledgeMap Concept Indexer (KMCI), MetaMap, National Center for Biomedical Ontology (NCBO) Annotator, MedEx and in-house NLP software [15]. For the second step, I found many of the models used in the classification process often repeated for different medical studies. I will list a few of them, shortly describe the model and explain why the researchers chose this model.

The adaptive LASSO procedure was used in predicting the probability of a bipolar diagnosis [16]. Adaptive LASSO procedure identifies important features and provides stable estimates of model parameters [17]. The model is applied to multidimensional datasets. It is used to select the most useful subset of features for modeling because it reduces the coefficients of uninformative features to zero. In this study was also implemented additional rule-based classifiers based on coded diagnostic, encounter, and medication information for each patient.

Another commonly used machine learning model is Support Vector Machine (SVM). In the task: identifying first episodes of psychosis [18] the authors focused on three algorithms: SVMs (in particular LibSVM) and Weka's Random Forest and JRip algorithms. Their selection was based on experiments in which multiple ML algorithms were tried. These three were chosen because they were the best in terms of both classification accuracy and speed. They considered the three algorithms to have insight into how different techniques interact with the choice of algorithm.

For risk prediction in an inpatient forensic psychiatry setting [15], logistic regression, decision trees, Bayesian networks, Naïve Bayes, support vector machines, and repeated incremental pruning for error reduction (RIPPER) have been used. These NLP techniques are often used for risk prediction [19]. This study, as one of many, shows that from the traditional machine learning algorithms the SVM obtains the highest results.

Many AI researchers use state-of-the-art models for different medical tasks. One of these models is Bidirectional Encoder Representations from Transformers

(BERT) [20] and its many variants. For ‘Violence risk assessment using dutch clinical notes’ task [21] a few models for sequence classification were used: SVM, Random Forests, and BERTje (a Dutch pre-trained BERT) [22]. Although, BERT-like models frequently show an advantage over other machine learning models, for this task BERTje did not achieved the best performance from the chosen models. Another example of medical task with BERT usage is ‘Predicting hospital readmission’ task [23]. Authors used ClinicalBERT (BERT variant for clinical text) [24] and compared the results with three other ML models: Logistic Regression, Bidirectional long short-term memory (BI-LSTM) and BERT. ClinicalBERT had the best accuracy and proved to be the best choice for the task.

Researchers tend to use a variety of algorithms, experiment with them using parameter adjustment, combining models with others, partial use of rule-based models (in the case of a relatively small structured dataset). Many models are used for different purposes, for decisions on completely different tasks.

2.2.1 Word Representation

Natural Language Processing is considered by researchers a very difficult area of AI. The main reason for this is the nature of human language. As we know, machines read only numbers, so the main goal in representing words is to write them using numeric vectors so that the meaning of words and the context in which they are used is preserved. There are many ways to represent words in NLP tasks.

The simplest method is the Bag-of-words (BoW) model. It is a technique of simplified text representation. It consists in transforming a sequence of words into a counted set of words. The words in the “bag” are formed from all the words in the training set, then each word has a fixed place in the vector, where the number of occurrences of the word in the document is counted and placed. This method is implemented in the scikit-learn package [25] and is called *Count_Vectorizer*. Using this method we can also create binary vectors where the only thing that matters is whether a word is in the document (1) or not (0).

An improved variant of the BoW model is TF-IDF (TF – term frequency, IDF

– inverse document frequency). The method performs better because it scales the frequency of occurrences of words in a document by the frequency of occurrences in the entire corpus. Words that occur many times in the corpus often do not provide information about a particular document. This method checks how important a word is by weighting it with the frequency of occurrences in the text. The formulas used to calculate TF-IDF are:

$$tfidf_{i,j} = tf_{i,j} \times idf_j$$
$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$
$$idf_j = \log \frac{|D|}{|\{d : t_i \in d\}|}$$

where:

$n_{i,j}$ - number of occurrences of a word t_i in a document d_j ,

$|D|$ - number of documents in a corpus,

$|\{d : t_i \in d\}|$ - number of documents containing at least one occurrence of the word.

The Word2Vec [26] method is more commonly used because the model converts words into their vector representations that can reflect the meaning of words. In Word2Vec word embeddings are created by walking through the corpus and updating individual word vectors according to the frequency and proximity of their occurrence. The vector representation allows to create relationships between words such as synonyms, antonyms, and analogies. On this basis, the Doc2Vec [27] model was introduced. It is a method to vectorize documents, not individual words.

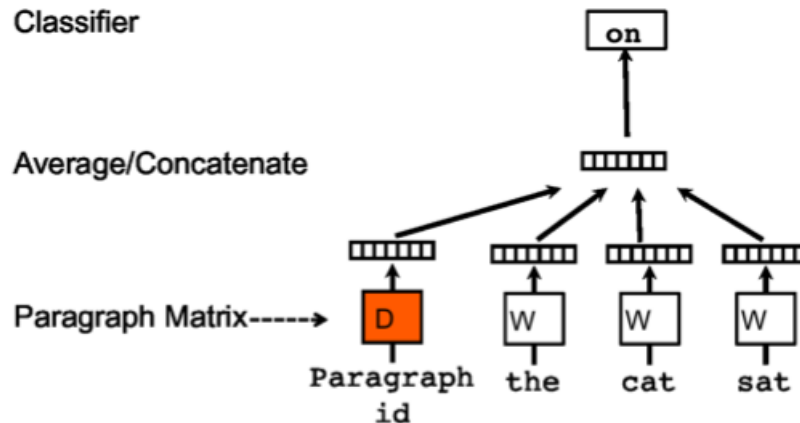


Figure 1: A framework for learning paragraph vector. Adapted from ‘Distributed representations of sentences and documents’ by Le and Mikolov [27].

In Figure 1 the simplified Doc2Vec model architecture is presented. Paragraph Vector is an algorithm that learns a fixed length feature representation from variable length piece of text such as sentences, paragraphs and documents. This algorithm represents each document with a vector that is trained to predict the words in the document [27].

2.2.2 Classification Methods

This section is divided into 3 types of classification methods: interpretable models, non-interpretable models and state-of-the-art models.

2.2.2.1 Interpretable Models

In machine learning the interpretability of the model means that the model is inherently interpretable due to its simple structure [28]. Interpretable models are readily used because of their transparency and comprehensibility. Examples of such models are: Naive Bayes, Logistic Regression or Decision Tree.

Naive Bayes classifier belongs to probabilistic classifiers. It is based on Bayes’ theorem, using conditional probabilities:

$$P(c|d) = \frac{P(d|c) \times P(c)}{P(d)}$$

where:

c - class,

d - document.

Naive Bayes Classifier is based on the assumption of mutual independence of predictors. It aims to match a given observation to the appropriate class with the highest probability. A formula is used to assign a document to a class:

$$c = \operatorname{argmax}_c P(c|d)$$

Logistic Regression is one of the simpler machine learning algorithms for classification. It can be applied to any binary classification problem. Logistic regression describes and estimates the relationship between one dependent binary variable and the independent variables. To use this model for multi-class classification, one-vs-rest strategy can be used. The idea is that one class is fitted, during model training, against all other classes. Logistic regression uses the logistic function (Figure 2) to obtain the output of a linear equation between 0 and 1 [29].

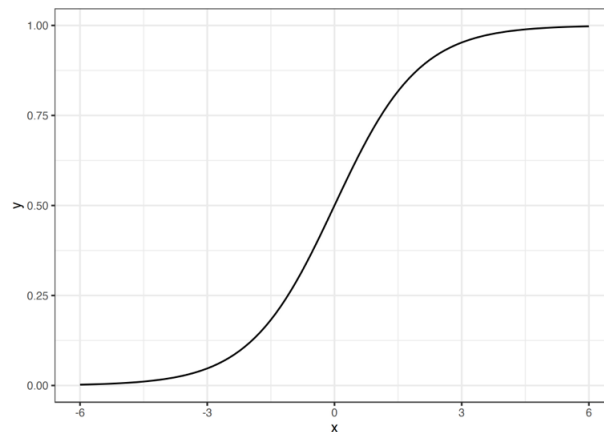


Figure 2: The logistic function. Reprinted from *Interpretable machine learning* by Molnar [29].

Decision Tree is a classification algorithm that aims to create a binary tree which assigns rules to each node. Branches separate individual leaves by features (in the case of NLP, features are words, n-grams or embeddings) and each branch leads us to the predicted class or next rule. This algorithm is intrinsically explainable if the inputs are words or n-grams, unless the tree is very deep, because it becomes unreadable to humans.

2.2.2.2 Non-interpretable Models

For non-interpretable models, I will discuss two of them: Support Vector Machine and Artificial Neural Network models.

The goal of the SVM is to find a place in the multidimensional space for each observation from the training set, and then separate them with a hyperplane according to class, with as much margin between them as possible. Classification using the SVM focuses on finding the side of the separating plane where the test example should be located. In case the data is not linearly separable, we can use a Support Vector Machine with a kernels function, which transforms the data into the required form.

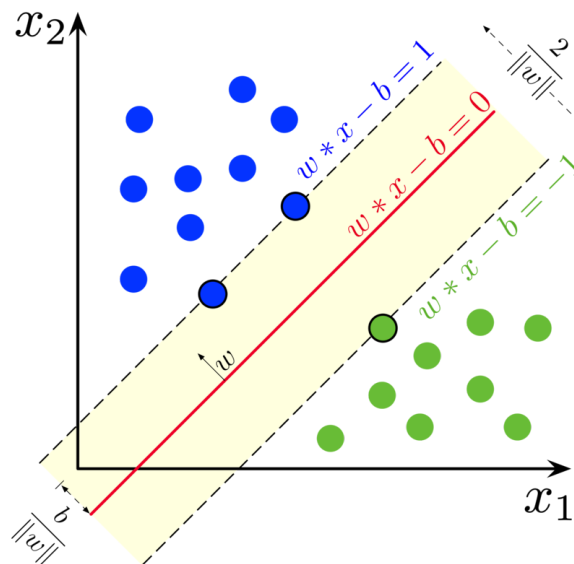


Figure 3: Support Vector Machine. Reprinted from *Support-vector machine* by Wikipedia [30].

The least human interpretable models include all models based on artificial neural networks (ANN). They consist of multiple layers of interconnected nodes, each node carrying a value and each connection carrying a weight. The simplest architecture is a dense neural network that is distinguished by the fact that each neuron has connections to each neuron in previous and next layers. This architecture is shown in the Figure 4.

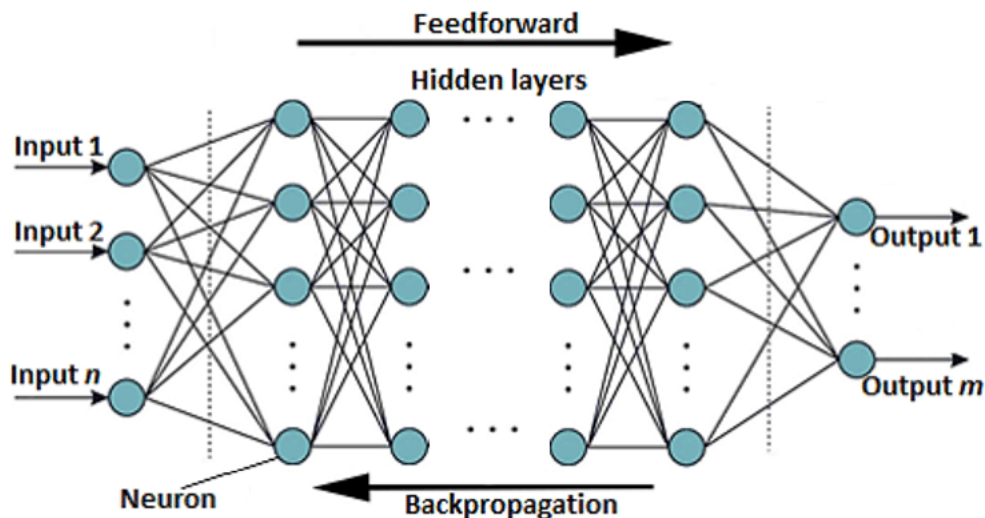


Figure 4: Fully connected (dense) artificial neural network. Adapted from *Deep learning collocation method for solid mechanics: Linear elasticity, hyperelasticity, and plasticity as examples* by Abueidda, Lu and Koric [31].

The network consists of an input layer, hidden layer(s) and an output layer. The number of hidden layers and neurons in each layer depends on the implementation. The training of an artificial neural network involves a back propagation algorithm that minimizes the error function, each time data flows through the network. The ANN model is often called a black box model because of its complex calculations inside the layers and its uninterpretable nature.

2.2.2.3 State-of-the-Art Models

From 2017, the neural network model's architecture that have most revolutionized Natural Language Processing research is the Transformer [32].

Transformer is based on attention mechanisms and is designed to solve sequence-to-sequence problems. It can be used for many different NLP tasks, such as language modeling, translation, or classification. Transformer performs these tasks quickly by removing the sequential nature of the problem - instead of passing word-by-word to the network, an entire sentence is passed. Its architecture is shown in Figure 5.

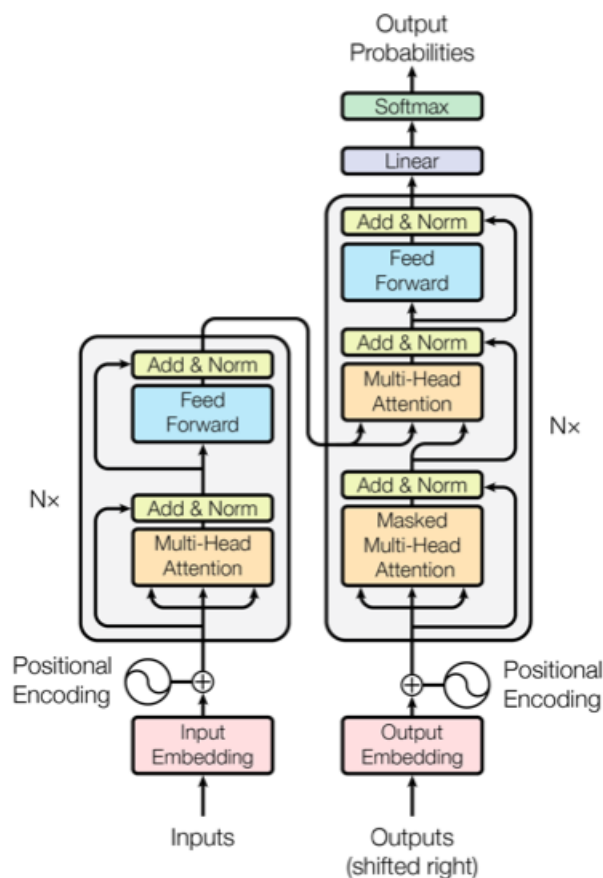


Figure 5: The Transformer - model architecture. Adapted from ‘Attention is all you need’ by Vaswani, Shazeer, Parmar *et al.* [32].

