Text mining of clinical outcomes for medical research: how accurate should it be?

Author:
Zwierd Grotenhuis
Student number:
6271669

Supervisor:
Dr Tuur Leeuwenberg
Second examiner:
Dr Pablo Mosteiro Romero

A thesis submitted in fulfillment
of the requirements for the degree of
MSc. Artificial Intelligence

Department of Information and Computing Sciences
Graduate School of Natural Sciences
College of Science

Universiteit Utrecht

Utrecht University
Utrecht, The Netherlands

October 2022

Author:
Zwierd Grotenhuis

Supervisor:
Dr Tuur Leeuwenberg
Second examiner:
Dr Pablo Mosteiro Romero

Institute:
Utrecht University

CONTENTS

ABSTRACT

In medicine, clinical prediction models are often developed to estimate future risk of patients regarding a certain health outcome (e.g., in-hospital mortality). To develop these models, historic structured data is needed about patient characteristics and the relevant health outcomes. Sometimes the to be predicted health outcome was not recorded in structured data but may be extracted from the textual notes by using text mining. If a text mining model is developed to extract outcome variables from clinical notes, that model can be used to generate the training data for the prediction model. Contemporary research often applies text mining, but the impact of text mining quality on prediction model performances in this setting remains unclear. We performed a simulation study that charted this relationship in a case study of in-hospital mortality prediction in ICUs. We created a logistic regression and neural network prediction model and trained it on data extracted by multiple text mining models with a wide range of performance. We varied the performance of the text mining models by changing the size of the training data used to develop them and by shifting the decision boundary. We found that analysis can be done to determine whether the text mining model performs well enough, or whether more data might be needed for text mining training purposes. We also concluded that shifting the decision boundary of the text mining model can be a viable way to increase prediction model performance, especially when a low amount of training data is used. The knowledge gained in this project may be used to create better performing prediction models using text mining models when training data is limited.

1 INTRODUCTION

1.1 PROBLEM STATEMENT

Since the introduction of machine learning into clinical research, many prediction models have been built for diagnosis and prognosis of clinical outcomes [1, 2]. Traditionally, most of these models are developed using structured data entries, such as lab values, demographic data, height, weight and blood type. More recently, researchers have been incorporating text mining into their prediction models as well, to take advantage of the huge amounts of free text entries (also called 'clinical narratives') available [3–6]. These free text entries consist of clinical notes as well as discharge summaries and can be used to mine outcome variables. For example, Dormosh *et al.* [7] extracted medical outcomes from electronic health records and then trained a prediction model on that extracted outcome. This allowed them to develop a prediction model that predicts an outcome variable that was not originally recorded in a structured manner. However, it remains unclear how accurate text mining models need to be for the prediction models to perform well, and how big of an influence the text mining model has on the prediction model.

1.2 OBJECTIVES

In this project, the main goal is to answer the general question "How accurate does text mining to extract outcome data need to be to create a valuable clinical prediction model?". We will be answering our research questions with a case study, creating a clinical prediction model for in-hospital mortality based on variables collected in the first 48 hours of admission to the intensive care unit. A text mining model will analyze the clinical notes, and extract whether or not mortality is recorded in these clinical notes. The clinical prediction model is trained on the output data of this text mining model. The performance of the clinical prediction model using text mined outcomes will be compared to the clinical prediction model using ground truth data. We will examine how different text mining performances compare to give an insight into how text mining performance relates to prediction performance for a task like mortality prediction. We will do this by artificially adjusting the performance of our text mining model. A quick overview of the pipeline of creating one of these prediction models can be seen in Figure 1.1. A more in-depth explanation of the project setup can be read in Section 3.

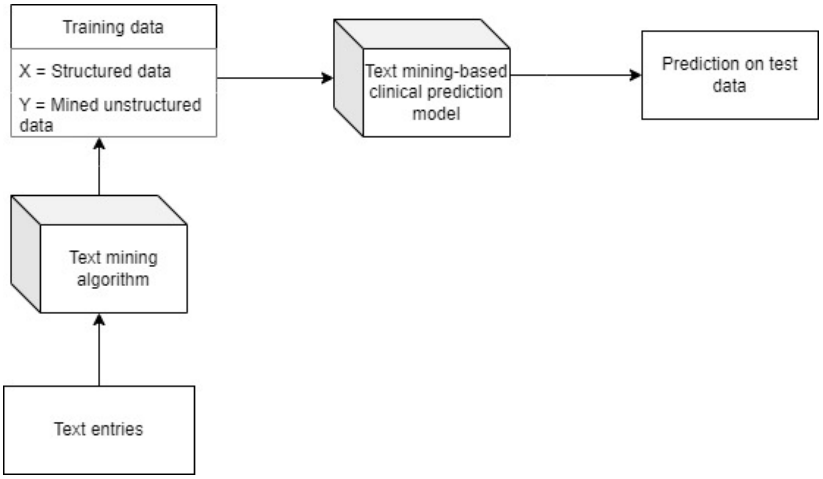


Figure 1.1: Diagram of one of the prediction models

1.3 RESEARCH QUESTIONS

The research questions are indicated below. RQ1 is the primary research question. To comprehensively answer the primary research question, we need the answers to secondary research questions RQa, RQb and RQc.

1. When using a text mining model for extracting clinical prognosis outcome data, how does the performance of the text mining model (precision, recall, F1 score) affect the performance of the prediction models (discrimination and calibration) trained on that mined data?
    a) How does changing the decision threshold and the training data size of the text mining algorithm affect the performance metrics (precision, recall, F1 score) of that text mining model for extracting information from clinical notes?
    b) How does changing the decision threshold of the text mining model affect the performance metrics (discrimination, calibration) of a clinical prediction model trained on text-mined data?
    c) To which extent is the choice of prediction modeling method (logistic regression, feedforward neural network) relevant in the performance (discrimination, calibration) of the prediction model trained on text-mined data?

1.4 THESIS OUTLINE

In this thesis we will first discuss the existing literature in Section 2, and how our research relates to it. In Section 3 we will discuss our method of answering the research questions posed in section 1. Section 4 will show the results of the experiments done, and Section 5 consists of the interpretation of the results and what they could mean for the clinical field, as well as a discussion of limitations and future research topics.

2

LITERATURE REVIEW

In this section, we look at some of the existing research regarding our research question. We identify how medical research is organized, give a brief summary of prediction modeling methods that will be relevant to this research and how they are used in clinical settings and discuss different performance measures that are often compared in the clinical field. We also give a short history of clinical text mining usage and provide some scientific embedding regarding the methods we used in our research. We also provide some examples to signify why our research is relevant and needed.

2.1 MEDICAL RESEARCH

Medical research is a huge field with many different subcategories. There is a distinction between primary and secondary research, where secondary research consists of aggregating the best sources to strengthen or weaken a certain conclusion. Aggregating sources can also increase the statistical analysis power for a conclusion. Primary research includes observational as well as experimental studies. In experimental studies, an intervention is done, e.g. a drug is administered or treatment is performed. The purpose of these studies is to determine the (side-)effects of certain drugs, surgery or other procedures. In observational studies, data is recorded from patients and the impact of a dependent variable is measured. Observational studies can be retrospective (historical) or prospective. In a retrospective study, previously recorded data is analyzed for certain patterns or outcomes, whereas a prospective study records the variables that are needed to draw a conclusion in the future. In other words, a prospective study generates new data whereas a retrospective study uses pre-existing data. An example of an observational study is a cohort study, where a group of people is followed during a study. Sometimes two groups of people are followed based on a common characteristic, after which one of the groups will be exposed to a certain risk factor to determine or explore the effect of this risk factor on a certain outcome [8]. In this project, the main focus is on the methods in observational research, specifically within epidemiology and prospective medical prediction research.

2.2 CLINICAL PREDICTION MODELS

Since 1950, physician's technology usage and dependency in the medical field have evolved, leading to more and more data being captured and stored about patients. In

the late 1950's the development of the HELP [9] system started, which is one of the earliest systems that aimed to help physicians in medical decision making and diagnosis. The development of this system continued for over 15 years. In 1974, the term *medical informatics* was coined, which is the field of studying information technology in the healthcare industry [10]. Since then, government mandates as well as the necessity to collaborate between different medical institutions have increased the data stored by those institutions for cooperation purposes [11]. Most of this data is stored in Electronic Medical Records (EMRs). The data is partly stored in a neat, ordered structure, which could directly be used to develop statistical models. These statistical models can then be used to help practitioners make informed decisions, or they can be used to predict, for example, a patient's mortality risk. Those prediction models can also be used to identify relevant predictor variables for a patient's mortality risk. For example, in research done by Valgimigli *et al.* [12] it was found that activation of a certain receptor correlates with an increased risk of mortality in patients with acute myocardial infarction. Previously, the risk of heart failure mortality would be estimated by the average mortality rates across all patients, or by the physician's experience. However, with access to the stored data from the EMRs and increased computational power, a more accurate estimate can be given to assist in decision making regarding whether or not a patient should be discharged [13], whether they should be admitted to an ICU [14], or whether they require end-of-life care (for high-risk patients) [15].

When developing medical prediction models, some things can happen that make the prediction model unsuitable for widespread clinical adoption. First of all, people often confuse predictions with causality. For example, a rule-based model for prediction of pneumonia risk was developed by Caruana *et al.* [16], using machine learning methods. One of the rules that the model learned was that a patient with asthma has a lower mortality risk. This is unexpected from a causal viewpoint, since asthma patients often require extra care when dealing with pneumonia. However, since this is known by physicians, many of the patients with asthma who presented with pneumonia were directly transferred to the ICU. The care they received there was so effective that their risk of dying from pneumonia was lower than that of an average pneumonia patient [17]. So while the rule asthma $\rightarrow$ low mortality risk was true in the data set, it does not represent a causal rule which, if misinterpreted, might lead to physicians changing their treatment based on the model, which in turn would lead to an increase in asthmatic patients dying. It is important to realize that the prediction of the model assumes that the received care is unchanged. It should not be used to derive rules about when to reduce the treatment that a patient should receive. Secondly, data like cardiac stroke volume or blood pressure is not directly measurable. Instead, indirect measures are used and it is hard to determine how well these are calibrated [18], which introduces a form of measurement error. This measurement error might make it difficult to transfer a prediction model from one institution to another, since different ways of measuring might have been used.

Collins *et al.* [19] noted that the reporting of prediction model studies is poor, and that this makes it hard to assess whether those models are biased or useful for clinical adoption.

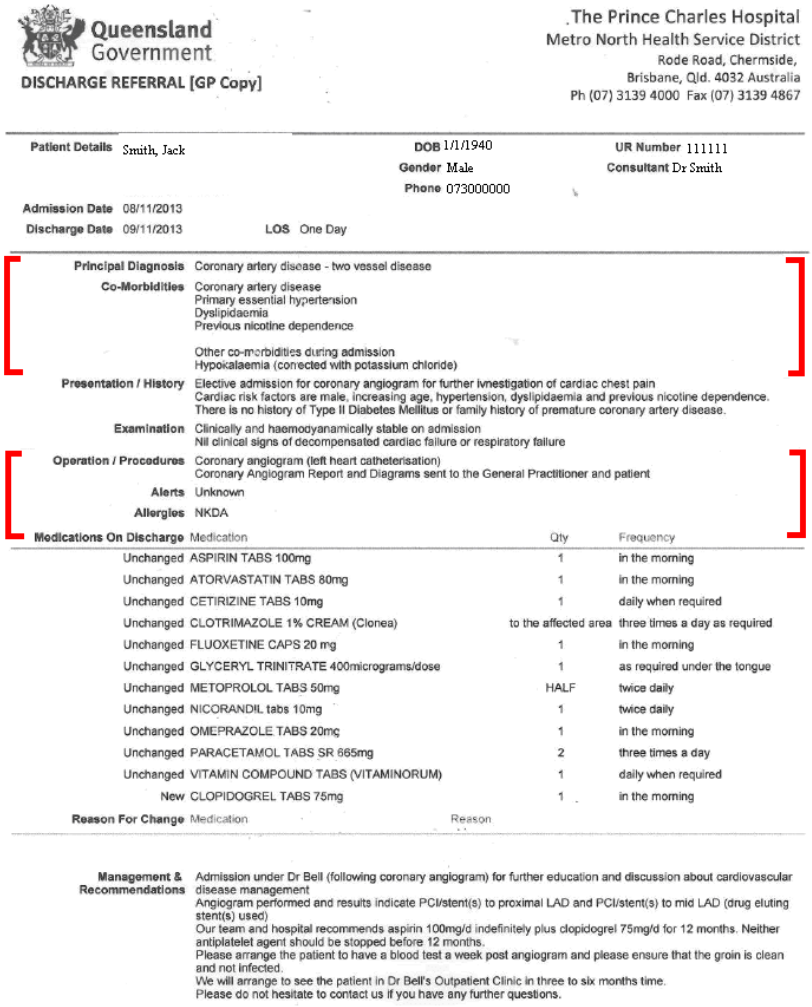
Based on this observation they created the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) initiative, which made a statement consisting of a 22 items checklist that should be included when reporting a multivariable prediction model for diagnosis or prognosis. These items include guidelines on model development, model performance and model specification, as well as how to report on the methods used, what to include in the discussion of a paper, and what should be in the introduction. In a review by Andaur Navarro *et al.* [20] from 2022, it was found that a mean of 37.9% of the TRIPOD items was adhered to in 152 papers published between 1 January 2018 and 31 December 2019.

With that, the need for a standardized way to develop prediction models (also when accompanied by text mining models) has been illustrated. This is where our research becomes relevant. If we can gain insight into the relation between text mining performance and text-mining based prediction model performance, developers can use this information to create better performing, more transparent models, and physicians may estimate more accurately whether a model might be useful in practice.

2.3 TEXT MINING IN CLINICAL RECORDS

Unstructured, free text entries are a less explored part of EMR analysis. These entries contain text that is not ordered as nicely as entries such as blood values or glucose levels. However, they still contain valuable information that might not be included in quantifiable entries. An example of an unstructured text entry in EMRs is discharge summaries. Discharge summaries often include a lot of information about the hospital stay, such as discharge diagnosis, treatment received in hospital, results of investigations and the follow-up required [21].

Text mining in these free text entries is challenging. Lots of abbreviations are used, and many of those abbreviations are not homogeneously used by different medical instances. Clinical narratives also often contain values for medical measures, such as blood pressure [22]. Terminology between two hospitals might vary, despite the same thing being meant. So not only is the lexicon very specific to the medical field, within the medical field there are also big differences in how each of the medical instances uses certain words. The notes can be, depending on the institution, grammatically coherent or only consist of keywords, meaning there is no grammatical structure at all, increasing the difficulty of text mining. The fact that there is a lot of jargon also makes it likely that some text mining algorithms will not perform as well, since there might not be a pre-trained representation for unknown words in algorithms that use contextualized embeddings like BERT [23]. An example of a discharge summary can be found in Figure 2.1.