Input Embedding is a layer which is responsible for embedding the input words in a multidimensional space and representing them with vectors of numbers. Positioning Encoders are vectors that give context based on the position of a word

in a sentence or sequence. Then in Multi-Head Attention the attention vector is calculated. This vector captures the contextual relationships between words in a sentence. The obtained vectors for the words are passed to a simple MLP (Multi Layer Perceptron) network, which is used for each attention vector to transform it and pass it to the next encoder layer or to the decoder. Next, embeddings are created for the output in Output Embedding layer. In Masked Multi-Head Attention “masks” are applied. In Multi-Head Attention layer input and output vectors are combined. In this step, we check to what extent each vector of words is related to each other. The vectors are then passed to the Feed-Forward network. Its job is to simplify the task (e.g. translation) of the vector so that it is easier for the transformer to process the result of the pairings. Finally, the result from the network is passed to the Linear layer, which transforms the previous results into a dimension that is exactly the number of words from output. In the next step, the Softmax function converts the results into probabilities, which is understandable and interpretable for humans.

Due to self attention operations the Transformer is unable to deal with the long sequences. To be able to cope with processing long documents the Longformer (Long-Document Transformer) was introduced [33]. The main difference between the Transformer and the Longformer is computational complexity, which has been changed from quadratic - $O(n^2)$ to linear - $O(n)$, what enable to process long sequences.

2.2.3 Evaluation Metrics

One of the key and final steps in machine learning model deployment is model evaluation. In the case of the classification task, we have several metrics that allow us to evaluate and help us compare the models with each other. The most commonly used are: accuracy, precision, recall and f1-score. To better understand the metrics I will introduce and formalize a few values:

- True Positive (TP) - a correctly predicted positive class.
- True Negative (TN) - a correctly predicted negative class.

- False Positive (FP) - an incorrectly predicted positive class.
- False Negative (FN) - an incorrectly predicted negative class.

The visualization of the values is presented in the Table 1.

		Actual	
		Positive	Negative
Predicted	Positive	TP	FP
	Negative	FN	TN

Table 1: Confusion matrix for classification.

These values are necessary to calculate the previously mentioned metrics.

Accuracy represents the number of correct predictions over all predictions and is calculated by:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision represents the number of true positives over all positive predictions and is calculated by:

$$Precision = \frac{TP}{TP + FP}$$

Recall represents the number of true positives over all correct predictions and is calculated by:

$$Recall = \frac{TP}{TP + FN}$$

F1-score is a harmonic mean of precision and recall and is calculated by:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

2.3 Explainable AI in Clinical Practice

Explainable AI methods are increasingly used in many fields where black box models are implemented. Machine learning is currently regarded as the most promising application of AI. As mentioned in Section 1 the algorithms used hide their inner workings. It is impossible to know precisely where and how a particular decision was made. The explainable AI models allow to indicate the features that were the main reason for the model to make a decision. AI systems should be understandable both to the users who use these systems and to the developers themselves so that they can control how the predictions are made.

To better understand the importance of explainability methods in AI systems for health care, I reviewed articles about the role of explainability and its usage with clinical notes. Looking at articles that rely on artificial intelligence in medicine, a much smaller proportion of them use explainable methods. In this section, I will discuss in detail the information gathered from several articles on the topic of the explainability of Artificial Intelligence in the medical field.

I will start by discussing the role of explainability and the relevant terminology in the field (section 2.3.1), I will present two XAI methods (section 2.3.2), then move on to describing explainable methods used for two different medical tasks (section 2.3.3), and at the end I will show the disadvantages and incompatibilities that may be associated with the use of explainable AI (section 2.3.4).

2.3.1 Role of Explainability and Formalization

Following the work of Markus et al. [34], I will outline the definitions of explainability, interpretability, and fidelity, which in the remainder of this work will prove extremely useful.

Definition 1: explainability

An AI system is explainable if it is intrinsically interpretable (here the AI system is the task model - the model generating predictions) or it is complemented with an interpretable and faithful explanation (here the AI system also contains a post-hoc

explanation).

Definition 2: interpretability

An explanation is interpretable if:

- a. the explanation is unambiguous, and provides a single rationale that is similar for similar instances (clarity),
- b. the explanation is not too complex, and is presented in a compact form (parsimony).

Interpretability describes the extent to which a human can understand an explanation [35].

Definition 3: fidelity

An explanation is faithful if:

- a. the explanation describes the entire dynamic of the task model, and provides sufficient information to compute the output for a given input (completeness),
- b. the explanation is correct, and is truthful to the task model (soundness) [36].

Other terms used in the literature are comprehensibility, intelligibility, transparency, and understandability. The authors [34] do not distinguish between the terms comprehensibility, intelligibility, and understandability from interpretability, while they define the concept of transparency as that a model is transparent if it is by itself interpretable.

The first distinction that exists for explainable AI methods are: explainable modelling (inherent explainability) and post-hoc explanations.

- Explainable modelling is the development of an AI model whose functions are directly accessible to humans. Such a model is internally interpretable.
- Post-hoc explanations are designed to understand and analyze the black boxes of black box machine learning models [37]. Post-hoc explanation methods are divided into model agnostic (explaining any type of model) or model specific (explaining only specific classes of models). Some provide global explanations (explaining for the model) and others provide local explanations (explaining for a single prediction).

Explainability AI methods can also be classified according to the type and scope of the explanation. In this classification, explanations can be model-based, attribution-based or example-based. Each type of explanation can be used to provide a global or local explanation. These classes differ in terms of the information provided. In the case of explainable modelling, there is no division in this regard because it provides both explanations.

- Model-based explanations include all methods that use a model for explanation. Model-based explanations belong to both explainable modelling and post-hoc explanations. In the case of explainable modelling, it provides transparency to the model's decision-making process and is preferred if the model is simple enough to be interpreted by humans. Post-hoc explanation may be less faithful but equally valuable.
- Attribution-based explanations classify or measure input features on the basis of explainability. They are also called feature/variable validity, significance or impact methods. Most post-hoc explanation methods fall into this class. For clinicians, this type of explanation is helpful to find out which features are responsible for the predicted outcome and to be able to compare with their own knowledge.
- Example-based explanations explain the model by using instances from the dataset or creating new instances, e.g. by selecting prototypes that are well predicted by the model and criticisms that are not well predicted by the model. They identify the most influential instances for model parameters or outputs. Relatively few methods of this class are available in the literature [34].

The figure 10 shows the distribution of methods by: approach, type of explanation, and scope.

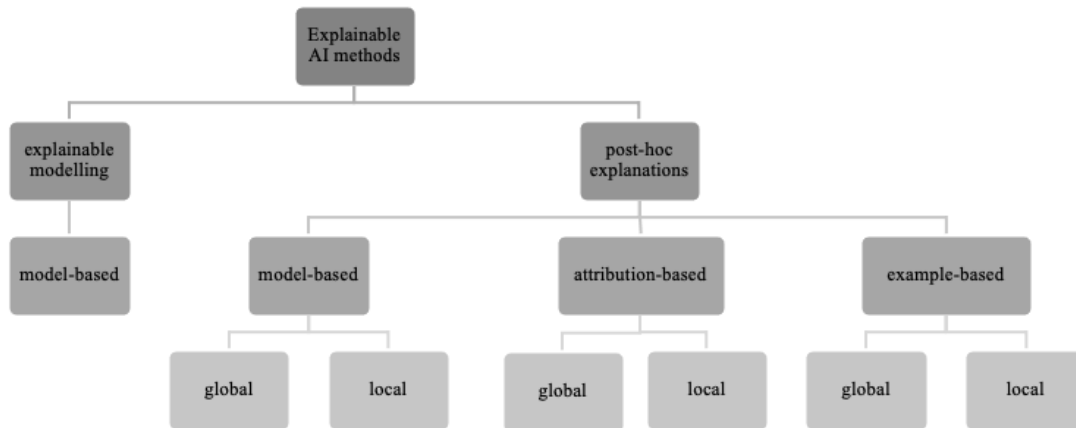


Figure 6: Classification of explainable AI methods.

In addition to understanding and selecting appropriate methods to implement interpretable models, it is also important to choose the correct evaluation method. One possible approach to dividing evaluation methods proposed by Doshi-Velez and Kim [38] is: application-grounded (experiments with end-users), human-grounded (experiments with lay humans) and functionality-grounded evaluation (proxies based on a formal definition of interpretability). Application-grounded and human-grounded evaluations may involve greater biases as compared to functionality-grounded evaluations, since the results of the former methods depend on a selected group of people.

Application-grounded approaches provide the greatest success, which is measured by testing whether the purpose for which the system was built was met, but this approach can be ineffective. Human-grounded evaluations can use both qualitative and quantitative metrics. Qualitative metrics are, for example, interviews or questionnaires. Quantitative metrics include measuring the performance of a human-machine task in terms of accuracy, response time needed, probability of deviation or ability to detect errors [39]. Human-grounded approaches provide valuable information but are expensive, time consuming and subjective [38].

In in the study by Markus et al. [34] they explain why and when explainability

can be useful and classify these reasons:

1. Not every clinical task can be improved with evaluations metrics. This means that black box model results are often unhelpful because the operation of the model is not understood. Explainability methods can detect and correct complex problems and verify the inappropriate or appropriate features used by model.
2. Clinicians' decisions are supported by a thesis. The same is expected from AI models. The decision making process has to be justified.
3. Explainability may be helpful in discovering certain relationships and insights for future tasks. Currently known knowledge by doctors and clinicals may be quite different from conclusions drawn from observations of AI models.

2.3.2 XAI Methods

The purpose of Explainable AI methods is to add transparency to black box models. There are many methods currently in use. In my work I will focus on two of them which can be implemented for many different models.

SHAP (SHapley Additive exPlanations) [40] is used wherever a balance between accuracy and interpretability of the model is needed. SHAP assigns an importance value to features. In case of Natural Language Processing field features are tokens or n-grams and the SHAP method can show which of them are the most important to predict given class. The authors of this method distinguished ML models and implemented corresponding XAI models. Currently we can use explainers such as: Tree Explainer for tree-based models, Linear Explainer for linear models, Deep Explainer for deep learning models, Kernel Explainer (model-agnostic approximation method) for any model.

LIME (Local Interpretable Model-Agnostic Explanations) [41] is another XAI model which, unlike the SHAP model, focuses on local explanations. The interpretable model is created based on the original instance from which selected components have been excluded. In the case of NLP, this is the removal of some

words.

These two models are often very helpful because they explain in a human-friendly way what features were most important for each class and create graphs that are easily readable and interpretable.

2.3.3 Selected Studies

In addition to the definitions, role and advantages that may arise in the study of explainable AI in medicine, I also focused on the methods used, datasets, and conclusions in two studies of explainability in medical practice.

One of the articles was about opening the black box in predicting stroke outcome [42]. In this research the authors focused on comparing modern ML methods with traditional methods to predict stroke outcome. They presented the first explainability comparison of Modern (nonlinear) ML frameworks: tree boosting (CatBoost) [43] and multilayer perceptrons (MLPs) to traditional (linear) ML frameworks: Generalized Linear Model (GLM), Lasso and Elastic Nets. The main contribution of the study was to use explainability techniques for ML methods to understand the predictions of the models. The deep Taylor decomposition [44] and SHAP values [40] were used for MLP and tree boosting respectively, and model coefficients for logistic regression.

Initially, the study consisted of 514 patients. They were triaged into receiving iv-tissue-plasminogen-activator (tPA) for thrombolysis therapy or conservative therapy. The modified Rankin Scale (mRS) was crucial to the decision about to continue the patient's participation in the study, because in their supervised machine learning framework, they tried to predict the final outcome of stroke patients in terms of dichotomized 3-months post-stroke mRS, where mRS equal to 0, 1, or 2 indicates a good outcome and mRS equal to 3, 4, 5, or 6 indicates a bad outcome. After pre-processing the final number of patients was 314. The features taken into account for the study were: age, sex, NIHSS (National Institute of Health Stroke Scale - stroke severity), history of cardiac disease, history of diabetes mellitus, presence of hypercholesterolemia, and thrombolysis treatment.

The dataset was split into training set (80%) and test set (20%). To adjust for imbalance of the target class, the training set was randomly sub-sampled to get an uniform class distribution. For tuning the models they used 10-fold cross validation method and repeated the process 50 times.

The model performance was tested using receiver-operating-characteristic (ROC) analysis by measuring the AUC (area-under-the-curve). The AUC values on the test set for all models were: GLM - 0.83, Lasso - 0.83, Elastic Nets - 0.81, Tree boosting - 0.81 and MLP - 0.83. All explainable models indicated age and NIHSS as the most important features.

In this article, the authors focused their discussion on two techniques: artificial neural networks and tree boosting. They wanted to show that these algorithms show high performance in medical tasks. In this study, tree boosting will achieve comparable performance to deep learning techniques, and are much easier to train. ANNs, on the other hand, although they were not better performing than the linear models, showed promising results in many other areas of healthcare [45]. They confirmed that modern techniques are able to provide reliable feature validity assessments comparable to their traditional counterparts for clinical predictive models. Researchers have highlighted the fact that modern methods are overly advertised and used where traditional techniques can produce the best results. The goal was to select the method that seems best suited to the classification task. The researchers' results and conclusions encourage the development of explainable clinical predictive models. Future work should use explanatory methods, further explore the differences between them, and test different predictive modeling frameworks.

The second research was carried out by Janizek, Celik and Lee [46]. They introduced their method for explainable prediction of synergistic drug combination for cancer medicine (possible drug combinations and their effectiveness). Because of the fact that the space of possible drug combinations is very extensive, machine learning in this study can be not only helpful but also revealing. The dataset for this study consisted over 22,000 samples, where each sample is one of the 583 two-drug combinations tested in 39 cancer cell lines.

They used a relatively new ML library - XGBoost [47] used to implement “efficient, flexible and portable” gradient boosted trees. The TreeCombo model, which is based on extreme gradient boosted tree (XGBoost). It is used to predict the synergy of new drug combinations using the chemical and physical properties of drugs and gene expression levels in cell lines as features. In contrast, they used the TreeSHAP model to interpret the predictions. The algorithm computes exact tree solutions for SHAP values. The feature attribution values are guaranteed to be unique, accurate and consistent solutions (i.e., their value never decreases when the true impact of that feature is increased). Experimenting with different models was an important part of the study. They compared TreeCombo to: Elastic Net - a linear regression method with regularization (scikit-learn implementation), Random Forest - which uses tree ensembles like TreeCombo (scikit-learn implementation), and DeepSynergy [48] - which uses deep neural networks (Keras implementation [49] with TensorFlow backend [50]). Testing was done using 5-fold cross validation.

For each model, they tuned the parameters. For the TreeSHAP model the parameters that proved most efficient were: maximum tree depth of 6, learning rate of 0.05 and 1000 estimators, with an early stopping parameter. TreeSHAP was then used to calculate feature importance values. They checked the feature importance for each prediction in each test fold and trained the models using only the n most important features, for different numbers n . TreeCombo performance decreased only slightly even though most of the least important features were removed.

To evaluate TreeCombo’s performance, they compared it with: ElasticNet, Random Forests, and DeepSynergy using two different evaluation measures: (1) mean squared error (MSE) and (2) rank correlation of actual synergy results with predicted synergy results. The prediction quality was averaged over five test folds in a five-fold cross-validation experiment. The synergy distributions were well reproduced and the median MSEs were very similar between the predicted and actual results. To test how well their model predicted the synergy ranking of different drug combinations, they used Spearman correlation between TreeCombo

predictions and actual synergy results. The drug combination ranking was not predicted worse in the high MSE cell lines, and the correlations were fairly consistent across all cell lines (ranging from 0.6 to 0.75).

The main advantage of using a tree-based method for data modeling was the ease of model interpretation using the TreeSHAP method. SHAP values for all features were calculated for the five models trained for TreeCombo (one for each of the five test folds held). For each individual model, they selected the most important features. Using the best 1000 and best 2000 features they trained the models again. Performance was retained only using the smaller set of features (1000). There was a slight increase in average MSE for the 2000 features.

The purpose of this study was to identify drug combinations and to understand why the model predicts that the synergy of these combinations will be high. In the discussion section, the authors highlighted the importance of using the TreeSHAP method for tree-based outcome prediction. By doing so, they gained the ability to train a comparably efficient model using only 11% of the data. By using explainable methods, they were able to check features that were not relevant to their study.

2.3.4 Challenges

In the article written by Ghassemi, Oakden-Rayner and Beam [13], explainable AI models in health care are presented in a completely different way. The authors are giving the examples when explainability techniques caused problems for decision making. They disagree with the position that explainability methods for AI models will instill confidence in health care professionals and provide transparency for complex models that are intended to expedite diagnosis or treatment. According to them, this is a false hope for explainable AI and that current explainability methods are unlikely to achieve these goals for patient-level decision support.

One example described in the article is heat maps used in medical imaging. Heatmaps highlight which region of an image and how much it contributed to a particular decision. Because they are illustrative they provide a simple way to understand some of the limitations of post-hoc explanatory techniques [51]. Even

the hottest parts of the map contain both useful and non-useful information (from the perspective of a doctor). If the model locates a region it does not say exactly what is useful in that area. This is the interpretive gap of the model. The doctor gives the labels and explanations for the problem, so the model will not always deal with the current interpretation. In the case of heat maps, this interpretation is not useful for physicians because the areas detected by the model are not fully explained.

Another example in the limitations of interpretability of models is the problem with contextual language models that are trained on large datasets e.g. SciBERT [52] (a pretrained language model for scientific text). It turns out that sources which appear to be suitable for models often have inappropriate and problematic associations with certain human differences such as gender or race [53]. AI models often create mental shortcuts for decision making and thus can be biased and incorrect (in the terms of the model's behavior).

In addition, many other approaches have been developed to create explanations for complex medical data, such as feature visualisation and prototype comparisons. Feature visualisation involves creating synthetic inputs that most activate specific parts of a machine learning model [54]. Synthetic input data rarely corresponds to the same as human-interpreted data. An example would be synthetic inputs that look more or less similar. In this case, the model will make a decision, even though it may be premature. In contrast, the physician may have doubts and make a completely different decision, e.g. ordering other medical examinations.

Locally interpretable model-agnostic explanations and Shapley values also raise interpretability issues. In the case of image analysis, LIME tries to understand the decision by obscuring parts of the image and the explanation consists of a heat map that shows which elements are most important. These are interpretability gaps that involve the fact that LIME and SHAP are general models, not necessarily for image interpretation. For my research, where I am using textual data, the models are appropriate because they are designed for text classification purposes.

The authors explain and conclude that these examples show that explanations

do not guarantee effectiveness. The performance of explanations is rarely tested. The reliability of an explainable AI model is usually assumed in advance. Explanations of black-box models can be wrong so we should not accept their predictions without any scrutiny. The key is to focus on accurate and rigorous validation of explainable systems in as many diverse and disparate populations as possible, rather than requiring local explanations from a complex AI systems.

The examples described above show the problems that occur when applying explainable AI in medicine. These examples are mainly related to reading and interpreting image data, which in the case of my research will not be of much relevance as I am using text data. However, it is important to analyze and test the results of explainable models, so that the results themselves are not the only measure of quality.

3 Experimental Setup

This section describes experimental setup of the work. First, the datasets of the two target studies are described. Thereafter, the study design I use in the thesis is explained and its framework is illustrated.

3.1 Clinical Datasets

During the dataset search phase, I applied for an access to the data from the DBMI Data Portal (<https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/>). They have data and computing resources from the Department of Biomedical Informatics (DBMI) in the Blavatnik Institute at Harvard Medical School. The majority of these Clinical Natural Language Processing data sets were originally created at National Center Biomedical Computing (NCBC), founded by a former National Institutes of Health (NIH), known as i2b2: Informatics for Integrating Biology and the Bedside.

In the first stage of the study I focused on two articles: Identifying Patient Smoking Status from Medical Discharge Records [55] and Recognizing Obesity and Comorbidities in Sparse Data [56] for which datasets from the DBMI portal were available. I chose those two because they can be used for clinical decision support systems. As I want to implement decision-making systems in my study, analyzing these datasets was helpful to decide if I can use them in my research. Both datasets are split into train and test sets in advance. Each of them contains textual clinical notes and annotations, which are necessary for classification tasks. Both articles described results of NLP challenges. In my target study, I am using only one of these two datasets, which is Identifying Patient Smoking Status dataset, therefore the Recognizing Obesity dataset is described in Appendix A.

The following section describes two clinical datasets which are used for the core of this thesis. The first dataset was made available through the DBMI portal,

while the second was published along with a scientific article.

3.1.1 Smoking Status Dataset

The data consisted exclusively of discharge summaries from Partners HealthCare [57] (a healthcare organization from England) and were prepared for the task: automatic evaluation of the smoking status of patients based on medical records. They preprocessed these records to be de-identified, tokenized, broken into sentences, converted into XML format, and separated into training and test sets. Each record was assigned to a unique patient.

Table 2 presents the distribution of classes for smoking status. Looking at the size of the training and test sets, they are very small, considering the fact that they were used to create the AI systems for classification.

Status	Train	Test
UNKNOWN	252	63
NON-SMOKER	66	16
PAST SMOKER	36	11
CURRENT SMOKER	35	11
SMOKER	9	3
all	398	104

Table 2: Distribution of classes in the smoking status dataset.

The categories were defined as follows:

- A Past Smoker is a patient who was a smoker one year or more ago but who has not smoked for at least one year.
- A Current Smoker is a patient who was a smoker within the past year.
- A Smoker is a patient who is either a Current or a Past Smoker but whose medical record does not provide enough information to classify the patient as either.

- A Non-Smoker is a patient who never smoked.
- An Unknown is a patient whose medical record does not mention anything about smoking.

In case of the smoking status dataset, the majority of the train and test sets was labelled UNKNOWN. This category is the least interesting one for the purposes of the smoking challenge, because it does not include any smoking-related information.

Looking at Table 2 there are only 146 examples in the training set and 41 in the test set. This is very little considering that the model must learn some structure for each class. During analyzing results and discussion of the project I will keep in mind how many examples there are.

To visually represent the dataset (after data preprocessing), I used the word cloud presented on the Figure 7. It allows to find out what words were used most often in clinical notes.

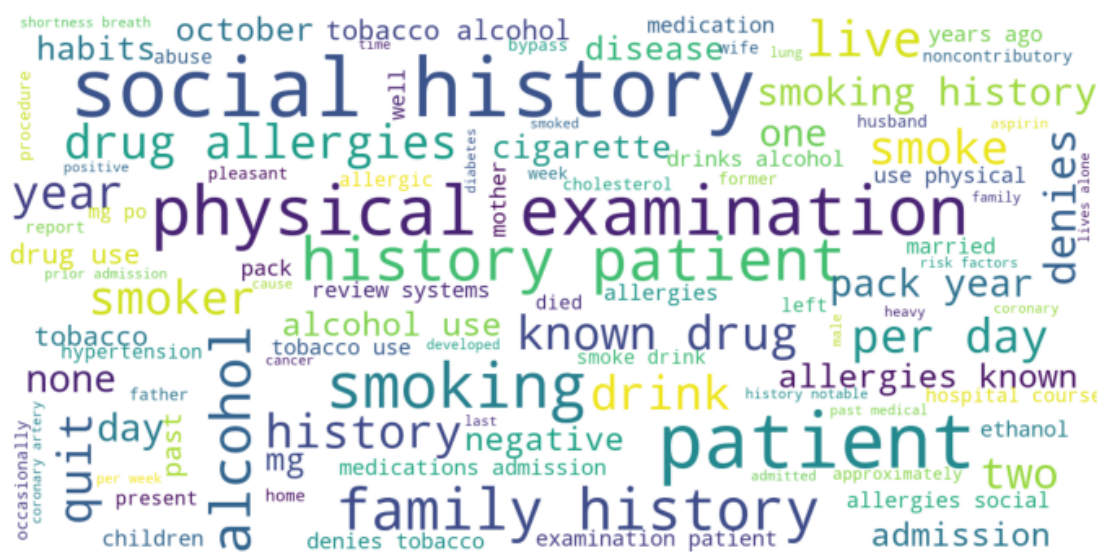


Figure 7: Word cloud for smoking status dataset.

Each clinical note in this dataset describes the patient's condition along with

much additional information, such as: medications the patient is taking, medical test results, information about the patient’s family history, current allergies and more. The most common sections that appeared in the clinical notes are presented in Table 3. The clinical notes in this dataset are relatively long, the longest has 3023 words, the shortest has 94 words, a mean of all clinical notes is equal 774.3 and a median 677.

Section name	Train data	Test data
DIS (stands for “discharge”)	398	104
HISTORY OF PRESENT ILLNESS	234	54
PHYSICAL EXAMINATION	228	49
HOSPITAL COURSE	207	51
PAST MEDICAL HISTORY	206	45
ADMISSION DATE	184	46
DISCHARGE DATE	184	46
PRINCIPAL DIAGNOSIS	162	36
ALLERGIES	151	44
LABORATORY DATA	126	25
SOCIAL HISTORY	114	31
DISPOSITION	111	30
MEDICATIONS ON ADMISSION	105	25
UNIT NUMBER	100	28
DISCHARGE MEDICATION	98	31
MEDICATIONS ON DISCHARGE	97	22
ASSOCIATED DIAGNOSIS	79	16
PRINCIPAL PROCEDURE	74	14
FAMILY HISTORY	67	13
REGISTRATION DATE	63	19

Table 3: Number of occurrences of the 20 most frequent sections in the smoking status dataset.

3.1.2 Acute Ischemic Stroke Dataset

The stroke dataset was published along with the Kim et al. article [58]. This paper assessed performance for classification of brain MRI radiology reports into acute ischemic stroke (AIS) and non acute ischemic stroke (non-AIS) phenotypes. Reports include only patients who had a detected stroke and who were admitted to hospital within 7 days of onset of neurological symptoms. This dataset is public and was reviewed by Institutional Review Boards and Ethics Committee at Chuncheon Sacred Heart Hospital. The hospital stores entire medical records. The dataset is based on all brain MRI reports performed between the 1st of January 2015 and the 31st of December 2016.

In the original method authors split the dataset into train set (70%) and test set (30%). With the scikit-learn “train_test_split” [25] method I have reproduced the same dataset split.

Table 4 presents the distribution of classes for acute ischemic stroke dataset. For this study I was also implemented word clouds for AIS examples (Figure 8) and non-AIS examples (Figure 9) individually to observe which words are the most frequent among the given class. These classes are also differ in the length of report. The shortest report for AIS class contains 17 words, the longest has 370 words, a mean is equal 76.3 words, and a median 73 words, whereas for non-AIS class the shortest has 3 words, the longest 257 words, a mean is equal 41.6, and a median 33 words.

Class	Train	Test
Non-AIS	1814	778
AIS	302	130
all	2116	908

Table 4: Distribution of classes in the stroke dataset.

3.2 Study Design

In my study, I use the framework from Figure 10 to analyze different scenarios for the two classification tasks discussed in the section 4. The purpose of my work will be to look at each step and transitions between steps of this framework, discuss them and find any contradictions that may arise depending on the task. I will use the literature study I discussed in section 2.3.1 so that decisions are most appropriate to the task, and the physician’s knowledge. I will also try to consider any additional steps or transitions that might help to make the right decision.

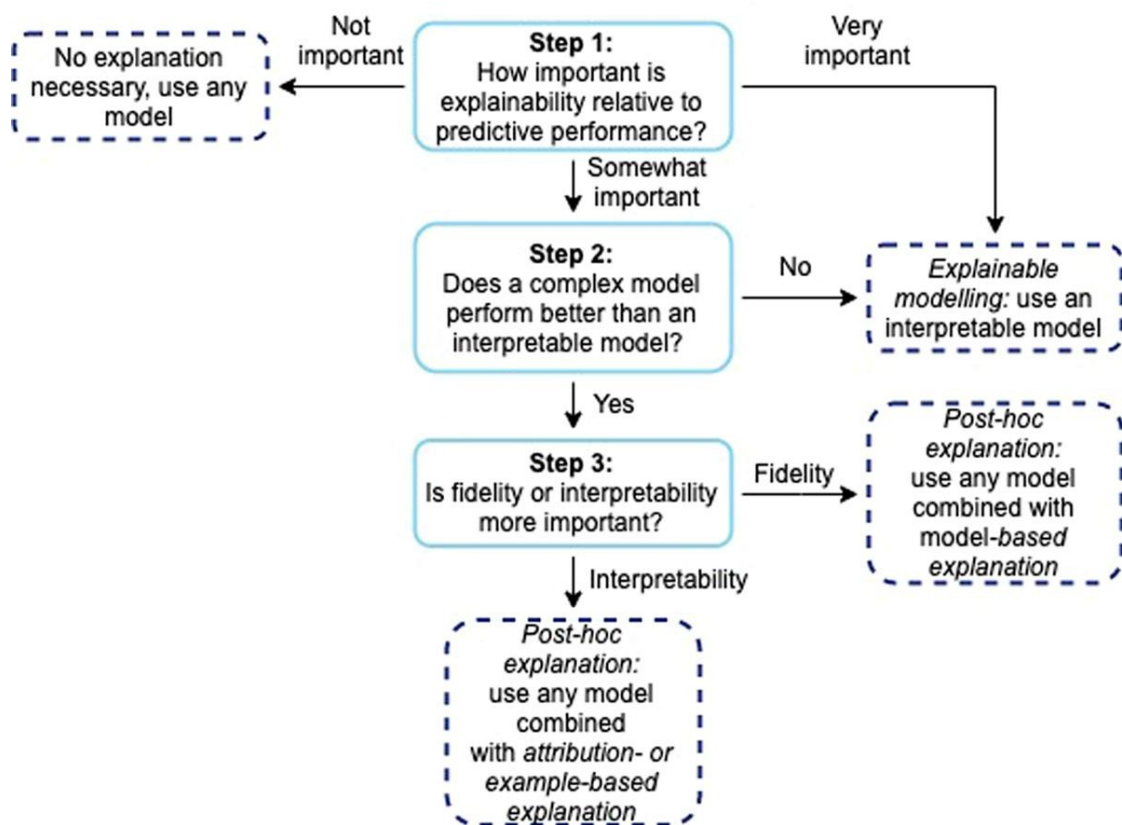


Figure 10: Proposed framework with recommendations to choose amongst explainable AI methods. Reprinted from ‘The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies’ by Markus, Kors and Rijnbeek [34].