Queensland Government
DISCHARGE REFERRAL [GP Copy]

The Prince Charles Hospital
Metro North Health Service District
Rode Road, Chermside,
Brisbane, Qld. 4032 Australia
Ph (07) 3139 4000 Fax (07) 3139 4867

Patient Details Smith, Jack — **DOB** 1/1/1940 — **UR Number** 111111
Gender Male — **Consultant** Dr Smith
Phone 073000000
Admission Date 08/11/2013
Discharge Date 09/11/2013 — **LOS** One Day

Principal Diagnosis Coronary artery disease - two vessel disease
Co-Morbidities Coronary artery disease
Primary essential hypertension
Dyslipidaemia
Previous nicotine dependence

Other co-morbidities during admission
Hypokalaemia (corrected with potassium chloride)
Presentation / History Elective admission for coronary angiogram for further investigation of cardiac chest pain
Cardiac risk factors are male, increasing age, hypertension, dyslipidaemia and previous nicotine dependence.
There is no history of Type II Diabetes Mellitus or family history of premature coronary artery disease.
Examination Clinically and haemodyanamically stable on admission
Nil clinical signs of decompensated cardiac failure or respiratory failure
Operation / Procedures Coronary angiogram (left heart catheterisation)
Coronary Angiogram Report and Diagrams sent to the General Practitioner and patient
Alerts Unknown
Allergies NKDA

Medications On Discharge	Medication	Qty	Frequency
Unchanged	ASPIRIN TABS 100mg	1	in the morning
Unchanged	ATORVASTATIN TABS 80mg	1	in the morning
Unchanged	CETIRIZINE TABS 10mg	1	daily when required
Unchanged	CLOTRIMAZOLE 1% CREAM (Clonea)	to the affected area	three times a day as required
Unchanged	FLUOXETINE CAPS 20 mg	1	in the morning
Unchanged	GLYCERYL TRINITRATE 400micrograms/dose	1	as required under the tongue
Unchanged	METOPROLOL TABS 50mg	HALF	twice daily
Unchanged	NICORANDIL tabs 10mg	1	twice daily
Unchanged	OMEPRAZOLE TABS 20mg	1	in the morning
Unchanged	PARACETAMOL TABS SR 665mg	2	three times a day
Unchanged	VITAMIN COMPOUND TABS (VITAMINORUM)	1	daily when required
New	CLOPIDOGREL TABS 75mg	1	in the morning

Reason For Change	Medication	Reason

Management & Recommendations Admission under Dr Bell (following coronary angiogram) for further education and discussion about cardiovascular disease management
Angiogram performed and results indicate PCI/stent(s) to proximal LAD and PCI/stent(s) to mid LAD (drug eluting stent(s) used)
Our team and hospital recommends aspirin 100mg/d indefinitely plus clopidogrel 75mg/d for 12 months. Neither antiplatelet agent should be stopped before 12 months.
Please arrange the patient to have a blood test a week post angiogram and please ensure that the groin is clean and not infected.
We will arrange to see the patient in Dr Bell's Outpatient Clinic in three to six months time.
Please do not hesitate to contact us if you have any further questions.

Figure 2.1: Example of a discharge summary, created by Maurice *et al.* [24]

Solutions have been created to standardize the text mining process for clinical research. In 2010, Savova *et al.* [25] built and evaluated the clinical Text Analysis and Knowledge Extraction System (cTAKES). This is a natural language processing system designed for the clinical field. It contains a sentence boundary detector, tokenizer, normalizer, Part-of-Speech tagger, a named entity recognition annotator, and a negation annotator. This system has been used to develop prediction models or information extraction [26, 27], and is still considered a state-of-the-art mechanism to make the development of prediction models using NLP easier. Similarly, Aronson and Lang [28] created a system to extract topics from biomedical texts. The two systems have different goals but have a slight overlap in that they can both analyze clinical notes or papers to extract biomedical knowledge [29]. And while a lot of research has been done on the performance of these systems by themselves, not much is known about their interaction with prediction models.

There are multiple ways to use text mining in clinical predictions. The first method is to mine a certain variable from the text and use this variable in a prediction model. In this case, the mined variable is used as input for the machine learning model. One such instance is the research done by Ford *et al.* [30]. In a meta-analysis, they found that including textual information in case-detection algorithms led to a significant increase in im-

provement in algorithm discrimination when combined with other (structured) predictors. They noted that there was no clear difference between rule-based and machine learning approaches to extracting information from the text.

Another way is to use text mining to fill in missing data. Missing data can be addressed by ignoring the samples containing missing data (not preferable), using averages, fitting a regression function to fill in the missing value or treating the missing data as a separate class so that the samples can still be used [31, 32]. Not every solution might work for every type of "missingness". Values might be missing completely at random or for a specific subgroup of samples. This requires appropriate handling, to reduce the risk of developing an over- or underestimating model [33]. This is quite a complex issue, and using text mining to fill in the gaps may help prevent it. For example, Erraguntla *et al.* [34] managed to fill in a gap of missing ICD 9 codes using text mining with an accuracy of around 70%. One limitation of this research is that the performance was validated using records for which the ICD code was known. There might be a significant difference between samples where the ICD code is known and samples where the code is missing. Two other examples are research done by Hylan *et al.* [35] and Dormosh *et al.* [7]. They use natural language processing to mine a specific outcome from clinical notes. They both use the same methodology; first, the text entries are mined, after which the positive outcomes are checked manually. A flaw in this methodology is that it will be unknown which entries might be missed by the text mining algorithm since negative outcomes are not analyzed. This might lead to structural flaws in the prediction models, since the characteristics of the falsely negative flagged samples are unknown, and might share a pattern. These two examples contribute to the motivation of our research; the reporting of their text mining method is incomplete which makes it unclear whether this model is suitable for practical use.

2.4 PREDICTION MODELING METHODS OVERVIEW

2.4.1 *Risk scores*

Risk scores are a traditional way to model the risk of a patient. Originating from the field of statistics, they are calculated to assess, for example, the risk of a patient for complications or death during surgery or an ICU stay [36]. Generally, they consist of some calculations using various relevant factors and produce a score that gives insight into the risk that is involved in their stay. These scores can then be used to adjust the care given to the patient. If a patient has a high-risk score, they might need some special medical attention to reduce the risk of complications or death. One of the earlier risk scores is the widely used Simplified Acute Physiology Score (SAPS), which was published in 1984 [37]. It uses 14 easily measured variables and is designed to be easily used across multiple pathologies, reducing the need for pathology-specific risk scores. As time progressed, the complexity of risk scores went up in an effort to score the highest accuracy, such as APACHE IV [38], which consists of over 120 predictor variables. Some other examples of risk scores include the ASA grade for postoperative mortality, APACHE I, II and III (Acute Physiology And

Chronic Health Evaluation) and POSSUM (Physiological and Operative Severity Score for the enUmeration of Mortality and morbidity). These risk scores precede contemporary machine learning methods and are generally less accurate than new state-of-the-art machine learning techniques [39–41]. However, they have been validated thoroughly and are easy to understand and use, given that they typically use fewer predictor variables than machine learning methods. For this reason, they still see quite a bit of use even today. Oftentimes, they are also calculated and used as a predictor value in machine learning-based prediction models.

2.4.2 *Prediction model performance measures and evaluation*

Performance measures for prediction modeling vary depending on the task. For regression tasks, the mean square error (MSE), root mean square error (RMSE) or mean absolute error (MAE) are common. For classification problems, we often look at accuracy, F1 scores, confusion matrices or receiver operating characteristic (ROC) curves (seen in Figure 2.2). In clinical settings, the accuracy metric is often not very insightful, since classes may be imbalanced (patients with a certain prognosis are often outweighed by those without). If 1% of the data has label 1, the accuracy can be 0.99 just by predicting the majority class. It is also important to note that recall (what percentage of the cases is detected?) is often more important than precision (what percentage of the detections are actual cases?). One reason for this is that algorithms are often used as a pre-screening method, so positive classifications will still be manually confirmed by a physician, whereas negative classifications might not be reassessed. Since the predictions are uncertain, a binary prediction gives less information than a risk score. It is hard to distinguish a label 0 with extremely low risk from one that borders the 1 label. ROC curves are analyzed frequently, to evaluate the trade-off between true positive and false positive rates. The more top-left the ROC curve is, the better the model's performance. To quantify a model's performance according to the ROC curve, the area under the curve (AUC or AUROC) can be calculated to give some insight.

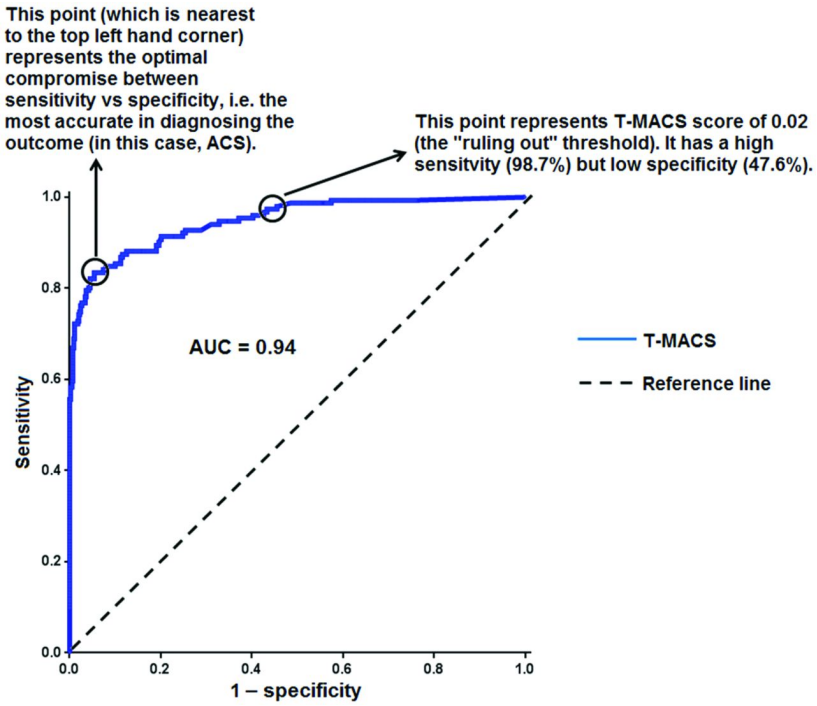
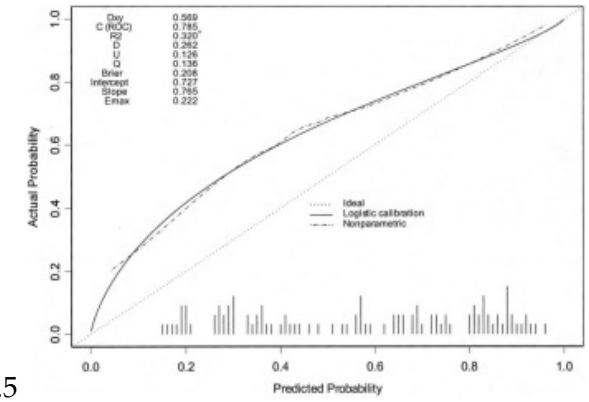


Figure 2.2: Example of a ROC curve, created by Hoo *et al.* [42], licensed by BMJ Publishing Group Ltd.

Steyerberg and Vergouwe [43] call for standardization within the development of clinical prediction models. Part of that standardization covers the evaluation of the model. They introduce the ABCD measures for validation, which include calibration (also argued for by Van Calster *et al.* [44]), discrimination and clinical usefulness. Since validation of clinical usefulness (D) is not in the scope of this project, it will not be summarized here. Instead, we focus on the ABC part. First of all is Alpha: calibration-in-the-large. This metric refers to the ratio of observed outcomes and predicted outcomes. The example Steyerberg and Vergouwe [43] use is: "if we predict a 5% risk that a patient will die within 30 days, the observed proportion should be ~5 deaths per 100 with such a prediction." Beta is the calibration slope. This slope determines how "extreme" predictions were. If low predictions are consistently too low, and high predictions are consistently too high, the calibration-in-the-large could appear to be perfect, but the predictions would still be consistently wrong. The calibration slope captures this notion. The closer to 1, the better. The calibration curve can be plotted to show how well-calibrated a model is. An example can be seen in Figure 2.3. It shows the predicted probability on the x-axis against the observed probability (% of positive outcomes for those samples) on the y-axis. A perfectly calibrated model has a 45-degree line from the origin of the graph. If the calibration curve is consistently above that, it means the prediction model underestimates risks, whereas a line below indicates an overestimating model. The final letter in our ABC model is for the concordance statistic: discrimination. This notion captures the idea that a model should be able to separate patients with positive outcomes from patients with negative outcomes. If a model predicts the average outcome for every patient, it would have a perfect calibration, but no discriminative ability, hence being useless. For binary prediction, this concordance statistic c is

equal to the area under the ROC curve. In this project, we will use these guidelines to compare the performance of the models we will create.



0.5

Figure 2.3: Example of a calibration curve, created by Herder *et al.* [45], licensed by Elsevier.

2.4.3 *Refitting to improve calibration*

In medical prediction models it has been observed that throughout the years, the calibration drifts despite the discrimination capability being maintained. Models start consistently over- or underpredicting after a few years, possibly due to a shift in prevalence of a certain outcome [46]. This can be addressed by refitting the model to increase calibration. First, a logistic regression model without regularization should be fit to the predicted linear probabilities for the validation set. To do this, the logit function ($log(\frac{p}{1-p})$) should be applied to the outcome vector as created by the prediction model, which inverses the logistic function applied by the logistic regression model. This new model consists of only an intercept and one coefficient [47]. Now let us call our original prediction model LR_1, and our newly fit model with only one coefficient LR_2. Then $LR_2 = \alpha + \beta LR_2$, where α is the intercept and β the slope. We can update our calibration intercept by replacing the original intercept of LR_1 with α added to the original intercept of model LR_1. The slope can be updated by multiplying each of the coefficients of LR_1 with the single coefficient of LR_2 ([44], supplementary information). Of course, the more refitting is being done, the more the model is fit to the validation data instead of the training data. This can lead to worse performance on the test set if the validation data set is too small. If we know the training data is biased, however (such as in our case where we text mine outcomes from clinical notes), we can use a validation set to recalibrate the model after training it. This might reduce the costs and effort needed compared to creating a fully unbiased training data set. This process of recalibrating can also be done on external datasets. This means that a model created at a certain medical institution can be taken and recalibrated to a different health care facility [48]. Other recalibration methods include Platt scaling [49] and isotonic regression [50], but these strategies are generally reserved for boosted trees, random forests and support vector machines [51].

2.5 TEXT MINING METHODS OVERVIEW

For our text mining algorithms, in this project, we primarily focused on tf-idf. We did not look at rule-based data mining approaches, because despite these being commonly used, they require expert knowledge in the domain, which we don't possess. Other algorithms we considered were Word2Vec, GloVe and Clinical BERT. Since we realized during our project that tf-idf allowed for a wide range in performance, we did not analyze other text mining algorithms.

For the purpose of this research, we will define a text mining algorithm as a model that extracts an outcome from text. We use the text mining algorithm on a clinical note, and the algorithm will output whether that note contains the output we inspect, e.g. mortality. This means that for tf-idf we will add our own logistic regression model to make predictions with the generated tf-idf vectors. More on this can be read in Section 3.

2.5.1 *tf-idf*

tf-idf (term frequency-inverse document frequency) is an algorithm that calculates how important a word is for that text based on how often it occurs, normalized for how often that word occurs in all texts. For example, the word "the" occurs often in most texts, so it gets a low score. "mortality" might occur often in texts for patients that expired, but not in texts for patients that survived, so it gets a high rating. To use this value in text mining, the top n most "important" words are encoded into a vector, and each sample document will have n features, with each feature having a count of that word for that document. The tf-idf value for a term within a document is calculated with the formula $tf(t,d) * idf(t,D)$, where $tf(t,d)$ is the count of term t divided by the number of unique terms in that document, and $idf(t,D)$ is calculated by taking the log of the result of the number of total documents divided by the number of documents that contain term t. tf-idf has been used to improve the accuracy of clinical prediction models, by combining traditional regression models using structured data with a tf-idf text mining approach. Research has shown that adding tf-idf (or other context-free algorithms like Bag-of-Words) as a way of extracting information from clinical notes from a clinical dataset to a regression model with other features increases the AUROC on a test split when compared to a baseline model using only structured data for predicting 1-month mortality [52]. Note that this does not give us information about the performance outside of the dataset, such as in other clinical facilities. Similarly, Klang *et al.* [53] came to the same conclusion with a significantly increased AUROC, but for 48-hour and in-hospital mortality. These studies lead us to believe that mortality can be sufficiently mined using tf-idf.