In this section I will discuss the steps and possible transitions in general, while in the sections 4.1 and 4.2 I will choose and describe the best scenario for a given task.

The graph starts with the question from step one, i.e: “How important is explainability relative to predictive performance?”. In this step, I will determine whether my task is based only on the best results without any knowledge of how the model works or whether it is at least somewhat or very important that an explanation of the model is present. Following the fact that my work focuses on the medical field, I will rule out the answer “not important”. If the answer is “very important” this framework says to use interpretable model. If the importance of explainability is only somewhat important then the transition leads us to the second step and the question: “Does a complex model perform better than an interpretable model?”. To discuss this question I will implement various complex and interpretable models and compare results. If a complex model perform better the transition leads to the third step and the question: “Is fidelity or interpretability more important?”. To answer this question, it will be essential to know how to distinguish between the two words. The definitions of the words: interpretability and fidelity were described in section 2.3.1. The information about model-based and example-based explanations are included in the same section.

At first glance, the graph appears clear, simple and complete. In the next sections, I will use this methodology for classification and describe any objections about this framework that will apply to the tasks.

4 Study Execution & Results

In this section I will discuss all the methods for preprocessing, methods of word representation, models used and their results. Additionally I will use the framework from section 3.2 to decide which path should be chosen to come to a reasonable decision.

4.1 Identifying Patient Smoking Status

As I mentioned in the previous section a question from the Step 1 has, according to the schema presented before, two possible answers: “Somewhat important” and “Very important”. I could start a short discussion when the first and when the second answer should be considered. While the problem in this task is a decision by the model what is a smoking status of a patient, both answers can be correct. The most important thing to choose an answer is to think in which case the model will be used. The purpose of this may be trivial, for example, to do some statistics among patients, using the model to suggest which medical examination a doctor should use, or prompting smoking status information to the doctor. On the other hand, the model can be used for decisive moments, for example, a decision about a lung cancer patient’s treatment process based on smoking status. I can assume that the decision will not determine the patient’s life by answering this question with the answer “Somewhat important”.

I will start with the replication of the study to show what I have done in the first stage of the work. In the next sections there will be all results and information for further discussion.

4.1.1 Replication of the study

The smoking status prediction challenge described in Section 3.1.1 focused on automatic evaluation of the smoking status of patients based on medical records,

i.e., smoking challenge. A total of 11 teams participated in the challenge.

For this task, I have replicated the methodology of one of the entrants into the challenge, to ensure data and method understanding and to use as a baseline for the main study in this thesis. For identifying smoking status I chose Cohen's approach [59] and implemented the same methods as he did. My choice of research group for replication was based on the fact that he is one of the researchers with the highest scores for this challenge.

In his approach, he created a word-level task. He used both rule-based and machine learning methods. This approach consisted of 5 steps, which I tried to replicate with high precision in my study.

1. Hot-spot Identification

The first step was to identify words (hotspots) that appeared frequently in the dataset: "smok", "cig", "tobac", "packs", "tob", "nicotine".

Examples that had at least one of the hotspots were taken into a new set taking only a window of 100 words before and 100 words after the identified hotspot. In case the report contains more than one hotspot only the first one found was taken into account. This made the texts shorter and focused more on the part around the word that may be important for classification.

2. Tokenization and Vectorization

In the second step the texts from new train set were preprocessed. The preprocessing consisted of:

- removing all symbols that were not alphabetic letters,
- lowercase texts,
- removing extra spaces,
- word tokenization,

- and filtering out all stopwords (except for “no”, “not”, which according to the author may have meaning for the classifier).

As done by Cohen, after tokenization preprocessed texts were encoded using a bag-of-words approach with binary counting.

3. Labelling all examples without hotspot as UNKNOWN

Examples without identified hotspot were automatically assigned to UNKNOWN class and were not taken into account for further classification.

4. SVM

Support Vector Machine with linear kernel function (as in the original study) was trained with preprocessed new train set (examples with hotspots). One-vs-the-rest (OvR) multiclass strategy were implemented to deal with 5-class classification.

5. Combine results from step 3 and 4

Predictions from two classifications (step 3 and 4) were combined. To show how Cohen’s approach affected the prediction of each class I used a confusion matrix displayed in Figure 11.

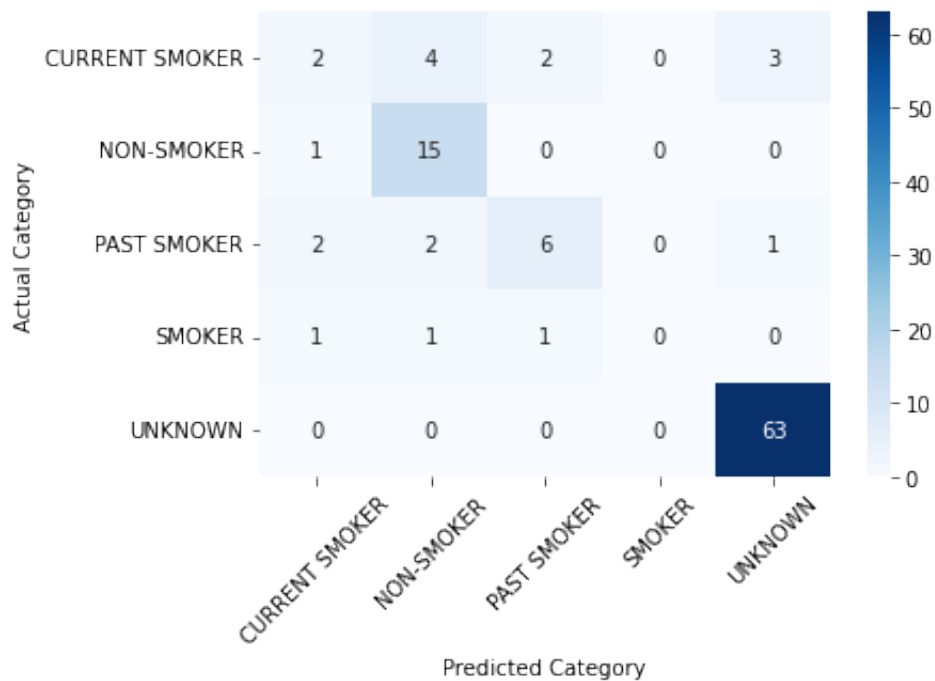


Figure 11: Confusion matrix for Cohen's approach applied to the task of identifying smoking status.

From confusion matrix, presented in Figure 11, we can see exactly how many examples from test set were categorized correctly and how many of them were categorized incorrectly for each class. Excluding the results for SMOKER class due to the small number of examples in dataset, SVM had the greatest difficulties with CURRENT SMOKER class. Only 2 of 11 examples were predicted correctly and the rest were assigned to three others categories.

I used the same evaluation metrics that were used in the challenge, namely precision, recall, f1 (all macro- and micro-averaged). A macro-average value computes a metric independently for each class and represents the arithmetic mean of all classes, whereas micro-average value computes a metric globally [25]. These two approaches are important to show a class imbalance. Micro-average value allowed me to see what the total results are for the model, when it does not matter how well a particular class is classified. The macro-average value, on the other hand, by calculating the average of the metrics for each class, shows whether

the model is able to predict all classes. In case a significant part of the dataset belongs to one class this value displays, for example, whether the model does not assign this class for all examples.

All results are presented in Table 5. The first two rows show Cohen’s results, the rest display results for my replication. It is noticeable that Cohen’s results differ from those obtained by me. After looking more closely and counting the metrics manually using confusion matrix, Cohen’s model correctly predicted only 5 more examples.

	precision	recall	f1
Cohen’s macro avg	0.64	0.65	0.64
Cohen’s micro avg	0.88	0.88	0.88
CURRENT SMOKER	0.33	0.18	0.24
NON-SMOKER	0.68	0.94	0.79
PAST SMOKER	0.67	0.55	0.60
SMOKER	0.00	0.00	0.00
UNKNOWN	0.94	1.00	0.97
macro avg	0.52	0.53	0.52
weighted avg	0.78	0.83	0.80
micro avg	0.83	0.83	0.83

Table 5: Performance measures for Cohen’s approach applied to the task of identifying smoking status.

The replication study allowed me to understand the dataset and methods used in the past for this task. I will use the model described in this section as a baseline for further experiments, and the same metrics: precision, recall, f1, so that I can compare the results obtained with those in Table 5. In addition, it is important to note that none of the research groups did not use the explainable models. This allows me to expand the study and implement explainable AI methods.

4.1.2 Interpretable Models

After replication I focused on interpretable models. For this part I implemented additional models from scikit-learn library [25]: Naive Bayes, Logistic Regression and Decision Tree. Default values were used for each model and tf-idf was implemented as a word representation method. For all models, preprocessing from Step 2 of Cohen’s approach (described in Section 4.1.1) was applied.

Table 6 presents the results of interpretable models. As can be seen in the table, the best micro- and weighted-averaged results were obtained by Logistic Regression. From the results for each category, it can be noticed that Decision Tree succeeded in predicting, to some extent, the largest number of classes among all models. Logistic Regression obtained lower averaged results for the PAST SMOKER class than Decision Tree, but better classified examples belonging to the NON-SMOKER and UNKNOWN classes. The models have significantly lower results than those obtained for Cohen’s approach described in the Section 4.1.1.

	Naive Bayes			Logistic Regression			Decision Tree		
	prec	recall	f1	prec	recall	f1	prec	recall	f1
CURRENT S.	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.09	0.13
NON-S.	0.00	0.00	0.00	0.37	0.44	0.40	0.28	0.31	0.29
PAST S.	0.00	0.00	0.00	0.25	0.09	0.13	0.30	0.27	0.29
SMOKER	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
UNKNOWN	0.61	1.00	0.75	0.73	0.94	0.82	0.66	0.75	0.70
macro avg	0.12	0.20	0.15	0.27	0.29	0.27	0.29	0.28	0.28
weighted avg	0.37	0.61	0.46	0.52	0.64	0.57	0.50	0.54	0.51
micro avg	0.61	0.61	0.61	0.64	0.64	0.64	0.54	0.54	0.54

Table 6: Performance measures for interpretable models applied to the task of identifying smoking status.

4.1.3 Non-interpretable Models

To see if other models will give better results than the models in the previous sections I was experimenting also with more complex models: tf-idf with dense Neural Network, doc2vec with dense Neural Network, and Longformer. Neural network consists of 1 input and 2 hidden layers. The input layer has 256 nodes, the first hidden layer has 128 nodes, the second has 50 nodes. Each hidden layer uses ReLU activation function and includes Dropout regularization to remove simple dependencies between the neurons and increase the robustness of the model. The output layer has 5 nodes (due to the number of classes) and uses Softmax activation function. The architecture of the neural network model was inspired by previous classification assignments in courses I participated during my studies and was intended to be a simple architecture on which I could test the differences in results for complex and interpretable models. The models (for both tf-idf and doc2vec methods) were trained in 5 epochs with a batch size of 64 and optimizer Adam. For the Longformer model, I used the parameters shown in Table 7.

num epochs	train batch size	gradient accumulation steps	eval batch size	evaluation strategy	warmup steps	learning rate	weight decay
3	5	64	5	epoch	1500	2e-5	0.01

Table 7: Parameters for the Longformer model.

Table 8 provides macro-averaged, weighted macro-averaged and micro-averaged results for three other approaches. The models, while complex and often accurate in many tasks, in this case are not even close to the results from Table 5. It is important to remember that due to unequal distribution of classes and small number of instances in the dataset models had more difficulties with classification. Due to the higher results obtained for the neural network with the tf-idf method than doc2vec method, experiments with interpretable models and word representation using doc2vec were not part of the research.

	Tf-idf + NN			Doc2vec + NN			Longformer		
	prec	recall	f1	prec	recall	f1	prec	recall	f1
macro avg	0.12	0.20	0.15	0.21	0.25	0.22	0.20	0.22	0.20
weighted avg	0.37	0.61	0.46	0.48	0.53	0.50	0.41	0.59	0.47
micro avg	0.61	0.61	0.61	0.53	0.53	0.53	0.59	0.59	0.59

Table 8: Performance measures for non-interpretable models applied to the task of identifying smoking status.

Analyzing all three tables: 5, 6, and 8, there is no doubt that Cohen’s approach is the best one for identifying patient smoking status task. With these results, we can assume that the answer for the next question from the framework is “Yes”, because SVM model (which is not intrinsically interpretable) performs better than interpretable models. The framework of Markus et al. asks us next: “Is fidelity or interpretability more important?”. The answer to this question will be addressed in the next section.

4.1.4 XAI Methods

Going into next step from the framework we can assume that for medical field more important is interpretability because it is essential that human (doctor) can understand an explanation. It provides us to the last step: “Post-hoc explanation: use any model combined with attribution- or example-based explanation”. As the best results for this task were obtained by SVM (Cohen’s approach) I will use explainability AI only for this model.

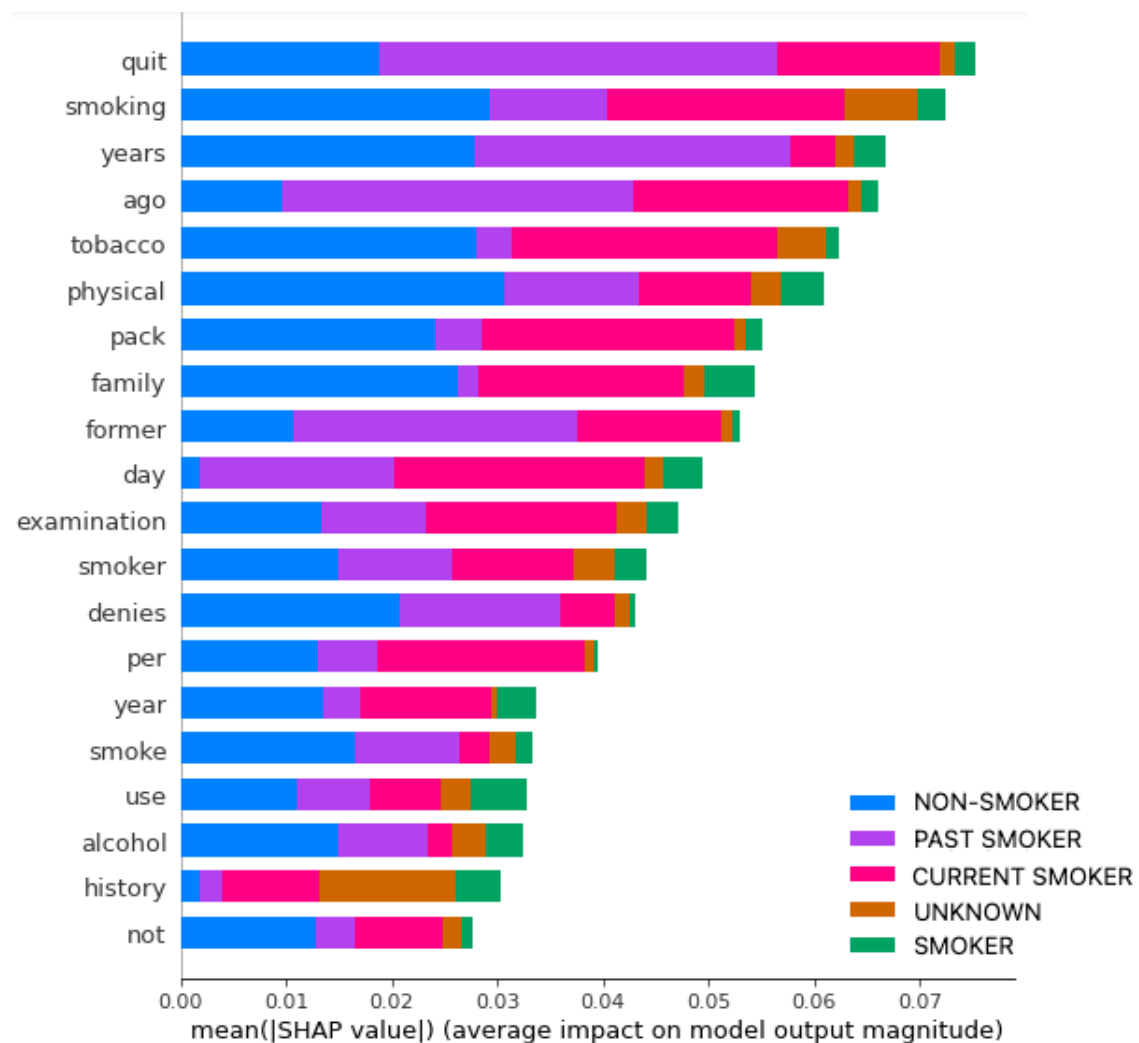


Figure 12: SHAP explanation for Cohen's approach applied to the task of identifying smoking status.

The first method I used, which is well known by the XAI researchers, is SHAP [29]. This method, as I mentioned in Section 2.3.2, explains the model globally. To use SHAP for the Support Vector Machine it is appropriate to apply the Kernel Explainer which uses a weighted linear regression to compute the importance of features. In Figure 12 the top 20 words importance are presented for each class.

Looking at some of the words given, we can see that the model correctly distinguishes classes at the word level. For the class NON-SMOKER the model indicates words: “not”, “denies”, “family”; for the class PAST SMOKER: “quit”, “years”, “ago”; and for the class CURRENT SMOKER: “day”, “per”, “pack ”. I did not take SMOKER and UNKNOWN classes into account in the graph analysis, because they represent a small percentage of all examples when training the model.

To show how XAI model deals with local predictions I used LIME [41]. I have taken 6 examples, 3 examples in which the model correctly predicted the class (one from each significant class: CURRENT SMOKER, NON-SMOKER, PAST SMOKER) and 3 examples in which the model incorrectly predicted the class. LIME aims to present prediction probabilities for each class, top words which that have had a negative or positive effect on the class, and part of the text with highlighted words.

Figure 13 presents LIME explanation for the correctly predicted CURRENT SMOKER class. Prediction probability is high at 70%. From the extracted text with highlighted words it may be concluded that the person is a current smoker. It can be seen mostly in part: “one pack per day smoker last forty years”.

The second example is LIME explanation for the correctly predicted NON-SMOKER class presented in Figure 14. The probability distribution is slightly more dispersed than in previous example: CURRENT SMOKER - 4%; PAST SMOKER - 16%; SMOKER - 15%, but the model almost as highly predicts the correct class - 63%. From the text shown in the figure above the most highlighted text excerpt: “no history tobacco use diabetes” best shows the correctly predicted class.

The last LIME explanation example for the correctly predicted PAST SMOKER class is shown in Figure 15. The model with a high probability - 74% predicted correct class. The greenest part from text with highlighted words is: “quit fifteen years ago” which shows exactly what we expect from the LIME model in this example.

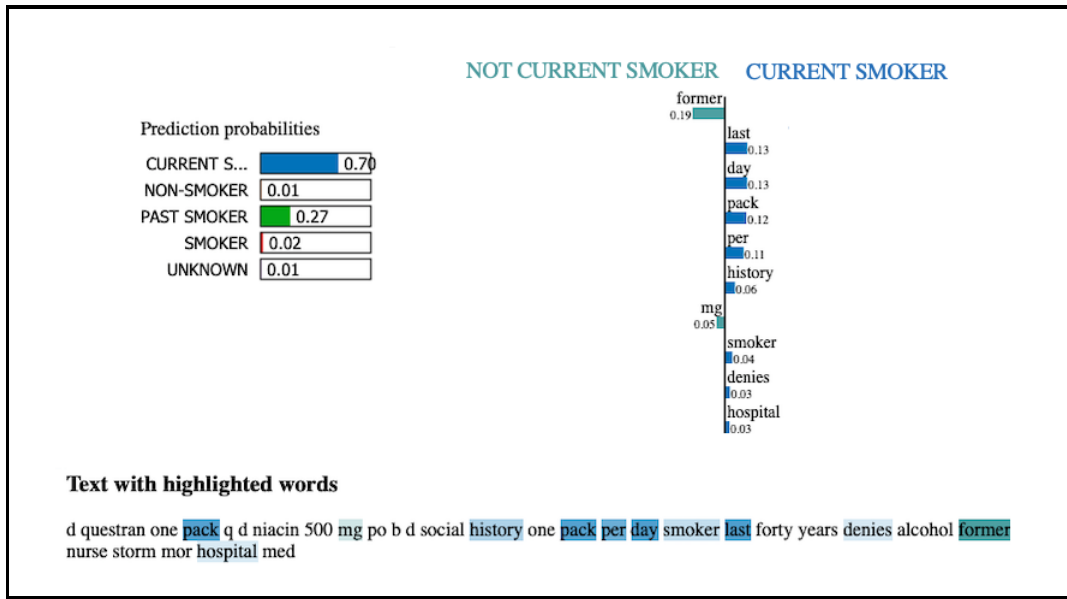


Figure 13: LIME explanation for Cohen's approach for correct smoking status classification (current smoker class).

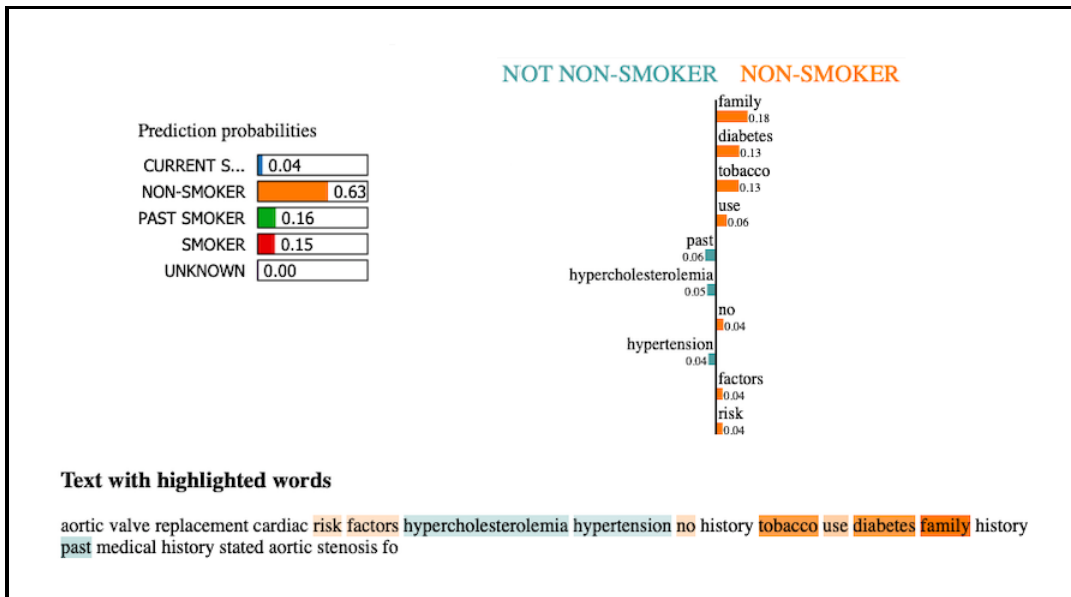


Figure 14: LIME explanation for Cohen's approach for correct smoking status classification (non-smoker class).

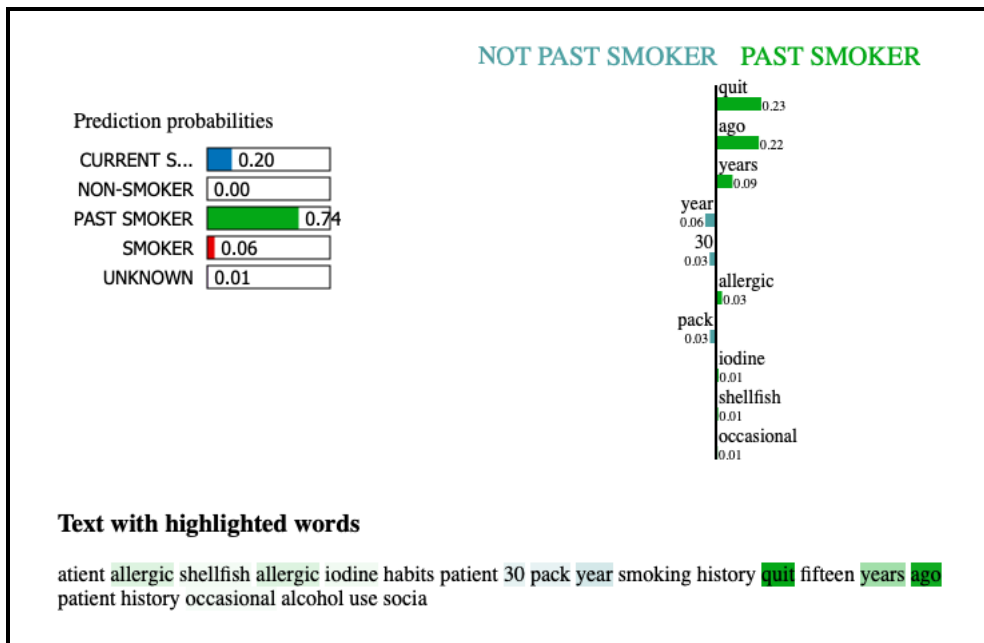


Figure 15: LIME explanation for Cohen's approach for correct smoking status classification (past smoker class).

In order to better understand why the model did not handle some examples and misjudged their class, I decided to select and show LIME explanations for the incorrectly predicted each of the 3 classes.

Figure 16 presents LIME explanation for the incorrectly predicted CURRENT SMOKER class. The example belongs to the PAST SMOKER class. The model predicted the CURRENT SMOKER class with very high probability (70%). From the text with highlighted words, many words indicate that the patient is a smoker. The most highlighted words are phrases such as: "heavy smoker" and "per day times". This is a good example of why the model was wrong in its assessment. There are many indications that the person may have been a smoker looking at this excerpt from the text.

Figure 17 shows LIME explanation for the incorrectly predicted NON-SMOKER class. In this example, we can see that the distribution of probabilities is spread over 3 classes with approximate values between 30 and

35%, which shows that the example was difficult for unambiguous classification. This example should be labelled as CURRENT SMOKER, what can be indicated by words that with a high value of feature importance pointed to the side of not non-smoker in the graph: “smoker”, “heavy”, “former”.

The last LIME explanation example for the incorrectly predicted PAST SMOKER class is shown in Figure 18. Although the correct class is CURRENT SMOKER, the probability with which the model indicated the PAST SMOKER class is 78%. Looking at the text that is depicted and highlighted in the figure, it explains the behavior of the model. An excerpt from the text that most prevailed on classifying the example into the wrong class was “smoking two years ago”, which is very confusing for the model.

Taking all graphs into consideration it can be stated that these explanations are human-interpretable and can be easily used and analyzed by doctors. Despite the small dataset the results and explanations are satisfactory and faithful for both correct and incorrect examples.

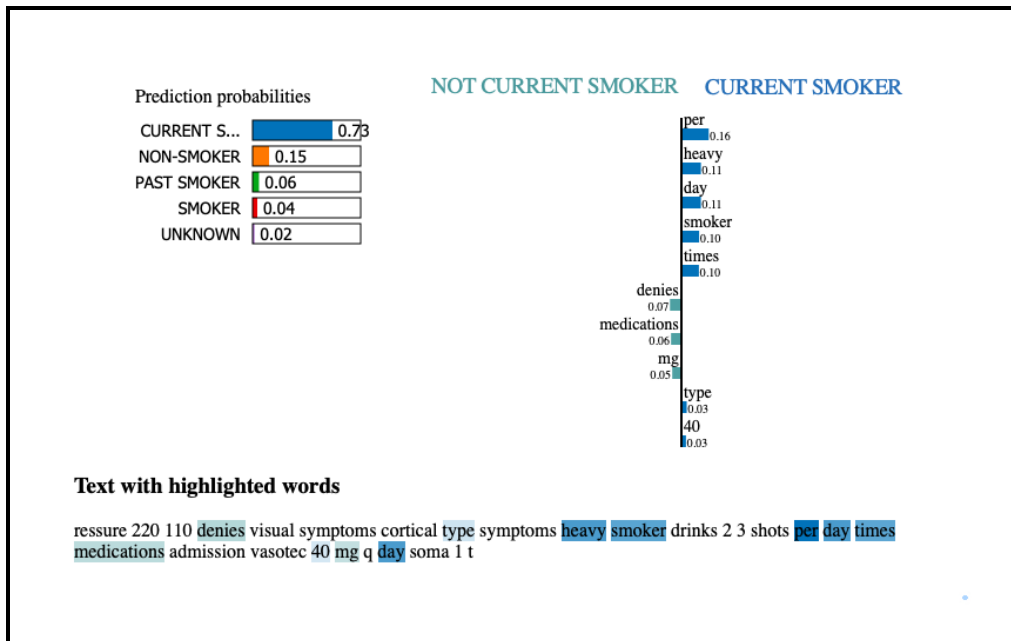


Figure 16: LIME explanation for Cohen's approach for incorrect smoking status classification (current smoker class).