2.5.2 *Clinical BERT*

BERT (Bidirectional Encoder Representations from Transformers) [54] is an algorithm that could be used for text mining in our context. It is a language model built on the idea

of Transformers [55], which is a way to use an attention mechanism without relying on recurrence. This means that the model does not have to go left to right, but instead can go both ways at the same time. This allows for the model to have a higher contextual awareness. An attention mechanism relates every token in the input sequence to every other token. BERT is trained on general text like Wikipedia articles, but can be fine-tuned for specific tasks or inputs, by using the base model and training it on a specific dataset. This is called transfer learning. With a fine-tuned model, performance can be increased on clinical data. As with tf-idf, BERT and several of its other pre-trained variants like Clinical BERT, MedBERT and BioBERT have been widely used for clinical prediction since their introduction [56–58]. Since Clinical BERT has been trained on MIMIC-III [59], the data set we will be using, this model could also be used for this research. When using Clinical BERT, it should be kept in mind that there is a 512 token limit. In the research by Huang *et al.* [56], they dealt with this problem by splitting the clinical notes from a specific patient into sub notes of ~318 words and then filling in a formula to "average" the prediction in a way.

Using this formula resulted in an increased performance of 3-8% when compared to taking the mean of the predictions for readmission prediction. We assume mortality prediction is similar, so the same method of calculating the outcome variable probability could be used in a project like this as well.

Recently, Carlini *et al.* [60] showed that it is possible to extract training data from language models like GPT-2 [61]. This caused concern for other language models, such as BERT, and of course ClinicalBERT. For this reason, we might ask ourselves whether there is a privacy risk in releasing models that are trained on privacy-sensitive data. Lehman *et al.* [62] tried to replicate this conclusion using a BERT model pretrained on the MIMIC-III dataset with artificially reintroduced patient names, but the conclusion stated that they were mostly unable to do so using simple methods. However, they noted that this does not rule out the possibility of extracting patient data with a more advanced technique, or future techniques that might not yet be developed. For this reason, we believe that model weights trained on the MIMIC-III data set should not be shared.

2.5.3 *Text mining performance measures and evaluation*

Since the original data contains labels of 0 or 1 for in-hospital mortality, we want to mimic this for our text-mined dataset, so we will be classifying our samples with a 0 or 1. This means that we will use binary classification metrics. The specific classification metrics we will use are precision, recall and F1 score. Precision indicates what percentage of the detections are actual cases and is calculated with $\frac{TP}{TP+FP}$, where TP is the number of True Positives, and FP is the number of false positives. Recall indicates what percentage of the cases is detected and is calculated with $\frac{TP}{TP+FN}$. FN is the number of False Negatives. The F1 score is calculated with $2 \cdot \frac{Precision*Recall}{Precision+Recall}$. This score is calculated by taking the harmonic mean of the precision and recall, meaning that a model with a low precision or recall will have a low F1 score, whereas a balanced precision and recall will lead to a higher F1 score.

2.6 USED PREDICTION MODELING METHODS

In this project, we will be taking a look at two different prediction modeling methods, logistic regression (LR) and feedforward neural networks (FFNN). We do this to determine whether our results are generalizable across different prediction modeling methods. If the trends that show for logistic regression carry over to the feedforward neural network, that gives us an indication that the results are at least generalizable over these two algorithms. While we often see that neural networks outperform logistic regression models in traditional AI tasks, research has shown that neither of them necessarily outperforms the other for clinical prediction tasks [63–66], which could indicate that clinical prediction tasks generally do not include non-linear predictors.

2.6.1 *Logistic regression*

The first prediction modeling method, which can perhaps be considered the main modeling method for this research due to its widespread usage, is logistic regression. It is one of the most popular techniques across medicine, marketing, credit scoring, and public health [67]. The basic idea is similar to that of a linear classifier, where a vector of weights is multiplied by a vector of inputs, giving the formula $\mathbf{w}^T\mathbf{x}_j$. To introduce an intercept to this formula, $\mathbf{x}_{(j,0)} = 1$. With gradient descent, the optimal vector for $\mathbf{w}$ is found by updating the weights iteratively until a minimum in the loss function is found. logistic regression applies the logistic function to the linear regression function, which leads to outcomes in the space of [0, 1]. This makes the outcome analogous to a probability estimate. The loss function is changed to the maximum likelihood estimation, so our model is chosen such that the training data is most probable given that model compared to other models, or in other words, the model that explains the data best [68]. A logistic regression model often includes a regularization term that prevents or reduces overfitting by penalizing complex models.

2.6.2 *Feedforward neural networks*

A feedforward neural network is a network consisting of layers of nodes (some hidden, meaning that they are not connected to the output nodes) that are connected, moving in one direction so that each node is passed no more than once. Each node has a weight and an activation function which determines whether and how the signal is propagated throughout the network. An activation function that results in either 0 or 1 is called a perceptron, but other activation functions, such as the logistic function can also be implemented. Having multiple of these nodes with nonlinear activation functions means that the entire network can reshape the inputs into a nonlinear output, which makes it able to approximate the training data more closely. As a result, training an FFNN is more complex and takes more resources, and has a higher risk of overfitting if no preventative measures are taken.

2.7 COMMON OUTCOME VARIABLES IN CLINICAL PREDICTION MODELING

When regarding prediction models in clinical settings, there is a distinction to be made between diagnosis and prognosis [69]. Diagnosis is the identification of a certain present disease. For example, detecting whether someone has Alzheimer's disease based on an MR image of the hippocampus using machine learning methods [70]. This is a form of diagnosis. On the other hand, we have prognosis. Prognosis refers to estimating the risk of future outcomes. Some prognosis outcomes that can be analyzed include mortality rate within a given time (mortality within 30 days or within a year are commonly analyzed outcomes). Other prognosis examples include, given a cancer patient, whether or not cancer will reoccur, or whether or not the patient will survive the cancer [71]. Both diagnostic and prognostic variables can be used as outcome variables, depending on the goal.

In this project, we will focus on in-hospital mortality. Other outcome variables we considered were sepsis (an extreme response to an infection), 30-day and 1-year mortality.

2.7.1 *In-hospital mortality*

One of the outcome variables commonly predicted is in-hospital mortality. This outcome is included in the MIMIC-III dataset and we expect it to be relatively easy to mine from the text entries since it should almost always be included in the discharge summary and mortality is expected to be written in relatively context-free ways. We expect the problem to mainly be a word-detection problem (e.g. "Absence of followup" or "Discharge condition: Expired" in the discharge summary indicates that the person has passed away), where the context is not as important as in other text mining tasks. Even negations should be rare in discharge summaries when talking about the expiration of patients. For this reason, it is expected to be easily minable, even when supported by a rudimentary text mining algorithm such as tf-idf. We expect tf-idf to perform similarly on outcome variables like this when compared to a context-dependant algorithm like BERT or Clinical BERT.

2.7.2 *Sepsis*

A second outcome variable that could be predicted is sepsis. Because of its multi-factorial characteristic, it is considered quite challenging to predict. Text mining this outcome from text should once again mainly be a text-mining task. We expect it to be harder than in-hospital mortality detection since mortality is virtually always mentioned in the discharge summary whereas sepsis is less impactful on the continuation of treatment, so it might not always be mentioned. For this outcome variable, we expect that it requires a bit more nuance, so a context-dependant algorithm like Clinical BERT might outperform tf-idf.

3

METHODS

In this section, we look at the structuring of the research. We discuss the dataset itself (and preprocessing steps), the models that were built, the text mining and machine learning algorithms involved, how we artificially changed the performance of the text mining algorithms, and how we evaluated the differences between the models in relation to the text mining model's accuracy. The code used in our method is published, see Chapter A.

3.1 STUDY DESIGN

An overview of the project setup can be seen in Figure 3.1. The purple circle indicates how we answer RQa, the yellow circle indicates how we answer RQc, and the red circle indicates how we answer the main research question RQb and finally RQ1. The red rectangle indicates the reference pipeline, the green rectangle indicates the pipeline for our own text mining-based clinical prediction models. It is important to note that the green pipeline was traversed for each of the combinations of split size and decision threshold mentioned above.

To elaborate, our general setup can be split up into three sections:

1. Text mining part: We mined in-hospital 48-hour mortality from clinical notes for each patient. We adjusted the performance of the text mining model by adjusting the **decision threshold** and the training data **split size**.
2. Prediction model part: We trained a prediction model with the text mined outcomes.
3. Evaluation: The effect of changing the text mining performance on the performance of the prediction model was evaluated.

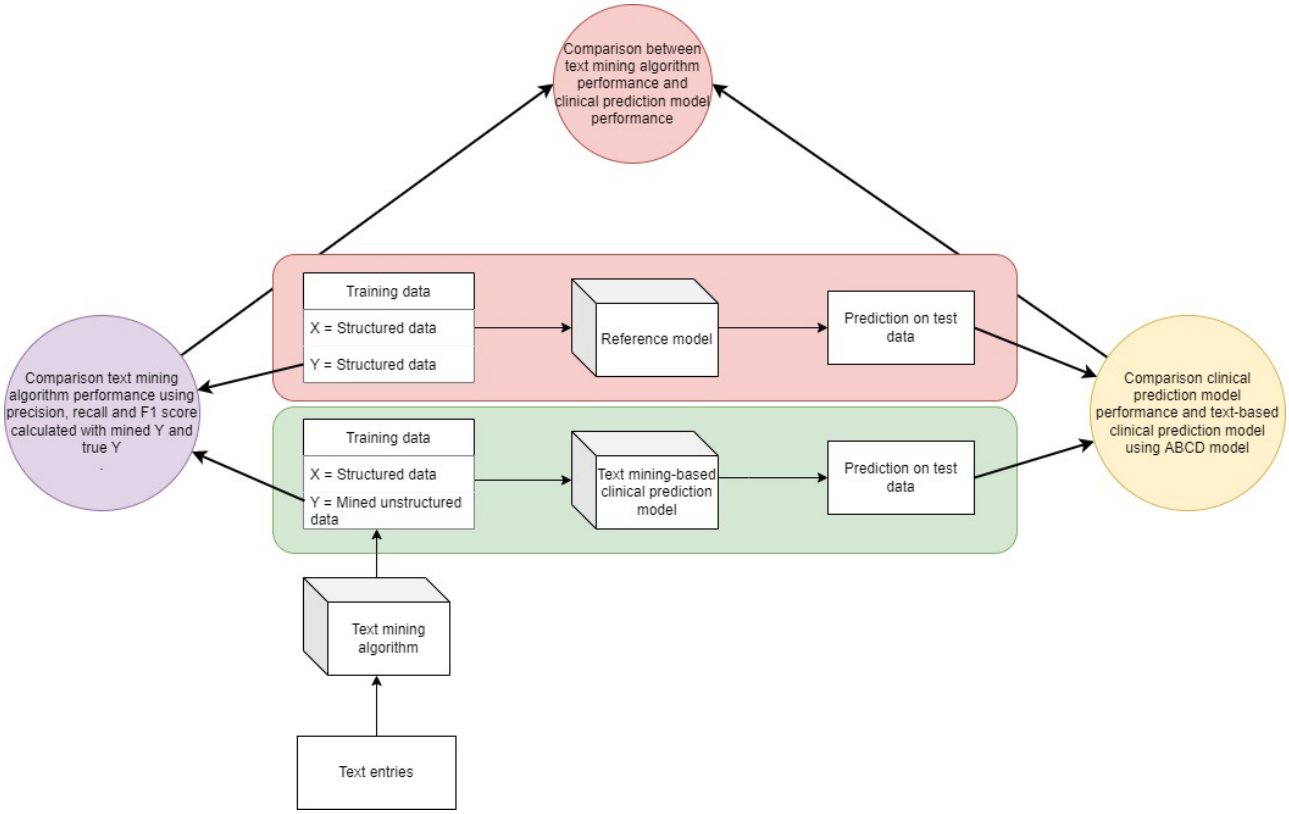


Figure 3.1: Diagram of project setup

Throughout this thesis, we use some consistent terminology, which will be explained below.

- *Text mining model*: A model that takes text as input and outputs whether or not that piece of text contains the outcome variable.
- *Clinical prediction model*: A model that takes certain inputs, and predicts an outcome about the health status of a patient.
- *Reference model*: A clinical prediction model (FFNN or logistic regression) that is trained without any text mining involved, to compare with the results of our text-based clinical prediction model. It is based on the model created by Harutyunyan *et al.* [72].
- *Text mining-based clinical prediction model*: A clinical prediction model that is trained using text mined outcomes.
- *Split size*: The ratio of data given to the tf-idf algorithm to learn from. E.g. a split size of 0.2 indicates that 20% of the total data made available for the tf-idf model is used.
- *Decision threshold*: The cut-off value for the text mining model. A decision threshold of 0.1 means that all predicted probabilities of label 1 above 0.1 will be classified as a 1. With this parameter we can increase or decrease the recall; a decision threshold of 0 means that all samples are classified as label 0 and a decision threshold of 1

means that all samples are classified as label 1. Any value between 0 and 1 slides the classification ratio.

3.2 DATA

The dataset we used is the MIMIC-III dataset [59]. This dataset contains information about ICU stays of patients at the Beth Israel Deaconess Medical Center in Boston, Massachusetts between 2001 and 2012. It includes vital signs, medications, laboratory measurements, observations and notes charted by care providers, diagnostic codes, whether a patient survived or not and the length of their stay, among others. Since the data of patients is private, all the data in MIMIC-III is de-identified, and the dates are also shifted forward or backwards in time to further anonymize the patients. It is one of the freely accessible databases, normally clinical dossiers are not freely distributable. A data usage agreement was signed to access the data for this study.

Each patient in the dataset has one or more hospital admissions, and each hospital admission contains one or more ICU stays. A clinical event is an observation, measurement or treatment of a subject. The dataset contains over 300 million events we can use to predict in-hospital mortality.

The dataset contains 53,423 ICU admissions for 38,597 unique patients. The in-hospital mortality is 11.5%. The main tables we are interested in are the "noteevents", which contains 2,083,180 rows of physicians' notes, and "chartevents" and "labevents", which contain all charted data for all patients. This includes things like heart rate, glucose levels, height, oxygen saturation, capillary refill rate and diastolic blood pressure. The table of features used can be seen in table 3.1. From these features, the minimum, maximum, mean, standard deviation and skew are recorded for the first 10%, 25% and 50% as well as the last 10%, 25% and 50% of time. This means there are a total of $17 * 7 * 6 = 714$ features. The features are normalized and missing values are replaced with the mean value for that feature in the training set.

Variable	MIMIC-III table	Modeled as
Capillary refill rate	chartevents	categorical
Diastolic blood pressure	chartevents	continuous
Fraction inspired oxygen	chartevents	continuous
Glascow coma scale eye opening	chartevents	categorical
Glascow coma scale motor response	chartevents	categorical
Glascow coma scale total	chartevents	categorical
Glascow coma scale verbal response	chartevents	categorical
Glucose	chartevents, labevents	continuous
Heart Rate	chartevents	continuous
Height	chartevents	continuous
Mean blood pressure	chartevents	continuous
Oxygen saturation	chartevents, labevents	continuous
Respiratory rate	chartevents	continuous
Systolic blood pressure	chartevents	continuous
Temperature	chartevents	continuous
Weight	chartevents	continuous
pH	chartevents, labevents	continuous

Table 3.1: Table with features used in the reference model. Adapted from Harutyunyan *et al.* [72]. (CC BY 4.0)

From this dataset, we used ICU stays to develop our model. For in-hospital mortality, we determined for each of the hospital admissions whether the patient expired during the hospital stay, using data from the first 48 hours of the hospital stay.

3.2.1 *Data split*

The data split in this project is less straightforward than usual data splits, so a diagram was made, seen in Figure 3.2. The MIMIC-III database consists of 42057 usable ICU stays, that is excluding patients with multiple ICU stays and ICU transfers, of which we separated 42% into the text mining component (blue), and 25% into the prediction model component (green). The text mining component consists of a train and test set, of which the train set varies based on the split size. After training the text mining model and creating a new outcome vector for the prediction model's training data, some data were excluded as done by the authors of the adapted model, Harutyunyan *et al.* [72]. The most relevant exclusion criteria were:

- Length of stay was missing
- Length of stay was shorter than 48 hours

- No events were captured before 48 hours

These criteria removed 21137 samples. Note that our text mining model is trained on data that contains these samples. This should not be a problem, as long as mortality is recorded and text notes are registered, our text mining model will be able to learn from it. After training the text mining model we also removed samples from the prediction model training set that did not contain text notes, since those could not be classified by the text mining model. The prediction model test data set did not require text notes (since text mining is only done for training purposes, our test data should contain the ground truth), so they are kept as-is.