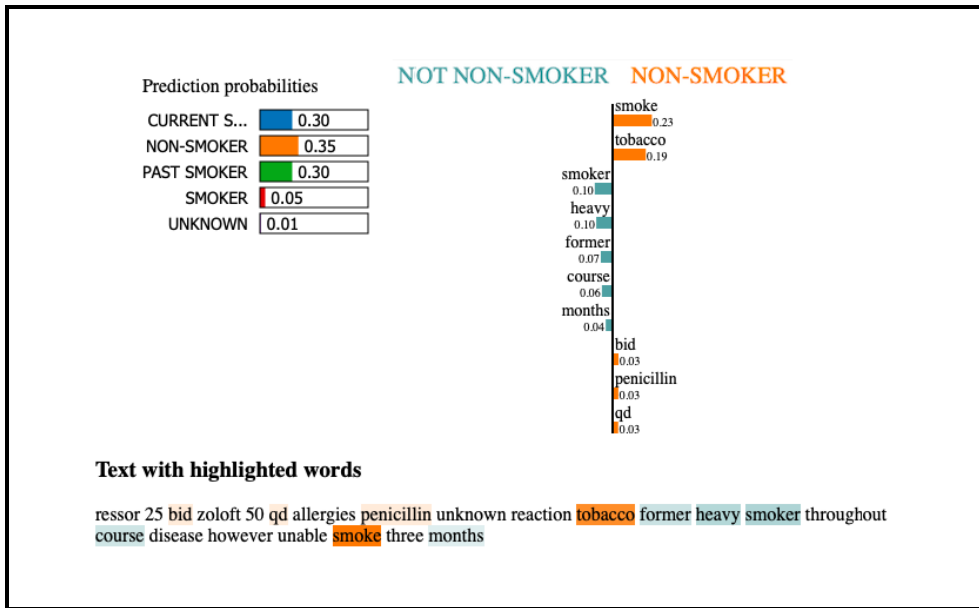


Figure 17: LIME explanation for Cohen's approach for incorrect smoking status classification (non-smoker class).

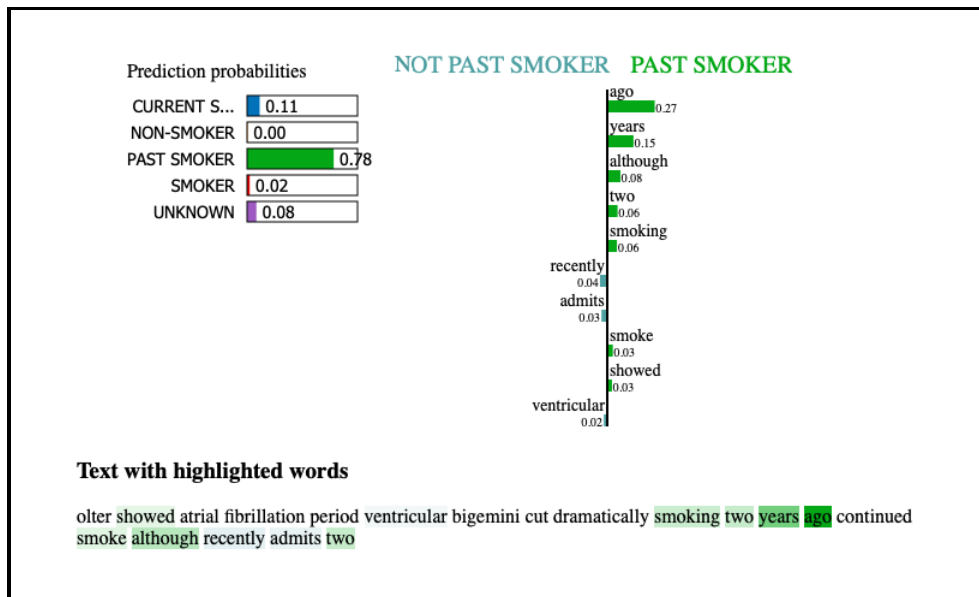


Figure 18: LIME explanation for Cohen's approach for incorrect smoking status classification (past smoker class).

4.2 Predicting Acute Ischemic Stroke

For the Acute Ischemic Stroke (AIS) prediction, for which the dataset was described in Section 3.1.2, the first question from the framework can be answered the same as for Identifying Smoking Status described in section 4.1. If we look at the model as a tool that is to help speed up the decision-making process and not to decide about the patient on its own, then we can always answer this question as “Somewhat important”. To answer the rest of the questions I will use the results from the following subsections.

4.2.1 Replication of the study

In this task I was also started with replication of the original study. Kim et al. applied 4 algorithms: logistic regression, naïve Bayesian classification, decision tree, and support vector machine. They assessed performance of the models by F1-measure. They used 4 types of word representation: unigram (a sequence of 1 word) + BoW, unigram + tf-idf, bigram (a sequence of 2 words) + BoW, and bigram + tf-idf. For each model, I used default values from the scikit learn library [25].

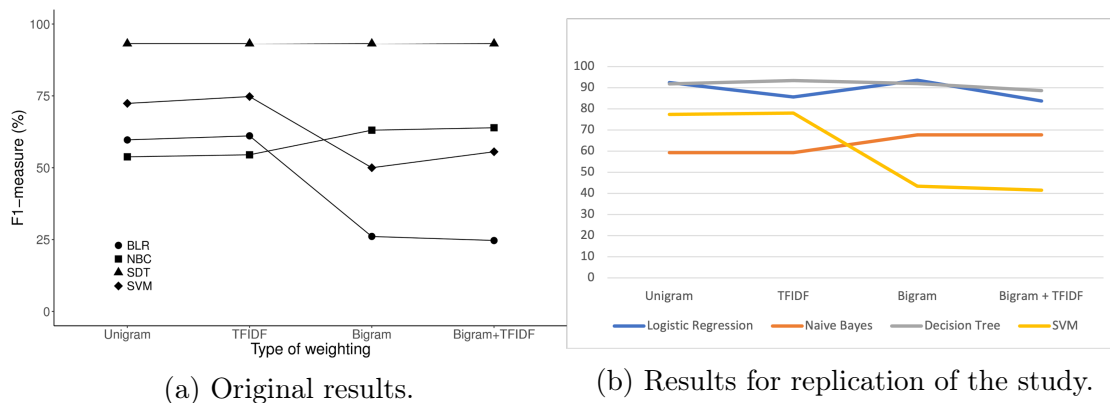


Figure 19: Comparison of ML and NLP algorithms for classifying the brain MRI reports. Reprinted from ‘Natural language processing and machine learning algorithm to identify brain MRI reports with acute ischemic stroke’ by Kim, Zhu, Obeid *et al.* [58]

Figure 19 presents graphs with results obtained by Kim et al. results (Figure 19a) and in the replication (Figure 19b). The best results were achieved by the decision tree model, which fluctuate around 90% for each vectorization method. As can be seen from the graphs, for the models: naive Bayes, decision tree, and support vector machine the trends in the graph look very similar. Only for logistic regression the results of my replication differ significantly from the original results. In my study the results for logistic regression are very high - approximately 90% with BoW and 85% with tf-idf. Kim et al. obtained around 60% for unigram methods and 30% for bigram methods. From the article, I have not be able to get details on the parameters implemented for the model, so the reason for the difference is unknown.

4.2.2 Interpretable Models

For predicting AIS task I implemented the same models as for identifying smoking status, which are: Naive Bayes, Logistic Regression and Decision Tree. Tf-idf was the word representation method for the models, and the models themselves were trained on default values. As for previous task, the preprocessing from Step 2 of Cohen's approach (described in Section 4.1.1) was applied.

The comparison of these three models are provided in Table 9. There are no concerns that the interpretable model with the best overall results is the Decision Tree. This model practically flawlessly predicts non-AIS class. For the AIS class, although the recall is 0.97, the precision is slightly lower than for the Logistic Regression model. Since the weighted macro-averaged and micro-averaged results are better for Decision Tree, this model will be used in next sections to compare the results of interpretable and non-interpretable models.

	Naive Bayes			Logistic Regression			Decision Tree		
	prec	recall	f1	prec	recall	f1	prec	recall	f1
Non-AIS	0.91	0.98	0.94	0.98	0.99	0.99	0.99	0.98	0.99
AIS	0.79	0.42	0.55	0.96	0.86	0.91	0.89	0.97	0.93
macro avg	0.85	0.70	0.74	0.97	0.93	0.95	0.94	0.97	0.96
weighted avg	0.89	0.90	0.89	0.97	0.97	0.97	0.98	0.98	0.98
micro avg	0.90	0.90	0.90	0.97	0.97	0.97	0.98	0.98	0.98

Table 9: Performance measures for interpretable models applied to the task of predicting acute ischemic stroke.

4.2.3 Non-interpretable Models

For further experiments I chose the same non-interpretable models as in the previous task. These models are: Neural Networks with tf-idf, Neural Network with doc2vec, and Longformer. The results of these 3 models are compared in Table 10. The NN model architecture consists of input Dense layer, two hidden Dense layers with ReLU activation function and Dropout after each hidden layer and output layer with Softmax activation function. Figure 20 illustrates the architecture with number of input and output parameters. The model was trained in 5 epochs, batch size - 64, and Adam optimizer. For Longformer model I used parameters shown in Table 7.

	Tf-idf + NN			Doc2vec + NN			Longformer		
	prec	recall	f1	prec	recall	f1	prec	recall	f1
macro avg	0.94	0.91	0.93	0.82	0.78	0.79	0.43	0.50	0.46
weighted avi	0.96	0.96	0.96	0.90	0.91	0.90	0.73	0.86	0.79
micro avg	0.96	0.96	0.96	0.91	0.91	0.91	0.86	0.86	0.86

Table 10: Performance measures for non-interpretable models applied to the task of predicting acute ischemic stroke.

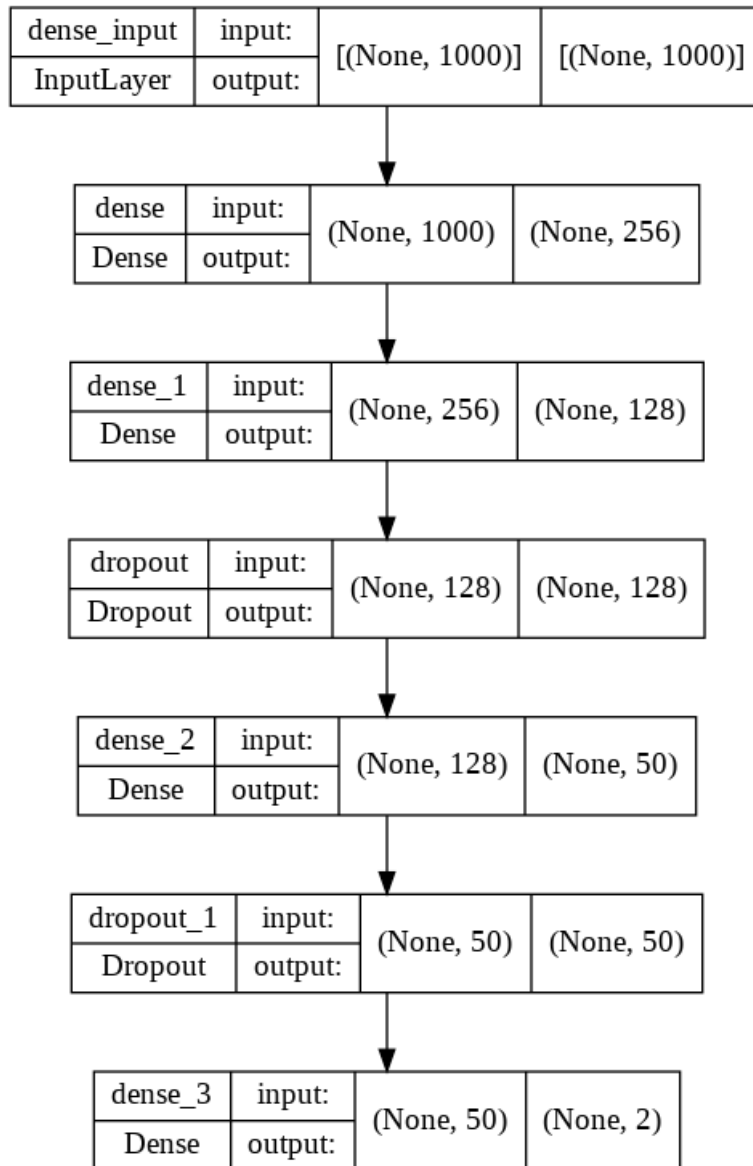


Figure 20: Neural network model architecture.

As can be seen in the Table 9 Longformer obtained the lowest results of all models. Significant differences in the results can also be seen for the NN model with doc2vec. The model which achieved the best results among all non-interpretable models and rivals the Decision Tree is the Neural Network with tf-idf word representation.

Going to results analysis the best model is Decision Tree from section 4.2.3. It provides us to the answer “No” for the question from step 2 of the framework. This contributes to the decision: “Explainable modelling: use an interpretable model”. In the next subsection I will present the visual interpretation of the model.

4.2.4 XAI Methods

For the Decision Tree models without specified `max_depth` parameter (default value) nodes are expanded until all leaves are pure. Figure 21 displays the first seven decision nodes of the Decision Tree. The whole graph is added in the appendix B. Although a Decision Tree model has a natural visualization, this graph is difficult to read, understand and analyze for a human.