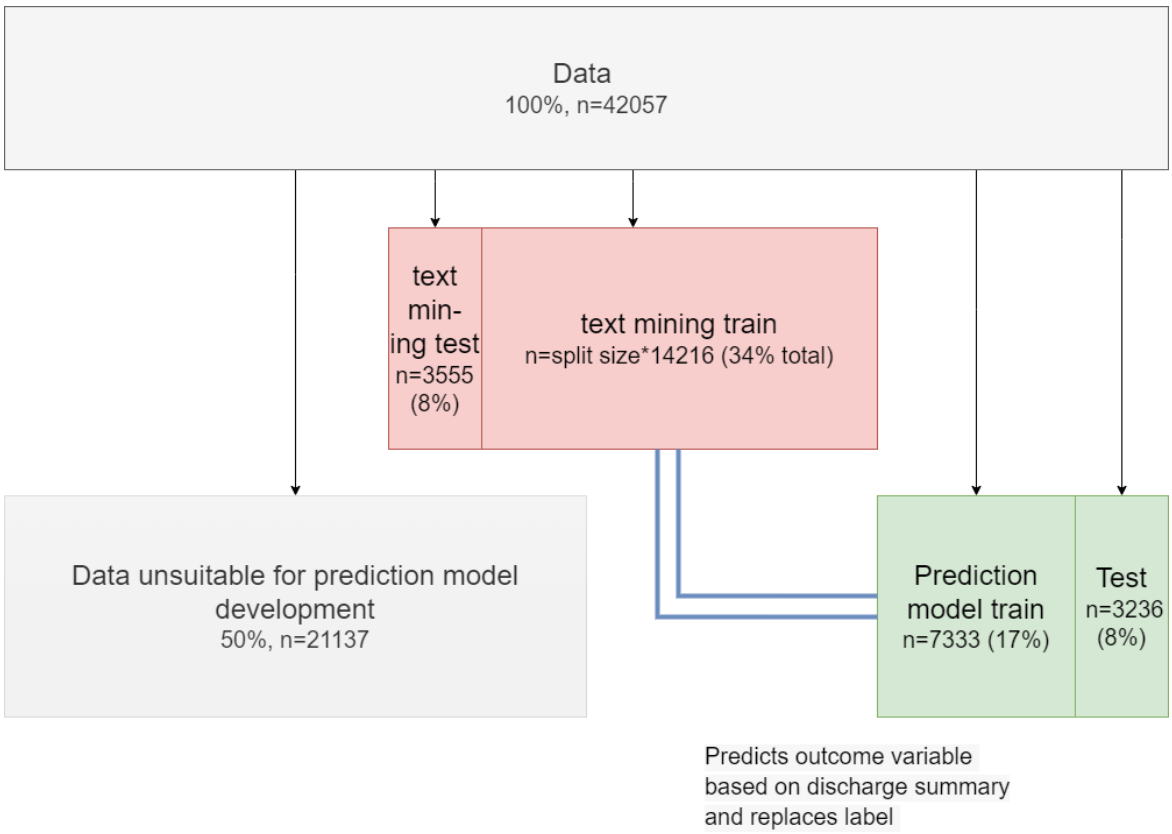


Figure 3.2: A figure showing how the data was split up and used.

3.3 REFERENCE MODEL

For our project we first constructed the reference models to compare our results with. For this reference model we trained a logistic regression model that predicts in-hospital mortality based on a number of structured features that can be extracted from the MIMIC-III dataset. This model used the ground truth labels; no text mining was involved. The model from Harutyunyan *et al.* [72] was used as our reference model. They proposed a strong baseline for prediction tasks, including mortality risk, using the features in table 3.1. We also created a second reference model, a feedforward neural network. It is the same as the first reference model, except using a different machine learning algorithm. We then split the data into a training, validation and test set, and did some minor optimizing.

Since this is a baseline model to compare our results to and the aim of this project is not to achieve the highest possible performance, we did not spend too much time on optimization. We determined the area under the ROC curve (AUROC), as well as gathered the calibration slope, intercept and calibration in the large for the reference models.

3.4 TEXT MINING MODEL DESIGN

Then, we took the same dataset and instead of using the outcome variables as provided by the dataset, we mined them from the aggregate of clinical notes per patient using tf-idf combined with a logistic regression network instead. The logistic regression used C = 1 and L2 regularization based on a grid search with parameters [0.001, 0.01, 0.1, 1, 10, 100, 1000]. The tf-idf vectorizer selects the 2000 most important words and converts each sample into a data set with 2000 features. The logistic regression model, with C=1 and L2 regularization, then determines for each sample whether the feature vector is associated with mortality. This is our text mining model. Some standard text preprocessing steps, such as stemming, lemmatization, number removal, lower casing, stop word removal and punctuation removal were performed to increase the performance of our tf-idf model. The two outcome variable vectors (original vector and the text mined vector) are compared using standard classification performance metrics, like F1 score, precision and recall. With performance metrics of these text mining algorithms we can answer RQa: "How does changing the decision threshold and the training data size of the text mining algorithm affect the performance metrics (precision, recall, F1 score) of that text mining algorithm for extracting information from clinical notes?"

3.5 PREDICTION MODEL DESIGN

This new outcome variable vector was then used to train two new clinical prediction models, with the same features as the reference model, resulting in a new text mining-based clinical prediction model that is expected to perform similarly to the original model. The two models created are a logistic regression model and an FFNN model. The reason we choose FFNNs and logistic regression is that they are some of the most used algorithms for models combining free text with structured data entries [6]. For logistic regression, since we have the reference model, we did not perform a grid search to optimize parameters, since we already had the optimized parameters as found by Harutyunyan *et al.* [72] (C=0.001, L2 regularization). For the FFNN, we manually optimized the reference model by testing different layer setups, and use the layer setup that we find works best for all of the text mining-based models as well. The final model passes the input of 714 features through a densely connected reLU layer with 16 units, then applies a dropout filter of 0.5, into another densely connected softmax layer with 2 units. We compare the performance measures we have recorded of both models, both discrimination (AUROC) and calibration (slope, intercept, CITL), to see if either of them outperforms the other.

3.6 PERFORMANCE ADJUSTMENTS OF THE TEXT MINING ALGORITHM

3.6.1 *Split size*

To answer our research questions we want to have a wide variety of text mining model performances. We want bad as well as good text mining performances, and see how it affects the prediction model. To achieve this, we need to adjust our text mining models so that the performance drops slightly or dramatically. The first way of changing the performance is by giving it less data to work with. Instead of using all samples, we only used a percentage of the data, which we expected would cause the precision and recall metrics to decrease. That decrease should also be observable in the text mining-based clinical prediction model's AUROC. We used split sizes in the range of 0.05 up to 0.95 with increments of 0.05.

3.6.2 *Decision threshold*

The second way to change the performance of the text mining algorithm is by changing the cut-off value for predictions to change the precision/recall ratio. This means that instead of any sample with a probability of over 0.5 getting a label 1 (as is the default), we changed it to some other value, such as 0.7 or 0.3. Increasing the decision threshold increases precision but reduces recall, while reducing the decision threshold reduces precision but increases recall. With these two adaptations, we generated a wide range of precision and recall values for our text mining algorithm. We used decision thresholds 0.1 up to 0.9 in increments of 0.1. The result of changing the split size and decision threshold is a big table mapping these values to each other, with for each combination the precision, recall, F1 score, AUROC, slope, intercept and CITL.

3.7 EVALUATION OF RESULTS

With the experiments done, we now have numerous data points that connect the split size and the decision threshold to the precision, recall and F1 score of the text mining algorithm. These in turn connect to the discrimination and calibration metrics of the text mining-based clinical prediction model. An example row with possible values can be seen in Table 3.2. The different F1 scores are from the different amounts of data used for training the text mining-based clinical prediction model and from adjusting the decision threshold, and the AUROC score is the result of using that text mining algorithm in conjunction with our text mining-based clinical prediction model. A data table consisting of 174 ($split\ sizes * decision\ thresholds = 18 * 9$) sample rows can be made for both the FFNN and the logistic regression models. With the information from this table, we can answer RQb.

Split size	Decision threshold	Precision	Recall	F1 Score	AUROC	Slope	Intercept	CITL
0.05-0.95	0.1-0.9	0-1	0-1	0-1	0.5-1	1	0	0

Table 3.2: A layout for what each row of our data looks like.

4

RESULTS

The results include plots about various performance metrics of the prediction models, and how they change given the adjustments in text mining performance.

To start this section off, we reestablish some terminology. When we talk about precision, recall or F1 scores, that means we are talking about the performance of the text mining algorithm. The text mining algorithm supplies us with 0 or 1 labels for the prediction model to learn from, so we can calculate the precision, recall and F1 score based on the true labels. When talking about AUROC or calibration metrics, this concerns the prediction model. The prediction model calculates a probability representing the risk of mortality. For terms like decision threshold, split size and reference model, refer to Section 3. The colored plots will always be color coded in a way that green equates to the highest possible outcome.

All plots will be added in full size to the appendix.[1]

4.1 TEXT MINING PERFORMANCE VARIATIONS

The first section of our results will be about the text mining algorithm, and how we adjusted its performance by changing the split size and the decision threshold. In Figure 4.1 we can see the effect of changing the training data split size on the F1 score of the positive label. Increasing the training data size increases the F1 score. The relation seems asymptotic. The first 20% of training data accounts for 60% of the increase in F1 score, after which the remaining 80% of data accounts for the last 40% of the increase in F1 score. Figure 4.2 shows the effect of a change in decision threshold on the performance metrics of the positive label. Interestingly, the optimal decision threshold seems to be below the standard threshold of 0.5, at around 0.3. The further the threshold from that point, the lower the F1 score. It also shows how the decision threshold changes the precision and recall of the positive label. Moving the threshold to the right increases precision at the cost of a lower recall, and moving the threshold left means that the recall increases at the cost of precision. Interestingly, the recall seems to decrease linearly with the decision threshold, while the precision has the shape of a logarithmic relationship with the decision threshold. An optimal F1 score emerges at the intersection of the precision and recall.

1 Calibration curves for each of the trained models were also made. Some examples will be put in the appendix. They show some insight into the performance of a single model, but the figures shown in this section summarize them, so we did not cover any of them individually.

Figure 4.3 shows the precision/recall curve for the text mining model. In color the F1 score is indicated. As expected, the top right of the curve has the highest F1 score, since the F1 score is the harmonic mean of the precision and recall.

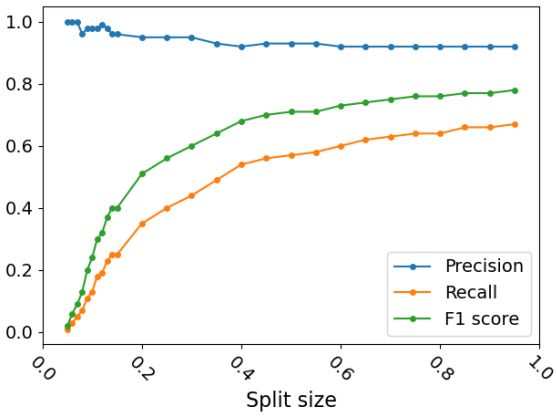


Figure 4.1: The precision, recall and F1 score for different split sizes with an equal decision threshold of 0.5.

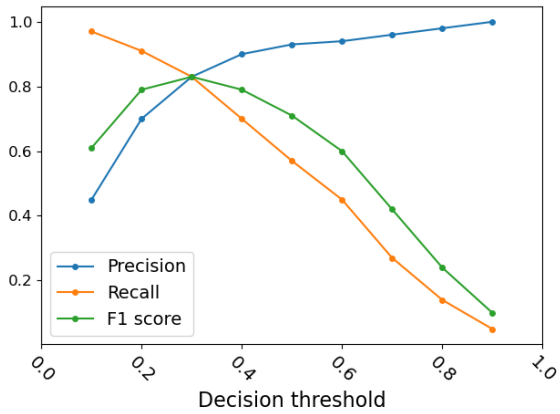


Figure 4.2: The precision, recall and F1 score for different decision thresholds with an equal split size of 0.5.

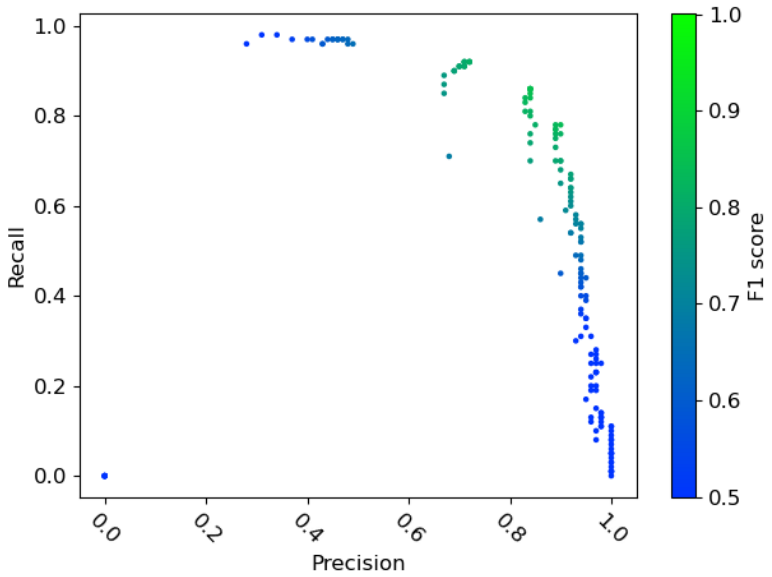


Figure 4.3: The precision, recall curve with the F1 score as color.

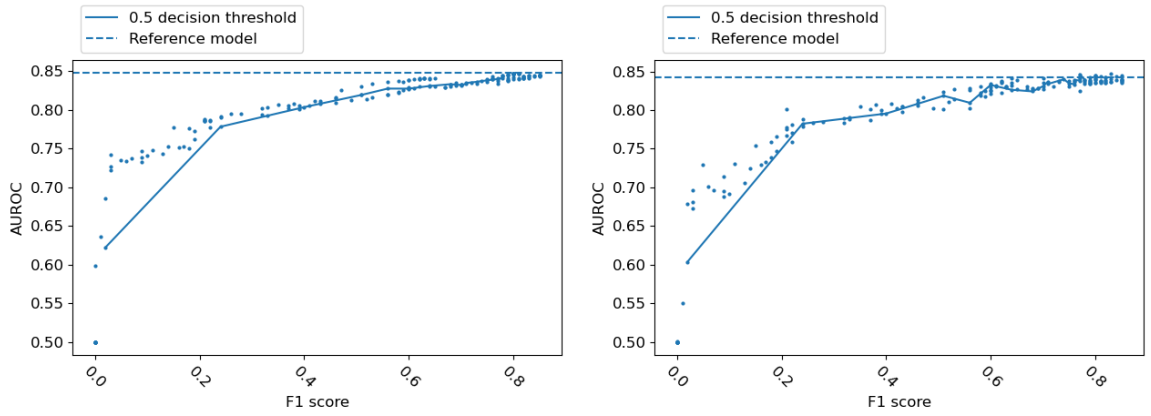
4.2 INTERACTION BETWEEN PREDICTION MODEL AND TEXT MINING MODEL

The prediction model's performance is analyzed in two components; discrimination and calibration, as encouraged by Collins *et al.* [19], Steyerberg and Vergouwe [43] and Van Calster *et al.* [44]. For discrimination we analyze the AUROC, for the calibration metrics we analyze the calibration slope, the calibration intercept and the calibration in the large.

4.2.1 *Discrimination*

4.2.1.1 *Influence of the F1 score*

In Figure 4.5 we can see the effect of the F1 score of the text mining model on the AUROC of both prediction models, with a line indicating the model with a 0.5 decision threshold. The trend is the same for both, where a higher F1 score leads to a higher AUROC. In the logistic regression model, this is true for every data point on the line, while the FFNN shows some noise and randomness, with some points having a lower AUROC despite a higher F1 score. A relatively low F1 score of around 0.5 already yields an AUROC above 0.8, which is generally considered good [73, 74]. We can also see that a high enough F score (in this case 0.8 seems high enough) can approximate the AUROC of the model using the true labels. It should be noted that this relation between F1 score and AUROC may be problem-dependent. The FFNN even manages to surpass the reference model for some data points, but this can be attributed to the property of the FFNN having a random initialization, meaning that the same parameters can lead to different models, and perhaps a lack of optimization on the reference model. The AUROC for both reference models, using the ground truth data instead of text mined data, are very close (logistic regression: 0.85, FFNN: 0.84).



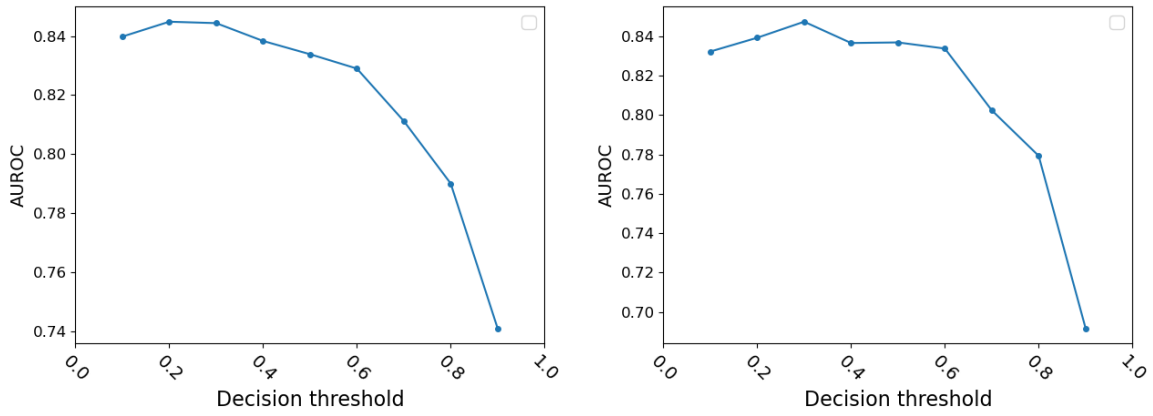
(a) Logistic regression

(b) FFNN

Figure 4.5: The AUROC of the LR model and the FFNN model for the F1 scores of the text mining model.

4.2.1.2 *Influence of shifting the decision threshold*

Figure 4.7 shows how changing the decision threshold of the text mining model changes the AUROC of the prediction model. Since we know from Figure 4.2 that a decision threshold of around 0.3 leads to a higher F1 score and a higher F1 score leads to a higher AUROC, this figure is somewhat implied, but it is still interesting to confirm that a decision threshold of 0.3 led to the model with the highest AUROC for both FFNN and logistic regression. Some explanations for this will be covered in Section 5.



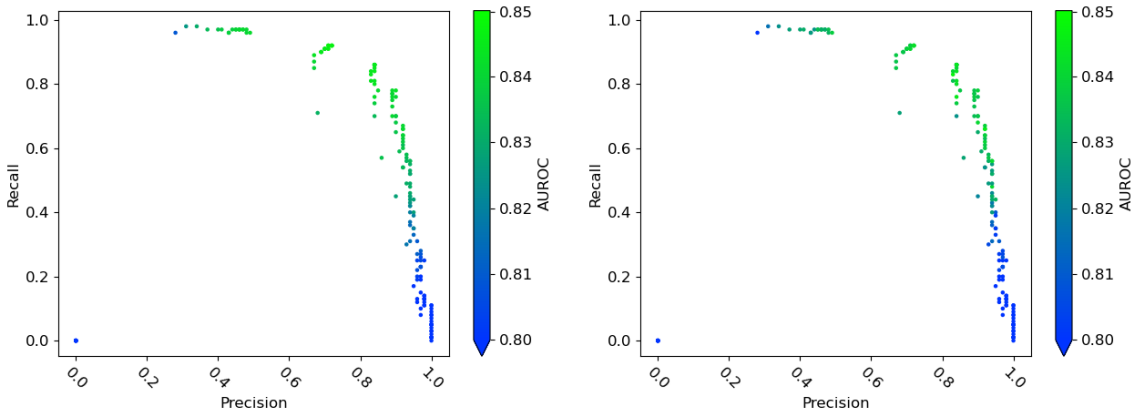
(a) Logistic regression

(b) FFNN

Figure 4.7: The AUROC of the logistic regression model and the FFNN model for the different decision thresholds of the text mining model, with an equal split size of 0.5.

4.2.1.3 *Precision/recall curves*

Figure 4.9 shows how the precision and recall of the text mining algorithm relate to the AUROC of both prediction models. Again, the plot for both the FFNN and the logistic regression model look very similar. We can see that a balanced precision and recall works well for a high AUROC, with an optimum around 0.82 for both the precision and recall, just like in Figure 4.3. However, a model with a slightly higher AUROC can be found on the curve where the recall is slightly higher than the precision. So in this case, retrieving more relevant samples was more important than only retrieving correct samples in order to create the prediction model with the highest AUROC. Some outliers can be seen in the bottom left. These occur when a combination of split size and decision threshold leads to a text mining model that labels all samples with outcome value 0, for example, split size 0.05 with decision threshold 0.9. This will lead to a recall and precision of 0, and the AUROC will be 0.5 since the model has no information to learn from, and will predict that all samples are of outcome 0.



(a) Logistic regression

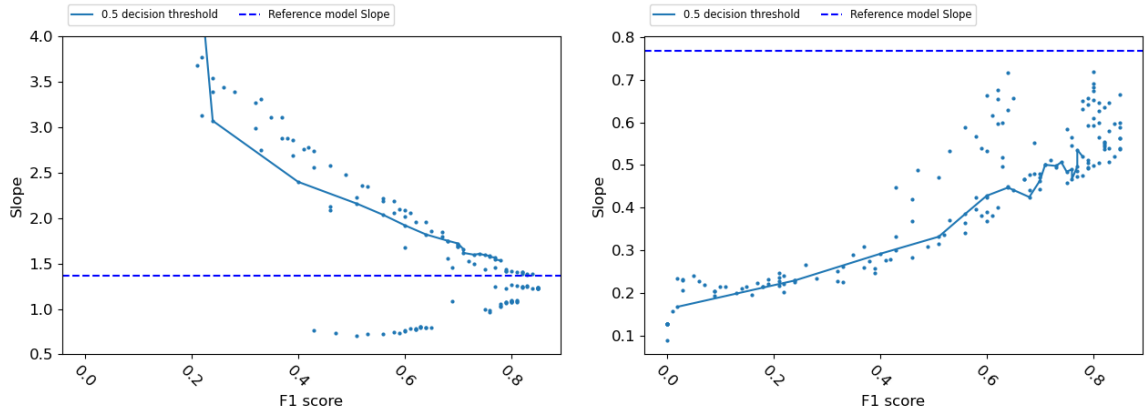
(b) FFNN

Figure 4.9: The precision/recall curve with the AUROC as color for the FFNN model (reference model AUROC: 0.84) and the logistic regression model (reference model AUROC: 0.85).

4.2.2 *Calibration metrics*

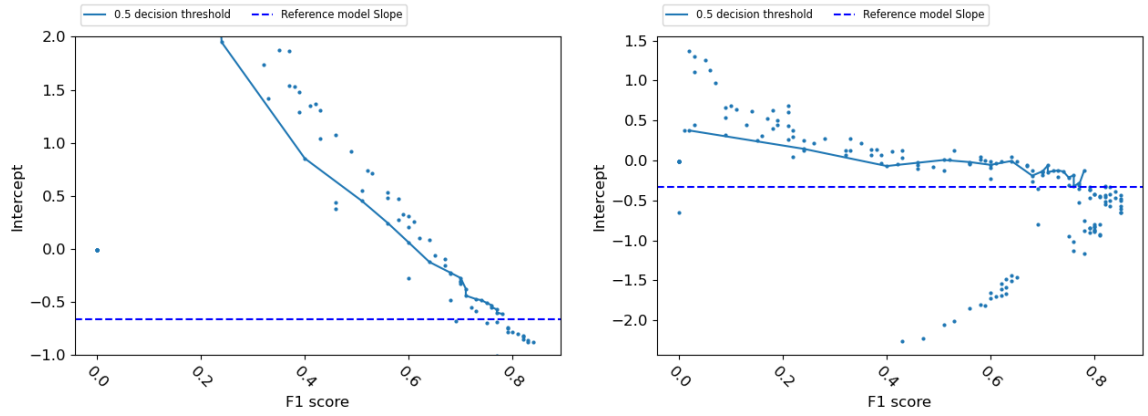
4.2.2.1 *Influence of the F1 score*

The same trend, where a higher F1 score leads to a model that more closely resembles the reference model, continues for the calibration metrics as seen in Figures 4.11, 4.13, 4.15 we can see how the F1 score of the text mining model influences the calibration metrics of the prediction models. Keep in mind that a perfect slope is 1, a perfect intercept is 0 and a perfect CITL is also 0. While for the AUROC comparisons a higher F1 score meant a better model, in this case, the reference model is not better on all metrics than some of the points in the graph, however this statement holds true for the points with the decision threshold fixed at 0.5. In Figures 4.11a and b, we see that the intercept decreases past the ideal intercept of 0 until they approach the reference model. In Figure 4.15a we also see that an increase in F1 score decreases the CITL beyond 0, even though a CITL of 0 would be ideal. For the Intercept and the CITL graphs for the FFNN, we see that a higher F1 score leads to a better performance in that metric.



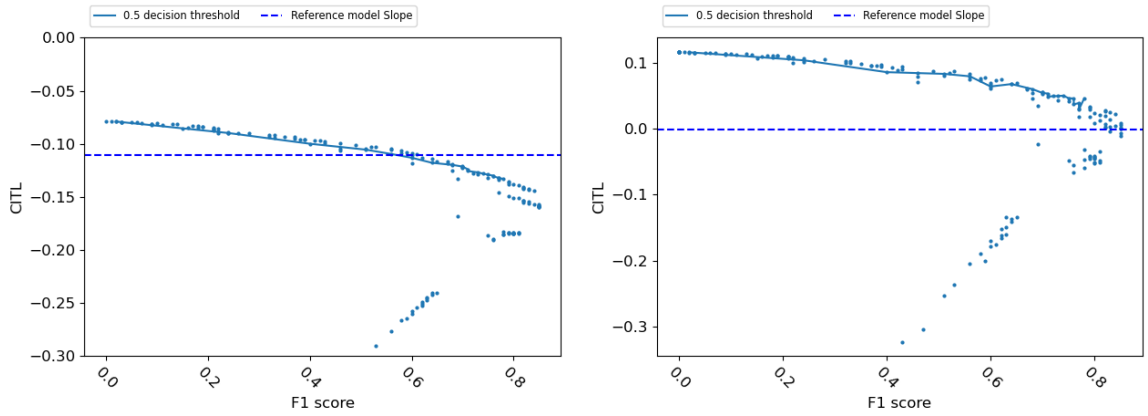
(a) Logistic regression (b) FFNN

Figure 4.11: The calibration slope of the prediction models for the F1 scores of the text mining model.



(a) Logistic regression (b) FFNN

Figure 4.13: The calibration intercept of the prediction models for the F1 scores of the text mining model.

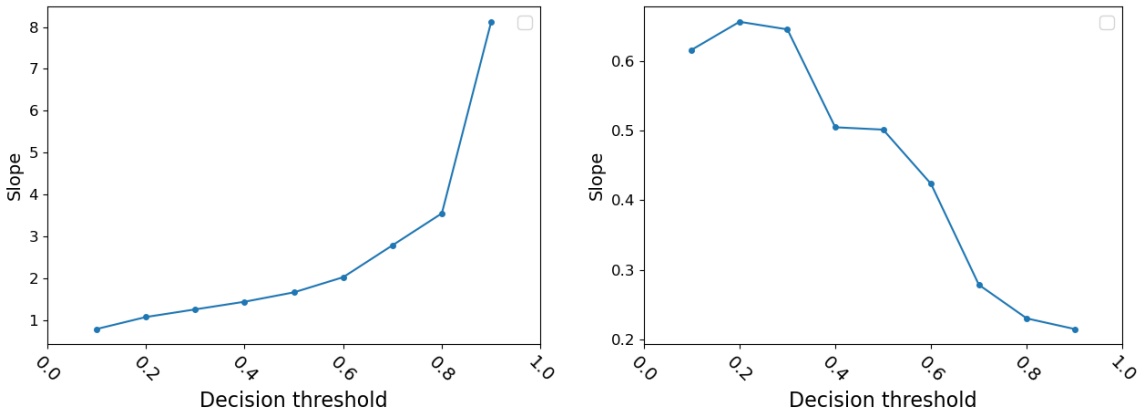


(a) Logistic regression (b) FFNN

Figure 4.15: The calibration in the large of the prediction models for the F1 scores of the text mining model.

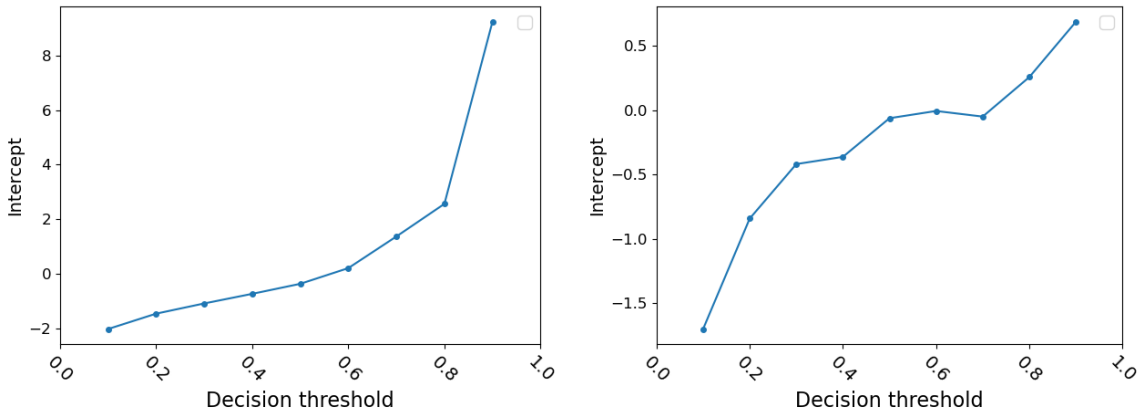
4.2.2.2 *Influence of shifting the decision threshold*

The influence of the decision threshold, as seen in Figures 4.17, 4.19 and 4.21, seems quite uniform across the intercept and calibration in the large for both prediction models, but the slope increases with a shift of the decision threshold to the right for the logistic regression model, while for the FFNN model the opposite is true. This means that the logistic regression model becomes more moderate (predicted probabilities closer to the average) the fewer samples it has to learn from, whereas the FFNN becomes more extreme. The logistic regression has the best slope of around 1 (optimal is 1) at decision threshold 0.1, while the FFNN has the best slope it can achieve at 0.2. The intercept increases for both models as the decision threshold shifts to the right, and both models see an intercept closest to optimal (0) around 0.5 or 0.6. The calibration in the large also increases for both models as the decision threshold shifts to the right, but the FFNN crosses the optimal CITL (0) at a decision threshold of 0.3, the logistic regression model only gets closer but never reaches it.