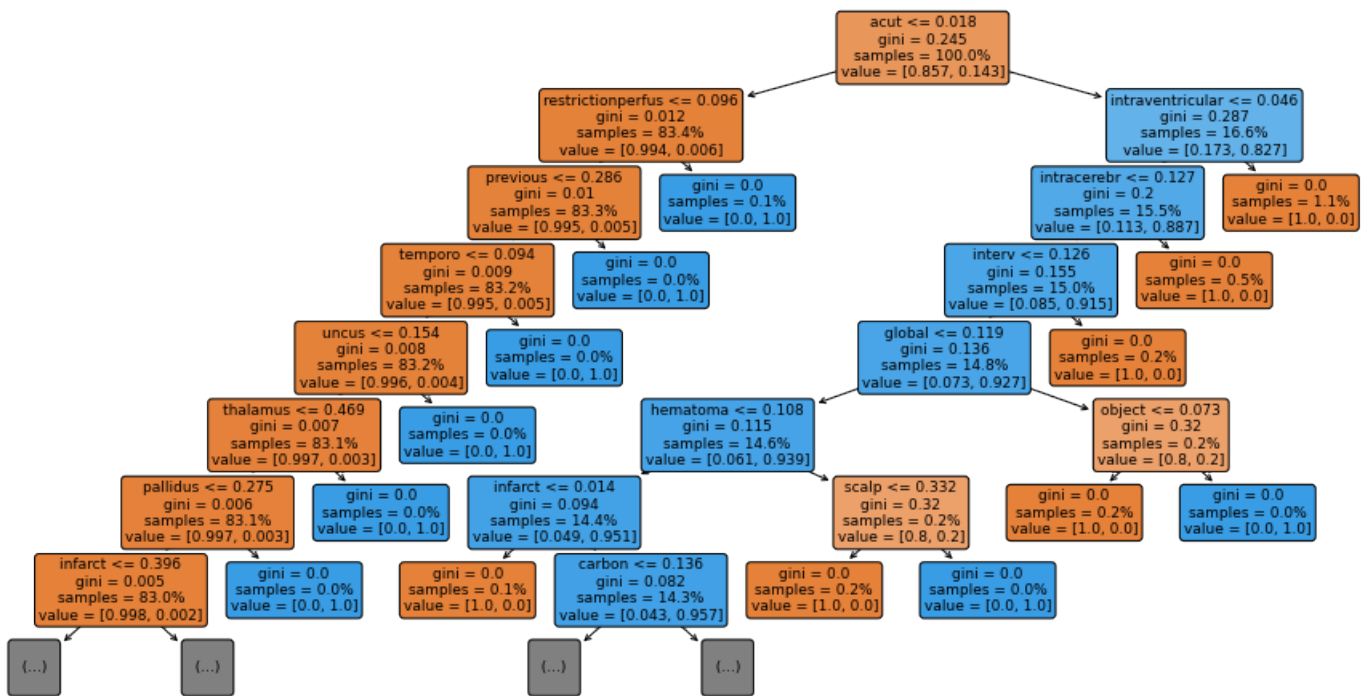


Figure 21: Decision Tree graph for decision tree model applied to the task of predicting acute ischemic stroke (short version).

Due to the fact that the graph is not human-interpretable, I also used SHAP for Neural Network with `tf-idf`. The graph is presented in Figure 22. Although,

this figure shows the most important features in the decision process it does not display the difference for impact for each class. The bars have exactly the same width for AIS and non-AIS classes, which makes no sense during analyzing the graph. Both the interpretation of the decision tree and the neural network are not human-interpretable.

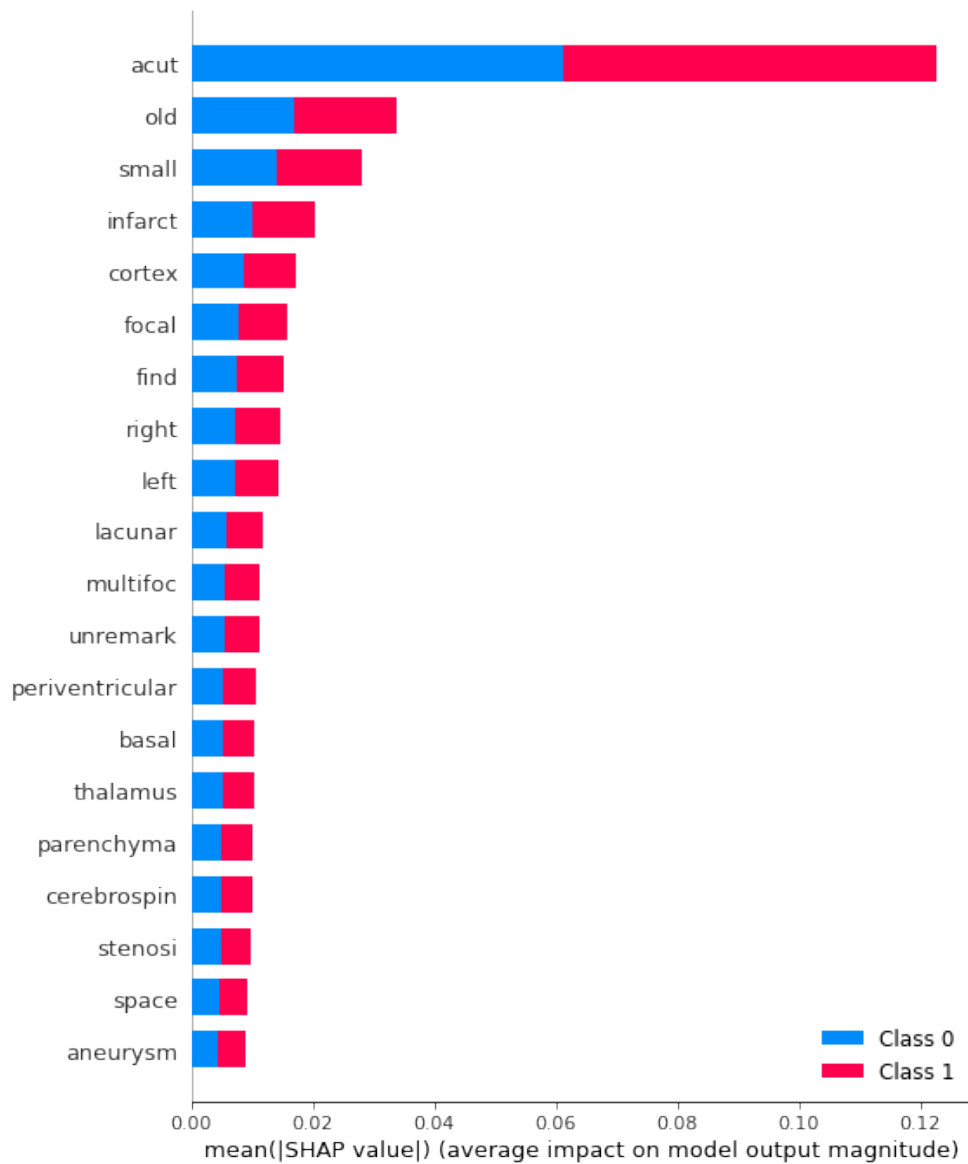


Figure 22: SHAP explanation for neural network model with tf-idf applied to the task of predicting acute ischemic stroke.

5 Discussion

In this section I will explore the results in depth, answer the research subquestions, propose an extended framework with recommendations to choose amongst explainable AI methods, and go into details about the meaning of my findings. In addition, I will point out the limitations of the study.

5.1 Findings

The two studies described in section 4 helped me to understand better how the framework from Figure 10 is reflected in reality and to find answers for the research questions. In the task on identifying smoking status, the path I followed considering all the recommendations and the fact that a complex model performed better than interpretable models led me to the last step: “Post-hoc explanation: use any model combined with attribution- or example-based explanation”. Whereas, following the same inference for the AIS prediction task, the framework led me to step: “Explainable modelling: use an interpretable model”, because an interpretable model performed better than complex models. These results seem reasonable, but the inference may be simplistic in some cases, especially when the field of research is medicine.

Moving on to the main part of discussion - answering the research questions, I will point them out and answer under each question.

Question 1: Is there a single framework for model selection and explanation?

My research work was focused on framework (Figure 10) proposed by Markus et al. [34]. For two tasks: ‘Identifying patient smoking status’ and ‘Predicting acute ischemic stroke’ this framework was easy to analyze, simple, and transparent. Nonetheless, I have noticed some shortcomings. It can be assumed that this framework is sufficient to investigate the models and make decisions but only

if we look at this framework as something flexible. It is good to give yourself space to think about whether the decision that the framework has guided you fits for a particular task, or whether certain steps are missing in a particular case. It should not be seen as something fixed, unchangeable.

Question 2: Can we assume that all explanations are human-interpretable?

From two studies described in section 4 it can certainly be answered that not all explanations are human-interpretable. Even if an explainable model is applied on a task that has a high performance measure, it is not guaranteed that a human will understand and appreciate the explanations provided by the model. In the first study ‘Identifying patient smoking status’ both SHAP and LIME explanations are relevant and easy to interpret. While, in the second study ‘Predicting acute ischemic stroke’ the explanations for decision tree are meaningless (SHAP explanation - Figure 22) and difficult to read (decision tree graph - Figure 24). Importantly, not only complex model can be uninterpretable and unintelligible, but even traditional interpretable models such as a decision tree can be very complex and illegible.

Question 3: Are small datasets (hundreds of examples) of clinical notes are sufficient to implement an effective model and its explainable model?

It is difficult to define a small dataset in machine learning. For different tasks other size of data can be sufficient. Although, if the quality of the data is good, fewer datapoints may be adequate. In machine learning tasks, the more data the better. Unfortunately, in many cases, data with hundreds of examples is given to the research. Usually, researchers have no influence on how much data they receive. The most important is to experiment with various preprocessing steps, word representation methods, machine learning models, and XAI methods. With the right methods, even smaller datasets can prove to be very useful.

During the research, in addition to finding the answers for research questions I also analyzed the framework from Figure 10 in detail. Based on my research of two medical tasks, I looked for ambiguities and possible gaps in the graph.

The extension of the framework is presented in Figure 23. I added “Extra step” after deciding on a post-hoc explanation with the question: “Is an explanation human-interpretable?”. The answers can be “Yes” which leads to the decision: “Keep the previous post-hoc explanation” and “No” which returns to “Explainable modelling: use an interpretable model” decision. In addition, as mentioned in the answer to Question 2, even interpretable models can be unreadable and non human-interpretable. In such a case, when a final decision is made based on the framework, we need to make a final check whether the explanation can help doctors understand the model or is pointless.

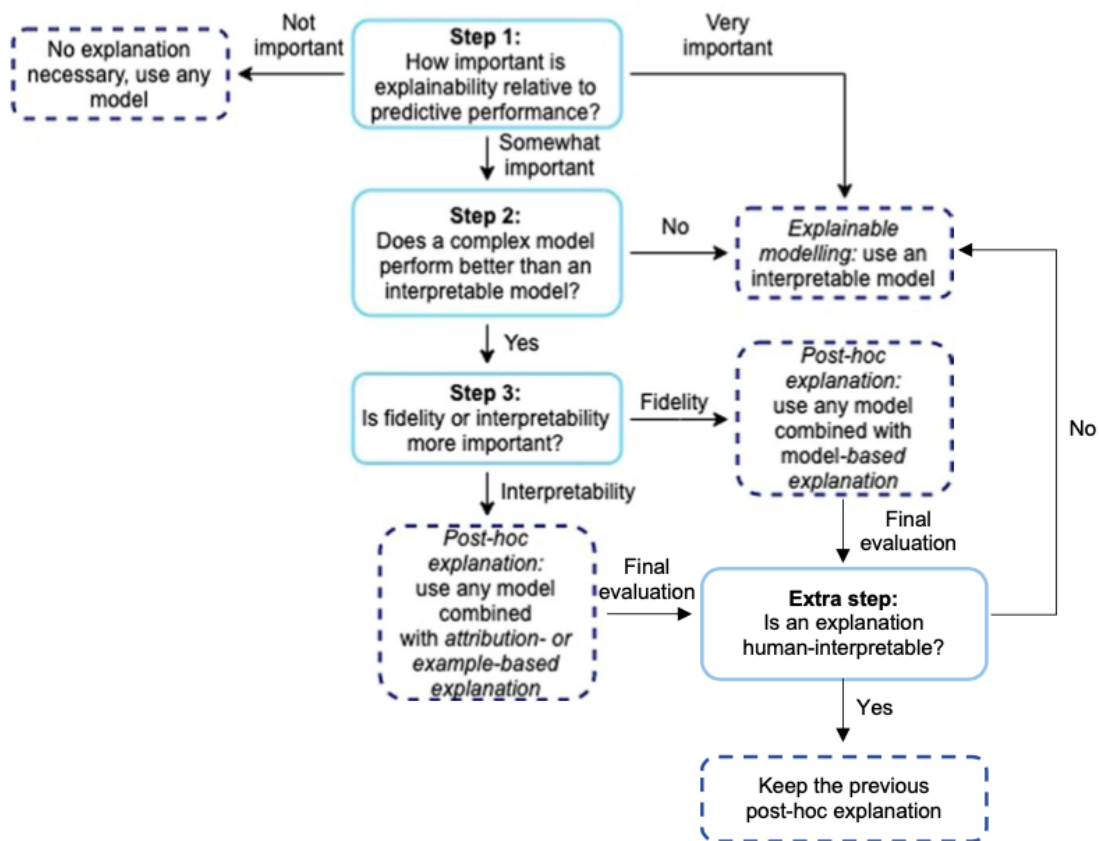


Figure 23: Extension of the framework from Figure 10, with one of the main findings of this study, namely, that explanation models do not always provide human-interpretable explanations. Therefore, evaluations of explanations should always be carried out.

5.2 Limitations

The first limitation of this study is the problem of accessing clinical datasets. As described in the Section 2.1, not many datasets from hospitals, institutes, and clinics are public and available for AI researchers. It was a first problem which occurred during the first stage of the research. The second limitation is a lack of parameters fine-tuning in the study. Since I mainly focused on XAI, I only used default values for all interpretable models, one architecture for the neural network model and one set of parameters for the Longformer model. The last limitation is the usage of only two explainable AI models: SHAP for attribution-based explanation and LIME for example-based explanations. The study could be expanded to include the use of other XAI methods.

Modification of the study could allow me to extend the project and draw new conclusions. In addition, more extra steps in the framework proposed by Markus et al. could be added or a completely new framework proposed.

6 Conclusions

This study set out to understand the views and experiences of explainable AI methods in clinical practice. All previous sections helped me to understand better the importance of explainable AI in medical field and answer the main research question: *Which explainability methods and NLP models should be used in the different classification tasks in clinical practice to obtain satisfactory results and doctors' confidence?*. These findings highlight the potential usefulness of a framework with recommendations to choose amongst XAI methods.

I will start with a recommendation of the framework during a study research, where explainable AI is an important part of the project. While I was working on the project, the framework helped me make decisions and accelerated the selection process. Notwithstanding the relatively limited steps and choices, this framework offers valuable insights into each decision, even if it does not give a precise recommendation on classification or XAI models. There are many classification models available for training and XAI models for explanation, but each dataset and machine learning task requires its own research and analysis process. A framework can only support and facilitate this procedure.

In this thesis, I proposed extension of the framework with an extra step at the end. Although, the framework was valuable during the study it should be considered with care during further research.