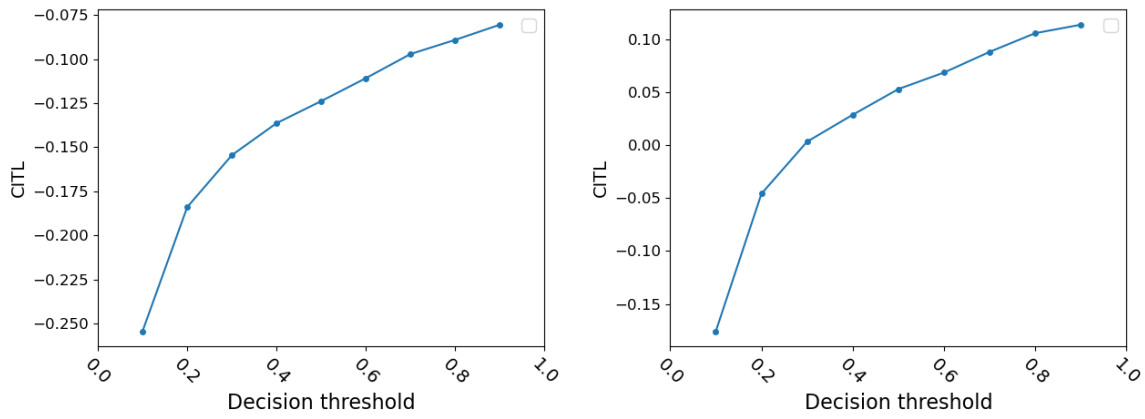
(a) Logistic regression (b) FFNN

Figure 4.17: The calibration slope of the prediction models for the different decision thresholds of the text mining model, with an equal split size of 0.5.



(a) Logistic regression (b) FFNN

Figure 4.19: The calibration intercept of the prediction models for the different decision thresholds of the text mining model, with an equal split size of 0.5.

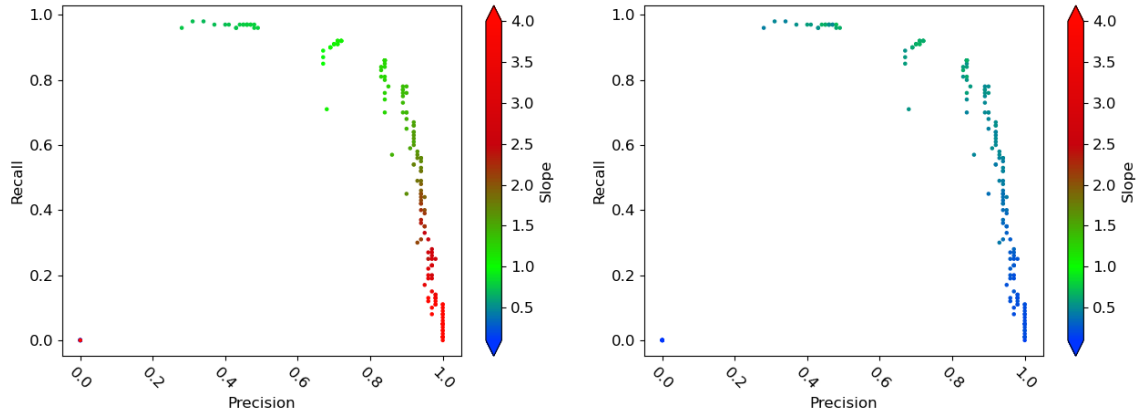


(a) Logistic regression (b) FFNN

Figure 4.21: The calibration in the large of the prediction models for the different decision thresholds of the text mining model, with an equal split size of 0.5.

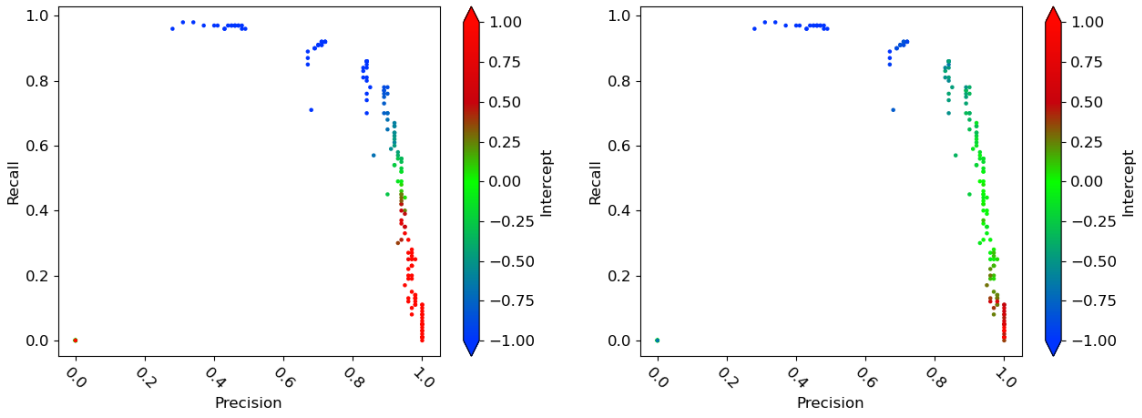
4.2.2.3 *Precision/recall curves*

In Figures 4.23 - 4.27, the precision/recall curves are shown for each of the calibration metrics. Keep in mind that the scale is not always the same for both models; for example, the logistic regression model has a bigger range of values for the slope than the FFNN. Still, the green dots represent the models with the best value for that metric. Important to pay attention to are the points in the top right. With some exceptions, the top right areas are greenest, indicating the best performance measures can be found when precision and recall are balanced. Interestingly, the CITL seems to increase with a higher precision for the logistic regression model, even at a high recall cost. This is not the case for the FFNN.



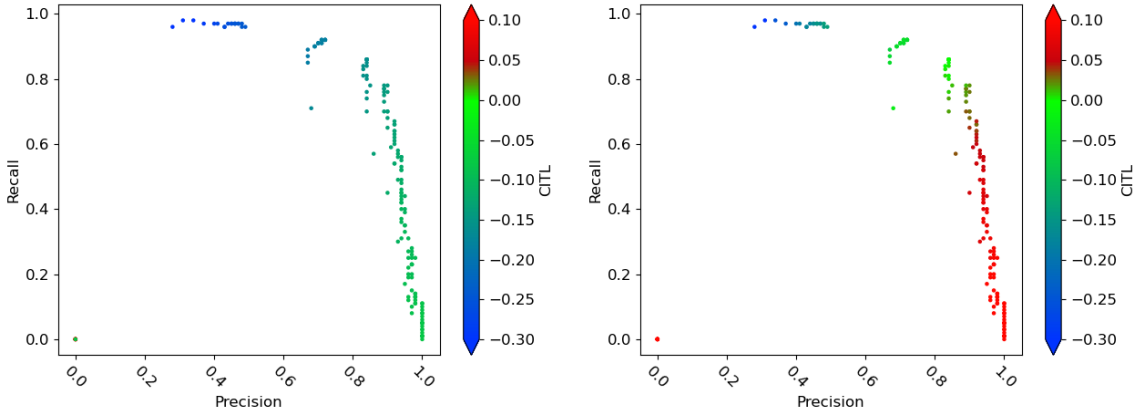
(a) Logistic regression (b) FFNN

Figure 4.23: The precision/recall curve with the calibration slope as color for the prediction models.



(a) Logistic regression (b) FFNN

Figure 4.25: The precision/recall curve with the calibration intercept as color for the prediction models.



(a) Logistic regression (b) FFNN

Figure 4.27: The precision/recall curve with the calibration in the large as color for the prediction models.

5

DISCUSSION

In this section, we discuss the results and talk about the implications they have for future research, and how they can be used in practice to create better text mining- and prediction models. We will also discuss some limitations of this project and how they could be improved upon.

We start by looking at the influence of changing the decision threshold and split size on the performance of the text mining algorithm (RQa). Then we will look at the effect of changing those attributes on the prediction model (RQb). We will also address the differences between our logistic regression model and our feedforward neural network (RQc). Combining these answers will allow us to answer the primary research question: "When using a text mining algorithm for extracting clinical prognosis outcome data, how does the performance of the text mining algorithm (precision, recall, F1 score) affect the performance of the prediction models (discrimination and calibration) trained on that mined data?".

5.1 RESEARCH QUESTIONS

5.1.1 *Secondary Research question a.*

As we could see in Section 4.1, increasing the split size leads to an increase in the F1 score. More data would yield a slightly higher F1 score, but more data does not seem to increase the F1 score after an F1 score of around 0.8. Since the increase in F1 score for the last 40% in training data is small, performance would not have decreased much (judging by precision, recall and F1 score) if there was 40% less data available for this particular problem. The recall also increases with more data, meaning that the text mining model retrieves more relevant information the more training data it has. The precision decreased when the split size increased. This is because our model with very little training data only rarely classifies a sample with a 1 label, so when it does happen it is quite 'certain' of its decision. As we increase the training data, the model learns more associations between the training data and the outcome variable, so it will label more samples with a 1 label. The precision remains relatively high throughout but shows a slight decline, which indicates that precision is quite a poor performance metric by itself.

The decision threshold had a large impact on the text mining performance as well. This makes sense since a high decision threshold resulted in almost no samples being labelled as

1, whereas a low decision threshold resulted in almost all samples being labelled as 1. This is reflected in the precision and recall values for each of the decision thresholds. A high decision threshold means that only samples with a high probability of having label 1 are given label 1. The recall will increase by default by increasing the number of samples that are classified with a 1 label unless all those classifications are of samples with an outcome of 0 (which of course is possible, but gets rarer the more classifications are performed). Interestingly, the optimal F1 score for label 1 is achieved at a decision threshold of 0.3, which could be an artifact of the fact that the text mining model minimizes the log loss. Since our goal is to maximize the F1 score of label 1, this is not always in agreement with the model with the lowest loss, since this loss function is biased towards the majority class. Adjusting the decision threshold manually is one way to fix this.

The extracted outcome variable we have chosen for this research is well-suited for text mining, due to the high F1 score that can be achieved (around 0.8). Even with a rudimentary algorithm as tf-idf, which ignores word order and context, we achieve quite good results. We presume this is in part due to the fact that mortality is almost always included in the discharge summary, so it should theoretically be minable for most of the discharge summaries. We also presume that discharge summaries are less context-driven because they contain many keywords and fewer sentences than texts for typical text mining problems. Since text mining models like BERT are particularly good at context-related tasks in contrast to tf-idf, we expect that in this case, the difference in performance would not be that big. This hypothesis would also explain why tf-idf performs well for this task. Other outcomes could give different results, which would be interesting to explore as well.

5.1.2 *Secondary Research question b.*

The decision threshold has a big influence on both the calibration and the discrimination of the prediction model. In our case, a lower decision threshold of around 0.3 gave the highest AUROC, and some calibration metrics were improved as well. One possible explanation for why a lower decision threshold leads to a higher AUROC is that it allows the text mining algorithm to include patients that are more likely to die than others. This gives the prediction model more high-risk samples to learn from which might be the reason that it can discriminate between the models better. From Table A.1, we can also see that moving the decision threshold can lead to models with limited text mining training data still giving a high AUROC score. Notice the first 9 rows, where the split size is 0.05. With a decision threshold of 0.5, an AUROC of only 0.62 is achieved, due to the limited amount of positive samples being retrieved (recall of 0.01). Moving the decision threshold to 0.2 leads to a big F1 score increase, and thus a big AUROC increase as well. The model can later be recalibrated to increase the calibration metrics.

5.1.3 *Secondary Research question c.*

Our results show us that the logistic regression and FFNN models perform extremely similar regarding AUROC. This is in line with research done by Dreiseitl and Ohno-Machado [75], who established based on a meta-analysis that neither model performs significantly better than the other. On the calibration metrics the FFNN performs better, with the exception that the logistic regression model has a better CITL even when the F1 score is lower. A downside to the FFNN models is that due to their random initialization, the results might vary from run to run. This means that analysis of the performance metrics is slightly harder and might sometimes require repeatedly training a model to rule out a better alternative. Deciding which model to use depends on which of these properties is valued more highly.

5.1.4 *Primary Research question 1.*

In Section 4 we saw that a better F1 score was associated with better discrimination and calibration closer to that of the reference model. Since increasing the text mining training data size leads to a better prediction model, more training data needs to be gathered and sometimes manually annotated, which may be costly. For this reason, we would recommend creating some training data, and plotting the precision, recall and F1 score against the amount of training data used (as in Figure 4.1). Based on the trajectory of the line it may be determined whether more training data is required. In our case study, a text mining model with a relatively low F1 score of 0.5 could already create a prediction model with an AUROC of 0.8 (Figure 4.5). If 0.8 would be considered a high enough AUROC for our problem, we can then read in Figure 4.1 that we only needed about 20% of our training data, which could save around 80% of the training data collection effort. In other words, if we incrementally increased our training data, figured out which decision threshold is best and analyzed the figures named above, we would have been able to conclude at 5% of our training data that we did not need to gather any more.

5.2 IMPLICATIONS FOR CLINICAL PREDICTION MODEL DEVELOPMENT

Analysis of the precision, recall and F1 score versus the discrimination and calibration can be done for any setting of interest, and can be done iteratively; create the plots, determine whether the F1 score line has flattened, and if not, gather more training data and repeat. It also shows the relation between F1 score and AUROC for the setting of interest, which can give insight into whether text mining is suitable, and whether a lacking prediction model performance can be attributed to a poor text mining model. It also opens up possibilities to use a text mining algorithm created by others. Using a text mining algorithm created by others without making any adjustments could result in calibration problems or a poorly performing prediction model. However, by validating the text mining model on a part of your training data some of these problems may be prevented. First, it should be checked

whether a sufficient F1 score can be attained by the text mining model and whether the decision threshold should be adjusted. Then, a prediction model may be created based on the data generated by the text mining model. The same training data used in validating the text mining model may be used if training data is scarce and hard to acquire. This prediction model should then be evaluated on AUROC and calibration metrics.

5.3 LIMITATIONS AND FUTURE RESEARCH

While the data gathered from this research provides a clear insight into the research questions posed, there are some limitations. Fixing these limitations would allow for a more generalized answer to the research questions. First of all, determining the relation between text mining performance and prediction model performance for other outcome variables (such as sepsis) could give new insights based on different settings of interest, and could show that the conclusions reached in this project are generalizable across different project settings. If the results are consistent with ours, it would solidify the conclusion of our research. Secondly, a different text mining algorithm should also be tested. This was initially planned but scrapped when we realized that tf-idf performs quite well on this task. Since we wanted the performance range of our text mining model to be as big as possible we thought we would need to include a more sophisticated text mining algorithm, but tf-idf by itself could create a large enough range. It would still be interesting to see if anything notably changes when another algorithm is used. It would also be interesting to perform a grid search for every split size and decision threshold combination so that each model is optimized for the data it uses. In future research this would be an interesting addition as it might slightly change the results. Another limitation is that we used internal data to test with. For our research purpose this is fine, but to make our findings more applicable to the medical field it would be interesting to see our analysis performed on a separate external data set from another medical institution to test and validate on. This situation, where a text mining model is created by an organization and adopted by another for usage in their own prediction model, would pave the way for broader text mining usage in clinical prediction models. It is therefore a valuable next step to consider.

BIBLIOGRAPHY

[1] W. Bouwmeester, N. P. Zuithoff, S. Mallett, M. I. Geerlings, Y. Vergouwe, E. W. Steyerberg, D. G. Altman and K. G. Moons, 'Reporting and methods in clinical prediction research: A systematic review', *PLoS medicine*, vol. 9, no. 5, e1001221, 2012.

[2] B. E. Keuning, T. Kaufmann, R. Wiersema, A. Granholm, V. Pettilä, M. H. Møller, C. F. Christiansen, J. Castela Forte, H. Snieder, F. Keus *et al.*, 'Mortality prediction models in the adult critically ill: A scoping review', *Acta Anaesthesiologica Scandinavica*, vol. 64, no. 4, pp. 424–442, 2020.

[3] B. Percha, 'Modern clinical text mining: A guide and review', 2021.

[4] H. Dalianis, *Clinical text mining: Secondary use of electronic patient records*. Springer Nature, 2018.

[5] R. Zhu, X. Tu and J. X. Huang, 'Utilizing bert for biomedical and clinical text mining', in *Data Analytics in Biomedical Engineering and Healthcare*, Elsevier, 2021, pp. 73–103.