References

- [1] J. Behler, ‘Perspective: Machine learning potentials for atomistic simulations,’ *The Journal of chemical physics*, vol. 145, no. 17, p. 170901, 2016.
- [2] T. Wuest, D. Weimer, C. Irgens and K.-D. Thoben, ‘Machine learning in manufacturing: Advantages, challenges, and applications,’ *Production & Manufacturing Research*, vol. 4, no. 1, pp. 23–45, 2016.
- [3] A. Becker, ‘Artificial intelligence in medicine: What is it doing for us today?’ *Health Policy and Technology*, vol. 8, no. 2, pp. 198–205, 2019.
- [4] C. Friedman, G. Hripcsak *et al.*, ‘Natural language processing and its future in medicine,’ *Acad Med*, vol. 74, no. 8, pp. 890–5, 1999.
- [5] S. Khedkar, P. Gandhi, G. Shinde and V. Subramanian, ‘Deep learning and explainable ai in healthcare using ehr,’ in *Deep learning techniques for biomedical and health informatics*, Springer, 2020, pp. 129–148.
- [6] D. P. Pragallapati, ‘Application of ai in medicine and its concerns,’
- [7] D. Doran, S. Schulz and T. R. Besold, ‘What does explainable ai really mean? a new conceptualization of perspectives,’ *arXiv preprint arXiv:1710.00794*, 2017.
- [8] F. Jiang, Y. Jiang, H. Zhi *et al.*, ‘Artificial intelligence in healthcare: Past, present and future,’ *Stroke and vascular neurology*, vol. 2, no. 4, 2017.
- [9] J. A. Sidey-Gibbons and C. J. Sidey-Gibbons, ‘Machine learning in medicine: A practical introduction,’ *BMC medical research methodology*, vol. 19, no. 1, pp. 1–18, 2019.
- [10] D. Ofer, N. Brandes and M. Linial, ‘The language of proteins: Nlp, machine learning & protein sequences,’ *Computational and Structural Biotechnology Journal*, vol. 19, pp. 1750–1758, 2021.
- [11] D. Delen, G. Walker and A. Kadam, ‘Predicting breast cancer survivability: A comparison of three data mining methods,’ *Artificial intelligence in medicine*, vol. 34, no. 2, pp. 113–127, 2005.

- [12] ‘Seer cancer statistics review. surveillance, epidemiology, and end results (seer) program (www.seer.cancer.gov) public-use data (1973—2000). national cancer institute, surveillance research program, cancer statistics branch, released april 2003. based on the november 2002 submission. diagnosis period 1973—2000, registries 1—9.’
- [13] M. Ghassemi, L. Oakden-Rayner and A. L. Beam, ‘The false hope of current approaches to explainable artificial intelligence in health care,’ *The Lancet Digital Health*, vol. 3, no. 11, e745–e750, 2021.
- [14] P. Lewis, M. Ott, J. Du and V. Stoyanov, ‘Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art,’ in *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, 2020, pp. 146–157.
- [15] D. Van Le, J. Montgomery, K. C. Kirkby and J. Scanlan, ‘Risk prediction using natural language processing of electronic mental health records in an inpatient forensic psychiatry setting,’ *Journal of biomedical informatics*, vol. 86, pp. 49–58, 2018.
- [16] V. M. Castro, J. Minnier, S. N. Murphy *et al.*, ‘Validation of electronic health record phenotyping of bipolar disorder cases and controls,’ *American Journal of Psychiatry*, vol. 172, no. 4, pp. 363–372, 2015.
- [17] H. Zou and H. H. Zhang, ‘On the adaptive elastic-net with a diverging number of parameters,’ *Annals of statistics*, vol. 37, no. 4, p. 1733, 2009.
- [18] G. Gorrell, S. Oduola, A. Roberts, T. Craig, C. Morgan and R. Stewart, ‘Identifying first episodes of psychosis in psychiatric patient records using machine learning,’ in *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, 2016, pp. 196–205.
- [19] J. Pestian, H. Nasrallah, P. Matykiewicz, A. Bennett and A. Leenaars, ‘Suicide note classification using natural language processing: A content analysis,’ *Biomedical informatics insights*, vol. 3, BII–S4706, 2010.
- [20] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, ‘Bert: Pre-training of deep bidirectional transformers for language understanding,’ *arXiv preprint arXiv:1810.04805*, 2018.

- [21] P. Mosteiro, E. Rijcken, K. Zervanou, U. Kaymak, F. Scheepers and M. Spruit, ‘Machine learning for violence risk assessment using dutch clinical notes,’ *Journal of Artificial Intelligence for Medical Sciences*, 2021.
- [22] W. de Vries, A. van Cranenburgh, A. Bisazza, T. Caselli, G. van Noord and M. Nissim, ‘Bertje: A dutch bert model,’ *arXiv preprint arXiv:1912.09582*, 2019.
- [23] K. Huang, J. Altosaar and R. Ranganath, ‘Clinicalbert: Modeling clinical notes and predicting hospital readmission,’ *arXiv preprint arXiv:1904.05342*, 2019.
- [24] E. Alsentzer, J. R. Murphy, W. Boag *et al.*, ‘Publicly available clinical bert embeddings,’ *arXiv preprint arXiv:1904.03323*, 2019.
- [25] F. Pedregosa, G. Varoquaux, A. Gramfort *et al.*, ‘Scikit-learn: Machine learning in python,’ *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.
- [26] T. Mikolov, K. Chen, G. Corrado and J. Dean, ‘Efficient estimation of word representations in vector space,’ *arXiv preprint arXiv:1301.3781*, 2013.
- [27] Q. Le and T. Mikolov, ‘Distributed representations of sentences and documents,’ in *International conference on machine learning*, PMLR, 2014, pp. 1188–1196.
- [28] G. Stiglic, P. Kocbek, N. Fijacko, M. Zitnik, K. Verbert and L. Cilar, ‘Interpretability of machine learning-based prediction models in healthcare,’ *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 5, e1379, 2020.
- [29] C. Molnar, *Interpretable machine learning*. Lulu. com, 2020.
- [30] Wikipedia, *Support-vector machine*, https://en.wikipedia.org/wiki/Support-vector_machine.
- [31] D. Abueidda, Q. Lu and S. Koric, *Deep learning collocation method for solid mechanics: Linear elasticity, hyperelasticity, and plasticity as examples*, Dec. 2020.
- [32] A. Vaswani, N. Shazeer, N. Parmar *et al.*, ‘Attention is all you need,’ *Advances in neural information processing systems*, vol. 30, 2017.

- [33] I. Beltagy, M. E. Peters and A. Cohan, ‘Longformer: The long-document transformer,’ *arXiv preprint arXiv:2004.05150*, 2020.
- [34] A. F. Markus, J. A. Kors and P. R. Rijnbeek, ‘The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies,’ *Journal of Biomedical Informatics*, vol. 113, p. 103 655, 2021.
- [35] G. Ras, M. van Gerven and P. Haselager, ‘Explanation methods in deep learning: Users, values, concerns and challenges,’ in *Explainable and interpretable models in computer vision and machine learning*, Springer, 2018, pp. 19–36.
- [36] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan and W.-K. Wong, ‘Too much, too little, or just right? ways explanations impact end users’ mental models,’ in *2013 IEEE Symposium on visual languages and human centric computing*, IEEE, 2013, pp. 3–10.
- [37] O.-M. Camburu, E. Giunchiglia, J. Foerster, T. Lukasiewicz and P. Blunsom, ‘Can i trust the explainer? verifying post-hoc explanatory methods,’ *arXiv preprint arXiv:1910.02065*, 2019.
- [38] F. Doshi-Velez and B. Kim, ‘Towards a rigorous science of interpretable machine learning,’ *arXiv preprint arXiv:1702.08608*, 2017.
- [39] J. Huysmans, K. Dejaeger, C. Mues, J. Vanthienen and B. Baesens, ‘An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models,’ *Decision Support Systems*, vol. 51, no. 1, pp. 141–154, 2011.
- [40] 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*, 2017, pp. 4768–4777.
- [41] 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*, 2016, pp. 1135–1144.

- [42] E. Zihni, V. I. Madai, M. Livne *et al.*, ‘Opening the black box of artificial intelligence for clinical decision support: A study predicting stroke outcome,’ *Plos one*, vol. 15, no. 4, e0231166, 2020.
- [43] ‘Catboost python package,’ <https://catboost.ai>.
- [44] G. Montavon, S. Lapuschkin, A. Binder, W. Samek and K.-R. Müller, ‘Explaining nonlinear classification decisions with deep taylor decomposition,’ *Pattern Recognition*, vol. 65, pp. 211–222, 2017.
- [45] A. Esteva, A. Robicquet, B. Ramsundar *et al.*, ‘A guide to deep learning in healthcare,’ *Nature medicine*, vol. 25, no. 1, pp. 24–29, 2019.
- [46] J. D. Janizek, S. Celik and S.-I. Lee, ‘Explainable machine learning prediction of synergistic drug combinations for precision cancer medicine,’ *bioRxiv*, p. 331 769, 2018.
- [47] T. Chen and C. Guestrin, ‘Xgboost: A scalable tree boosting system,’ in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.
- [48] K. Preuer, R. P. Lewis, S. Hochreiter, A. Bender, K. C. Bulusu and G. Klambauer, ‘Deepsynergy: Predicting anti-cancer drug synergy with deep learning,’ *Bioinformatics*, vol. 34, no. 9, pp. 1538–1546, 2018.
- [49] C. François *et al.*, ‘Keras,’ 2015, <https://keras.io>.
- [50] Martín Abadi, Ashish Agarwal, Paul Barham *et al.*, *TensorFlow: Large-scale machine learning on heterogeneous systems*, Software available from <https://www.tensorflow.org>, 2015. [Online]. Available: <https://www.tensorflow.org/>.
- [51] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, ‘Grad-cam: Visual explanations from deep networks via gradient-based localization,’ in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.
- [52] I. Beltagy, K. Lo and A. Cohan, ‘Scibert: A pretrained language model for scientific text,’ *arXiv preprint arXiv:1903.10676*, 2019.
- [53] H. Zhang, A. X. Lu, M. Abdalla, M. McDermott and M. Ghassemi, ‘Hurtful words: Quantifying biases in clinical contextual word embeddings,’ in *proceedings of the ACM Conference on Health, Inference, and Learning*, 2020, pp. 110–120.

- [54] C. Olah, A. Satyanarayan, I. Johnson *et al.*, ‘The building blocks of interpretability,’ *Distill*, vol. 3, no. 3, e10, 2018.
- [55] Ö. Uzuner, I. Goldstein, Y. Luo and I. Kohane, ‘Identifying patient smoking status from medical discharge records,’ *Journal of the American Medical Informatics Association*, vol. 15, no. 1, pp. 14–24, 2008.
- [56] Ö. Uzuner, ‘Recognizing obesity and comorbidities in sparse data,’ *Journal of the American Medical Informatics Association*, vol. 16, no. 4, pp. 561–570, 2009.
- [57] P. HealthCare, *Partners healthcare®*, 2015.
- [58] C. Kim, V. Zhu, J. Obeid and L. Lenert, ‘Natural language processing and machine learning algorithm to identify brain mri reports with acute ischemic stroke,’ *PloS one*, vol. 14, no. 2, e0212778, 2019.
- [59] A. M. Cohen, ‘Five-way smoking status classification using text hot-spot identification and error-correcting output codes,’ *Journal of the American Medical Informatics Association*, vol. 15, no. 1, pp. 32–35, 2008.
- [60] J. DeShazo and A. Turner, ‘Hands-on nlp: An interactive and user-centered system to classify discharge summaries for obesity and related co-morbidities,’ in *Proceedings of the i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*, 2008.

A Recognizing Obesity: replication of the study

In the first stage of the work, I analyzed another dataset and clinical task from the DBMI portal. The question of this challenge [56] was: Recognizing Obesity and Comorbidities in Sparse Data. A total of 30 teams participated in the Obesity Challenge. This task was focused on automatically extracting information on obesity and 15 of its most common comorbidities from patient clinical notes. The comorbidities were as follows: Diabetes mellitus (DM), Hypercholesterolemia, Hypertriglyceridemia, Hypertension (HTN), Atherosclerotic CV disease (CAD), Heart failure (CHF), Peripheral vascular disease (PVD), Venous insufficiency, Osteoarthritis (OA), Obstructive sleep apnea (OSA), Asthma, GERD, Gallstones/Cholecystectomy, Depression, Gout. Some diseases were present in a very small number of patients.

The data for the challenge consisted of 1237 discharge summaries from the Partners HealthCare Research Patient Data Repository and were annotated by two obesity experts from the Massachusetts General Hospital Weight Center. Each record was assigned to a unique patient.

For each patient, obesity and comorbidities was annotated as:

- Present (Y): the patient has/had the disease,
- Absent (N): the patient does/did not have the disease,
- Questionable (Q): the patient may have the disease,
- Unmentioned (-): the disease is not mentioned in the discharge summary.

On Table 11 we can see the distribution of classes in the dataset for obesity label. In this study are almost twice as many records in the training set and almost 5 times as many in the test set than in the smoking status survey (table 2).

Together, this gives us nearly 2.5 times more patients. As part of the data cleaning, I removed all “Questionable” and “Unmentioned” records, leaving only 553 records from the training set and 447 records from the test set.

Obesity	Train	Test
N	314	255
Y	239	192
-	176	60
Q	1	0
all	730	507

Table 11: Distribution of classes in the obesity dataset.

The data made available by i2b2 were annotated in two ways:

- textual assessments - the clinical note had to have reported explicit information about diseases,
- intuitive assessments - medical professionals’ reading of the information presented in the summaries and using their intuition to judge.

These annotations were saved in two independent datasets. In my replication, I focused on intuitive assessments because it is more difficult for ML models and is more interesting for the study.

The goal of the challenge was to find, compare, and evaluate NLP technologies to combat the obesity epidemic. Each group created systems for both textual and intuitive annotations.

To better understand the data and the learning process, I analyzed the work of one of the 30 research groups and their methods used in the study. I chose DeShazo et al. [60]. They used both rule-based classifier and Support Vector Machine to identify the patient’s diseases. The rule-based model was responsible for finding patterns in the text, which were then used as features for the SVM model.

During replication of this study I used only Support Vector Machine from scikit-learn library [25] with tf-idf word representation. I implemented simplified process to verify the relevance of the rule-based model. My results are not even close to those obtained by DeShazo et al., what can be seen in Table 12.

	Model	precision	recall	f1
micro	DeShazo et al.	0.95	0.95	0.95
	replication	0.59	0.59	0.59
macro	DeShazo et al.	0.97	0.62	0.63
	replication	0.6	0.54	0.48

Table 12: Performance measures applied to the task of recognizing obesity.

Comparing the results, the significant difference is visible. Results suggest that the effectiveness of the system was mainly based on the rule-based model. In this case, I was not able to create a replication of the system, but the results and conclusions of the study will help me with further work. As in the study described in section 4.1.1, no explainable methods were used here.

This task was not considered for further work due to finding and analyzing another dataset described in section 3.1.2 and obtained satisfactory results from the two tasks described in sections 4.1 and 4.2, which were sufficient to undertake a comprehensive discussion and answers to the research questions.