[6] T. M. Seinen, E. A. Fridgeirsson, S. Ioannou, D. Jeannetot, L. H. John, J. A. Kors, A. F. Markus, V. Pera, A. Rekkas, R. D. Williams *et al.*, 'Use of unstructured text in prognostic clinical prediction models: A systematic review', *Journal of the American Medical Informatics Association*, vol. 29, no. 7, pp. 1292–1302, 2022.

[7] N. Dormosh, M. C. Schut, M. W. Heymans, N. van der Velde and A. Abu-Hanna, 'Development and internal validation of a risk prediction model for falls among older people using primary care electronic health records', *The Journals of Gerontology: Series A*, 2021.

[8] D. A. Grimes and K. F. Schulz, 'Cohort studies: Marching towards outcomes', *The Lancet*, vol. 359, no. 9303, pp. 341–345, 2002.

[9] T. A. Pryor, R. M. Gardner, P. D. Clayton and H. R. Warner, 'The help system', in *Proceedings of the Annual Symposium on Computer Application in Medical Care*, American Medical Informatics Association, 1982, p. 19.

[10] M. F. Collen and M. J. Ball, *The history of medical informatics in the United States*. Springer, 2015.

[11] K. Feldman, N. Hazekamp and N. V. Chawla, 'Mining the clinical narrative: All text are not equal', in *2016 IEEE international conference on healthcare informatics (ICHI)*, IEEE, 2016, pp. 271–280.

[12] M. Valgimigli, C. Ceconi, P. Malagutti, E. Merli, O. Soukhomovskaia, G. Francolini, G. Cicchitelli, A. Olivares, G. Parrinello, G. Percoco *et al.*, 'Tumor necrosis factor-α receptor 1 is a major predictor of mortality and new-onset heart failure in patients with acute myocardial infarction: The cytokine-activation and long-term prognosis in myocardial infarction (c-alpha) study', *Circulation*, vol. 111, no. 7, pp. 863–870, 2005.

[13] J.-Y. Lin, C.-T. Cheng and K.-W. Chau, 'Using support vector machines for long-term discharge prediction', *Hydrological sciences journal*, vol. 51, no. 4, pp. 599–612, 2006.

[14] P. Ramírez, M. Ferrer, V. Martí, S. Reyes, R. Martínez, R. Menéndez, S. Ewig and A. Torres, 'Inflammatory biomarkers and prediction for intensive care unit admission in severe community-acquired pneumonia', *Critical care medicine*, vol. 39, no. 10, pp. 2211–2217, 2011.

[15] A. E. Barnato and D. C. Angus, 'Value and role of intensive care unit outcome prediction models in end-of-life decision making', *Critical care clinics*, vol. 20, no. 3, pp. 345–362, 2004.

[16] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm and N. Elhadad, 'Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission', in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 1721–1730.

[17] D. S. Watson, J. Krutzinna, I. N. Bruce, C. E. Griffiths, I. B. McInnes, M. R. Barnes and L. Floridi, 'Clinical applications of machine learning algorithms: Beyond the black box', *Bmj*, vol. 364, 2019.

[18] J. M. Bland and D. G. Altman, 'Statistical methods for assessing agreement between two methods of clinical measurement', *International journal of nursing studies*, vol. 47, no. 8, pp. 931–936, 2010.

[19] G. S. Collins, J. B. Reitsma, D. G. Altman and K. G. Moons, 'Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): The tripod statement', *Journal of British Surgery*, vol. 102, no. 3, pp. 148–158, 2015.

[20] C. L. Andaur Navarro, J. A. Damen, T. Takada, S. W. Nijman, P. Dhiman, J. Ma, G. S. Collins, R. Bajpai, R. D. Riley, K. G. Moons *et al.*, 'Completeness of reporting of clinical prediction models developed using supervised machine learning: A systematic review', *BMC medical research methodology*, vol. 22, no. 1, pp. 1–13, 2022.

[21] J. Wimsett, A. Harper and P. Jones, 'Components of a good quality discharge summary: A systematic review', *Emergency Medicine Australasia*, vol. 26, no. 5, pp. 430–438, 2014.

[22] G. Mujtaba, L. Shuib, N. Idris, W. L. Hoo, R. G. Raj, K. Khowaja, K. Shaikh and H. F. Nweke, 'Clinical text classification research trends: Systematic literature review and open issues', *Expert systems with applications*, vol. 116, pp. 494–520, 2019.

[23] T. Schick and H. Schütze, 'Rare words: A major problem for contextualized embeddings and how to fix it by attentive mimicking', in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 8766–8774.

[24] A. Maurice, S. Chan, C. Pollard, R. Kidd, S. Ayre, H. Ward and D. Walters, *Web supplement*, Aug. 2014.

[25] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler and C. G. Chute, 'Mayo clinical text analysis and knowledge extraction system (ctakes): Architecture, component evaluation and applications', *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 507–513, 2010.

[26] S. Gehrmann, F. Dernoncourt, Y. Li, E. T. Carlson, J. T. Wu, J. Welt, J. Foote Jr, E. T. Moseley, D. W. Grant, P. D. Tyler *et al.*, 'Comparing rule-based and deep learning models for patient phenotyping', *arXiv preprint arXiv:1703.08705*, 2017.

[27] S. Kulshrestha, D. Dligach, C. Joyce, M. S. Baker, R. Gonzalez, A. P. O'Rourke, J. M. Glazer, A. Stey, J. M. Kruser, M. M. Churpek *et al.*, 'Prediction of severe chest injury using natural language processing from the electronic health record', *Injury*, vol. 52, no. 2, pp. 205–212, 2021.

[28] A. R. Aronson and F.-M. Lang, 'An overview of metamap: Historical perspective and recent advances', *Journal of the American Medical Informatics Association*, vol. 17, no. 3, pp. 229–236, 2010.

[29] R. Reátegui and S. Ratté, 'Comparison of metamap and ctakes for entity extraction in clinical notes', *BMC medical informatics and decision making*, vol. 18, no. 3, pp. 13–19, 2018.

[30] E. Ford, J. A. Carroll, H. E. Smith, D. Scott and J. A. Cassell, 'Extracting information from the text of electronic medical records to improve case detection: A systematic review', *Journal of the American Medical Informatics Association*, vol. 23, no. 5, pp. 1007–1015, 2016.

[31] D. C. Howell, 'The treatment of missing data', *The Sage handbook of social science methodology*, pp. 208–224, 2007.

[32] H. Kang, 'The prevention and handling of the missing data', *Korean journal of anesthesiology*, vol. 64, no. 5, pp. 402–406, 2013.

[33] P. D. Allison, *Missing data*. Sage publications, 2001.

[34] M. Erraguntla, B. Gopal, S. Ramachandran and R. Mayer, 'Inference of missing icd 9 codes using text mining and nearest neighbor techniques', in *2012 45th hawaii international conference on system sciences*, IEEE, 2012, pp. 1060–1069.

[35] T. R. Hylan, M. Von Korff, K. Saunders, E. Masters, R. E. Palmer, D. Carrell, D. Cronkite, J. Mardekian and D. Gross, 'Automated prediction of risk for problem opioid use in a primary care setting', *The Journal of Pain*, vol. 16, no. 4, pp. 380–387, 2015.

[36] H. Jones and L. d. Cossart, 'Risk scoring in surgical patients', *British Journal of Surgery*, vol. 86, no. 2, pp. 149–157, 1999.

[37] J.-R. Le Gall, P. Loirat, A. Alperovitch, P. Glaser, C. Granthil, D. Mathieu, P. Mercier, R. Thomas and D. Villers, 'A simplified acute physiology score for icu patients.', *Critical care medicine*, vol. 12, no. 11, pp. 975–977, 1984.

[38] J. E. Zimmerman, A. A. Kramer, D. S. McNair and F. M. Malila, 'Acute physiology and chronic health evaluation (apache) iv: Hospital mortality assessment for today's critically ill patients', *Critical care medicine*, vol. 34, no. 5, pp. 1297–1310, 2006.

[39] D. L. Shung, B. Au, R. A. Taylor, J. K. Tay, S. B. Laursen, A. J. Stanley, H. R. Dalton, J. Ngu, M. Schultz and L. Laine, 'Validation of a machine learning model that outperforms clinical risk scoring systems for upper gastrointestinal bleeding', *Gastroenterology*, vol. 158, no. 1, pp. 160–167, 2020.

[40] G. Y. Lip, G. Tran, A. Genaidy, P. Marroquin, C. Estes and J. Landsheft, 'Improving dynamic stroke risk prediction in non-anticoagulated patients with and without atrial fibrillation: Comparing common clinical risk scores and machine learning algorithms', *European Heart Journal-Quality of Care and Clinical Outcomes*, 2021.

[41] A. C. Dimopoulos, M. Nikolaidou, F. F. Caballero, W. Engchuan, A. Sanchez-Niubo, H. Arndt, J. L. Ayuso-Mateos, J. M. Haro, S. Chatterji, E. N. Georgousopoulou *et al.*, 'Machine learning methodologies versus cardiovascular risk scores, in predicting disease risk', *BMC medical research methodology*, vol. 18, no. 1, pp. 1–11, 2018.

[42] Z. H. Hoo, J. Candlish and D. Teare, *What is an roc curve?*, 2017.

[43] E. W. Steyerberg and Y. Vergouwe, 'Towards better clinical prediction models: Seven steps for development and an abcd for validation', *European heart journal*, vol. 35, no. 29, pp. 1925–1931, 2014.

[44] B. Van Calster, D. J. McLernon, M. Van Smeden, L. Wynants and E. W. Steyerberg, 'Calibration: The achilles heel of predictive analytics', *BMC medicine*, vol. 17, no. 1, pp. 1–7, 2019.

[45] G. J. Herder, H. Van Tinteren, R. P. Golding, P. J. Kostense, E. F. Comans, E. F. Smit and O. S. Hoekstra, 'Clinical prediction model to characterize pulmonary nodules: Validation and added value of 18 f-fluorodeoxyglucose positron emission tomography', *Chest*, vol. 128, no. 4, pp. 2490–2496, 2005.

[46] S. E. Davis, T. A. Lasko, G. Chen, E. D. Siew and M. E. Matheny, 'Calibration drift in regression and machine learning models for acute kidney injury', *Journal of the American Medical Informatics Association*, vol. 24, no. 6, pp. 1052–1061, 2017.

[47] E. W. Steyerberg, G. J. Borsboom, H. C. van Houwelingen, M. J. Eijkemans and J. D. F. Habbema, 'Validation and updating of predictive logistic regression models: A study on sample size and shrinkage', *Statistics in medicine*, vol. 23, no. 16, pp. 2567–2586, 2004.

[48] M. E. Shipe, S. A. Deppen, F. Farjah and E. L. Grogan, 'Developing prediction models for clinical use using logistic regression: An overview', *Journal of thoracic disease*, vol. 11, no. Suppl 4, S574, 2019.

[49] J. Platt *et al.*, 'Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods', *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[50] B. Zadrozny and C. Elkan, 'Transforming classifier scores into accurate multiclass probability estimates', in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2002, pp. 694–699.

[51] A. Niculescu-Mizil and R. Caruana, 'Predicting good probabilities with supervised learning', in *Proceedings of the 22nd international conference on Machine learning*, 2005, pp. 625–632.

[52] P. Kocbek, N Fijacko, M. Zorman, S. Kocbek and G. Štiglic, 'Improving mortality prediction for intensive care unit patients using text mining techniques', in *Proceedings of SiKDD 2017 Conference on Data Mining and Data Warehouses*, vol. 29, 2017, pp. 31–32.

[53] E. Klang, M. A. Levin, S. Soffer, A. Zebrowski, B. S. Glicksberg, B. G. Carr, J. Mcgreevy, D. L. Reich and R. Freeman, 'A simple free-text-like method for extracting semi-structured data from electronic health records: Exemplified in prediction of in-hospital mortality', *Big Data and Cognitive Computing*, vol. 5, no. 3, p. 40, 2021.

[54] I. Tenney, D. Das and E. Pavlick, 'Bert rediscovers the classical nlp pipeline', *arXiv preprint arXiv:1905.05950*, 2019.

[55] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, 'Attention is all you need', *Advances in neural information processing systems*, vol. 30, 2017.

[56] K. Huang, J. Altosaar and R. Ranganath, 'Clinicalbert: Modeling clinical notes and predicting hospital readmission', *arXiv preprint arXiv:1904.05342*, 2019.

[57] L. Rasmy, Y. Xiang, Z. Xie, C. Tao and D. Zhi, 'Med-bert: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction', *NPJ digital medicine*, vol. 4, no. 1, pp. 1–13, 2021.

[58] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So and J. Kang, 'Biobert: A pre-trained biomedical language representation model for biomedical text mining', *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[59] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi and R. G. Mark, 'Mimic-iii, a freely accessible critical care database', *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[60] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson *et al.*, 'Extracting training data from large language models', in *30th USENIX Security Symposium (USENIX Security 21)*, 2021, pp. 2633–2650.

[61] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, 'Language models are unsupervised multitask learners', *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[62] E. Lehman, S. Jain, K. Pichotta, Y. Goldberg and B. C. Wallace, 'Does bert pretrained on clinical notes reveal sensitive data?', *arXiv preprint arXiv:2104.07762*, 2021.

[63] G. Clermont, D. C. Angus, S. M. DiRusso, M. Griffin and W. T. Linde-Zwirble, 'Predicting hospital mortality for patients in the intensive care unit: A comparison of artificial neural networks with logistic regression models', *Critical care medicine*, vol. 29, no. 2, pp. 291–296, 2001.

[64] B. Eftekhar, K. Mohammad, H. E. Ardebili, M. Ghodsi and E. Ketabchi, 'Comparison of artificial neural network and logistic regression models for prediction of mortality in head trauma based on initial clinical data', *BMC medical informatics and decision making*, vol. 5, no. 1, pp. 1–8, 2005.

[65] F. Jaimes, J. Farbiarz, D. Alvarez and C. Martínez, 'Comparison between logistic regression and neural networks to predict death in patients with suspected sepsis in the emergency room', *Critical care*, vol. 9, no. 2, pp. 1–7, 2005.

[66] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel and B. Van Calster, 'A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models', *Journal of clinical epidemiology*, vol. 110, pp. 12–22, 2019.

[67] P. R. Norvig and S. A. Intelligence, 'A modern approach', *Prentice Hall Upper Saddle River, NJ, USA: Rani, M., Nayak, R., & Vyas, OP (2015). An ontology-based adaptive personalized e-learning system, assisted by software agents on cloud storage. Knowledge-Based Systems*, vol. 90, pp. 33–48, 2002.

[68] I. J. Myung, 'Tutorial on maximum likelihood estimation', *Journal of mathematical Psychology*, vol. 47, no. 1, pp. 90–100, 2003.

[69] M. van Smeden, J. B. Reitsma, R. D. Riley, G. S. Collins and K. G. Moons, 'Clinical prediction models: Diagnosis versus prognosis', *Journal of Clinical Epidemiology*, vol. 132, pp. 142–145, 2021.

[70] S Li, F Shi, F Pu, X. Li, T. Jiang, S Xie and Y Wang, 'Hippocampal shape analysis of alzheimer disease based on machine learning methods', *American Journal of Neuroradiology*, vol. 28, no. 7, pp. 1339–1345, 2007.

[71] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis and D. I. Fotiadis, 'Machine learning applications in cancer prognosis and prediction', *Computational and structural biotechnology journal*, vol. 13, pp. 8–17, 2015.

[72] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. Ver Steeg and A. Galstyan, 'Multitask learning and benchmarking with clinical time series data', *Scientific data*, vol. 6, no. 1, pp. 1–18, 2019.

[73] D. E. Shapiro, 'The interpretation of diagnostic tests', *Statistical methods in medical research*, vol. 8, no. 2, pp. 113–134, 1999.

[74] D. W. Hosmer Jr, S. Lemeshow and R. X. Sturdivant, *Applied logistic regression*. John Wiley & Sons, 2013, vol. 398.

[75] S. Dreiseitl and L. Ohno-Machado, 'Logistic regression and artificial neural network classification models: A methodology review', *Journal of biomedical informatics*, vol. 35, no. 5-6, pp. 352–359, 2002.

APPENDIX 1

A.1 CODE REPOSITORY

The code can be found at https://github.com/zwierd99/ClinicalTM. Keep in mind that this code will not run without having the MIMIC-III database. However, for transparency and completeness the code is shared.

A.2 CALIBRATION CURVES

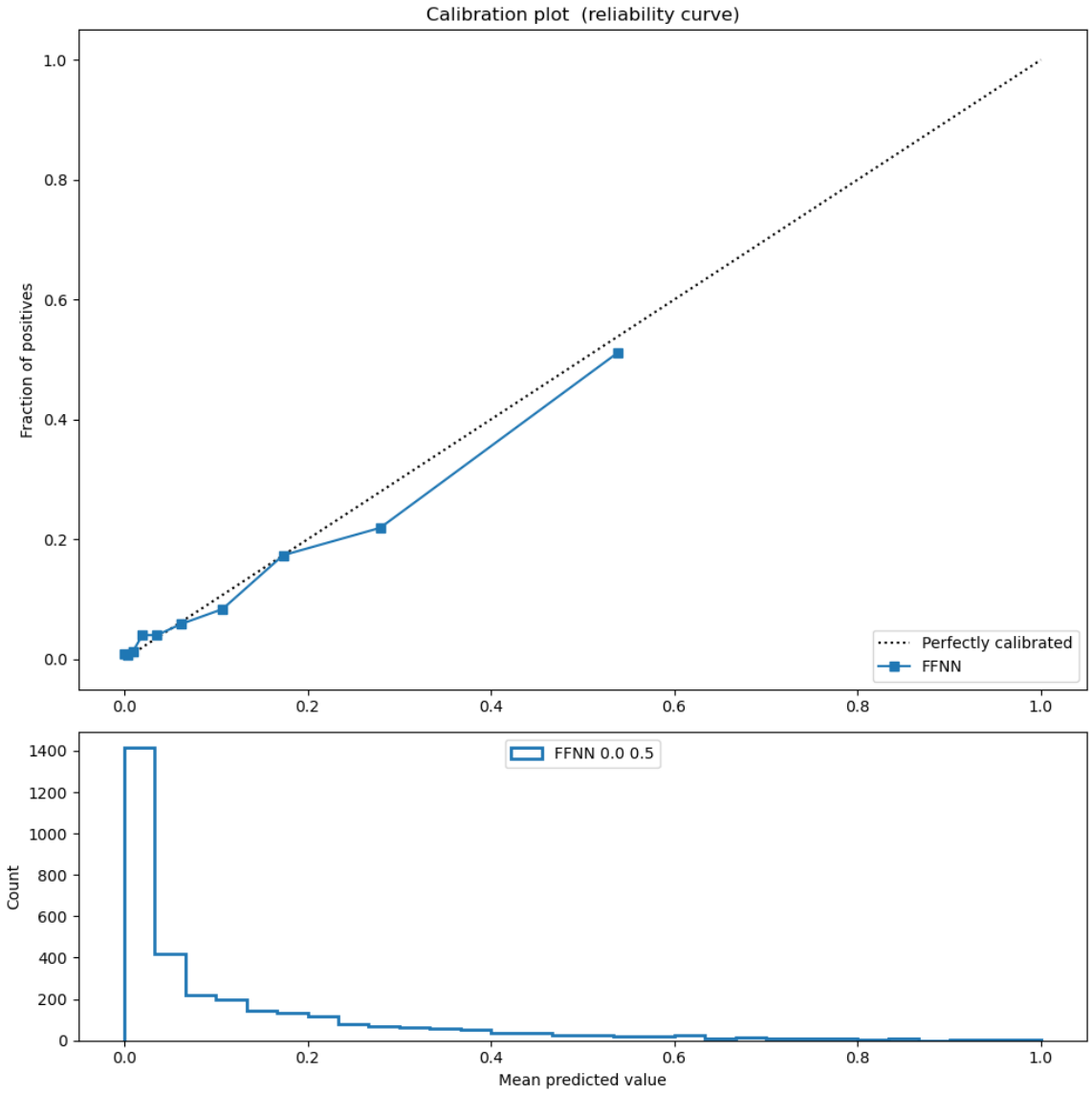


Figure A.1: The calibration curve of the FFNN reference model on the test set.

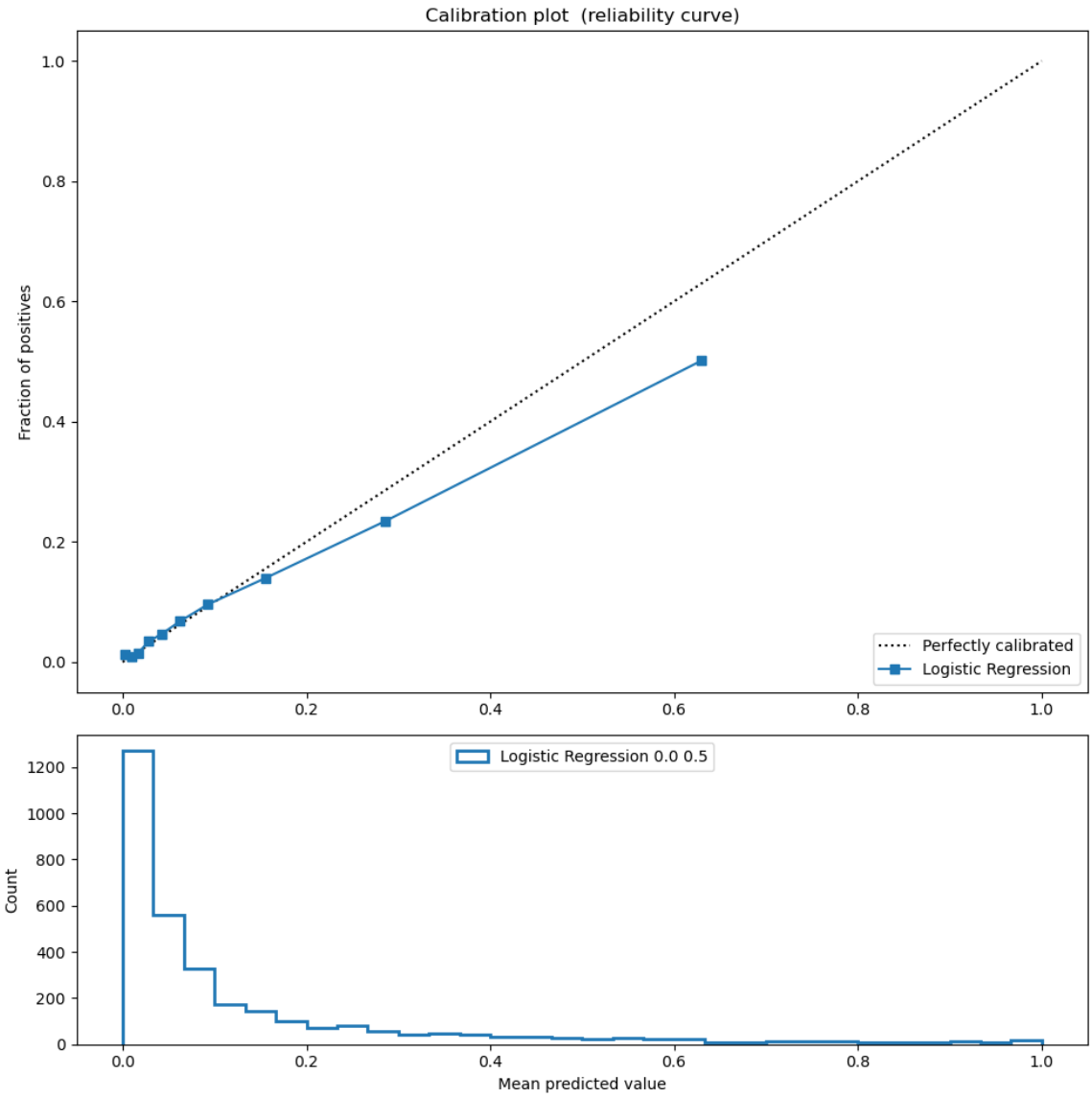


Figure A.2: The calibration curve of the logistic regression reference model on the test set.

A.3 DATA TABLE

Table A.1: Data generated from our experiments from the Logistic Regression model. The split size = 0.0 value is the reference model using the ground truth data.

Split size	Decision threshold	AUROC	Precision	Recall	F1 score	Slope	Intercept	CITL
0.00000	0.50000	0.84849	1	1	1	1.36956	-0.65799	-0.11040
0.05000	0.10000	0.80913	0.28000	0.96000	0.43000	0.76132	-2.36803	-0.34545
0.05000	0.20000	0.83182	0.68000	0.71000	0.69000	1.08360	-1.31554	-0.16791
0.05000	0.30000	0.81695	0.93000	0.30000	0.46000	2.09071	0.37949	-0.10626
0.05000	0.40000	0.75192	1.00000	0.08000	0.16000	4.08050	3.45686	-0.08473
0.05000	0.50000	0.62232	1.00000	0.01000	0.02000	17.36474	22.55112	-0.07898
0.05000	0.60000	0.50000	0.00000	0.00000	0.00000	0.12715	-0.00798	0.11557

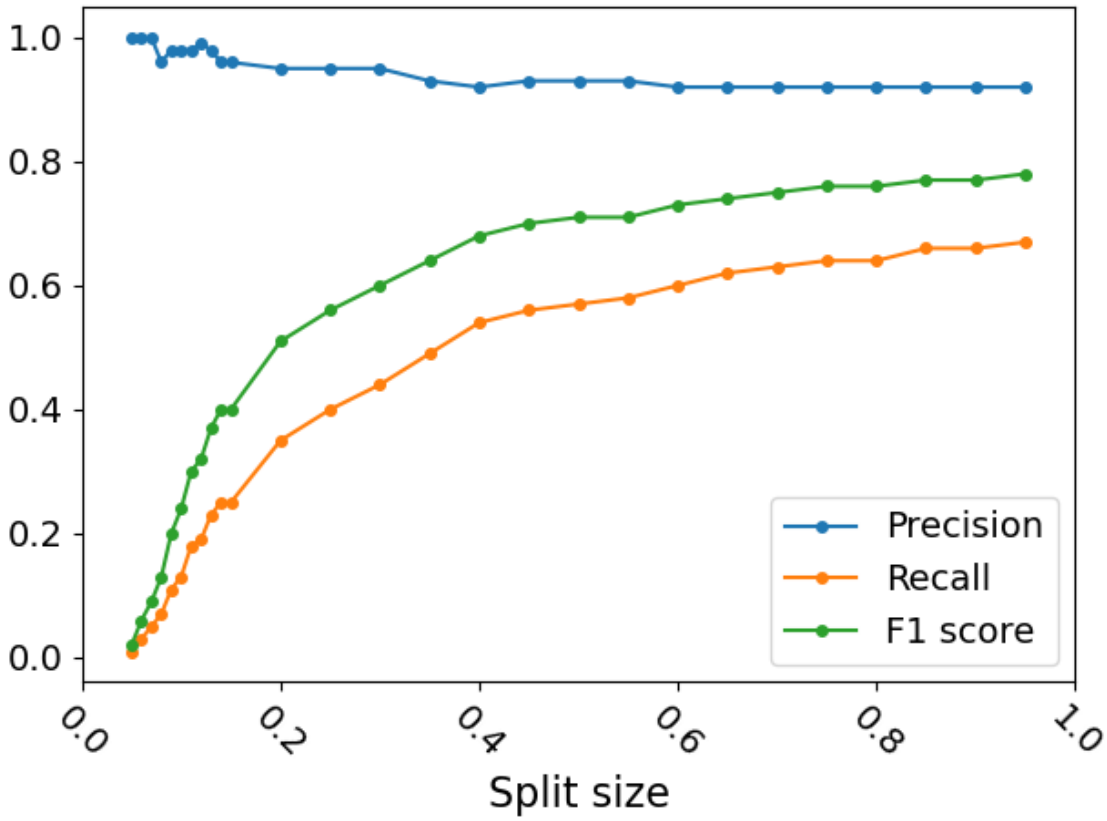
0.05000	0.70000	0.50000	0.00000	0.00000	0.00000	0.12715	-0.00798	0.11557
0.05000	0.80000	0.50000	0.00000	0.00000	0.00000	0.12715	-0.00798	0.11557
0.05000	0.90000	0.50000	0.00000	0.00000	0.00000	0.12715	-0.00798	0.11557
0.10000	0.10000	0.82546	0.31000	0.98000	0.47000	0.73134	-2.36795	-0.33126
0.10000	0.20000	0.83999	0.67000	0.85000	0.75000	0.99705	-1.52329	-0.18626
0.10000	0.30000	0.83583	0.86000	0.57000	0.69000	1.46123	-0.67984	-0.13313
0.10000	0.40000	0.81513	0.94000	0.31000	0.46000	2.13181	0.44234	-0.10493
0.10000	0.50000	0.77841	0.98000	0.13000	0.24000	3.07287	1.94853	-0.08975
0.10000	0.60000	0.73926	1.00000	0.05000	0.09000	6.21332	6.53938	-0.08154
0.10000	0.70000	0.68520	1.00000	0.01000	0.02000	19.58157	25.63759	-0.07905
0.10000	0.80000	0.50000	0.00000	0.00000	0.00000	0.12715	-0.00798	0.11557
0.10000	0.90000	0.50000	0.00000	0.00000	0.00000	0.12715	-0.00798	0.11557
0.15000	0.10000	0.83012	0.34000	0.98000	0.51000	0.70802	-2.27384	-0.30573
0.15000	0.20000	0.84090	0.67000	0.87000	0.76000	0.96783	-1.57457	-0.19022
0.15000	0.30000	0.83381	0.84000	0.70000	0.77000	1.24610	-1.00289	-0.14595
0.15000	0.40000	0.82787	0.90000	0.45000	0.60000	1.67501	-0.27240	-0.11785
0.15000	0.50000	0.80373	0.96000	0.25000	0.40000	2.40083	0.85419	-0.09985
0.15000	0.60000	0.77810	0.98000	0.13000	0.22000	3.13344	2.02506	-0.08964
0.15000	0.70000	0.74639	1.00000	0.05000	0.09000	5.86227	6.01723	-0.08204
0.15000	0.80000	0.72701	1.00000	0.01000	0.03000	20.11889	26.32341	-0.07921
0.15000	0.90000	0.50000	0.00000	0.00000	0.00000	0.12715	-0.00798	0.11557
0.20000	0.10000	0.83367	0.37000	0.97000	0.53000	0.72069	-2.22871	-0.29046
0.20000	0.20000	0.84359	0.67000	0.89000	0.76000	0.98563	-1.56402	-0.18972
0.20000	0.30000	0.83863	0.84000	0.74000	0.79000	1.22520	-1.06273	-0.14964
0.20000	0.40000	0.83332	0.92000	0.54000	0.68000	1.56177	-0.47946	-0.12481
0.20000	0.50000	0.81959	0.95000	0.35000	0.51000	2.16090	0.45316	-0.10542
0.20000	0.60000	0.79298	0.97000	0.20000	0.33000	2.75471	1.41814	-0.09393
0.20000	0.70000	0.77737	0.97000	0.08000	0.15000	4.35922	3.76834	-0.08572
0.20000	0.80000	0.73391	1.00000	0.03000	0.06000	8.80049	10.25465	-0.07998
0.20000	0.90000	0.59868	0.00000	0.00000	0.00000	51.64269	71.39243	-0.07858
0.25000	0.10000	0.83589	0.40000	0.97000	0.56000	0.72052	-2.16118	-0.27680
0.25000	0.20000	0.84422	0.69000	0.90000	0.78000	1.02828	-1.50652	-0.18573
0.25000	0.30000	0.84189	0.84000	0.76000	0.80000	1.26461	-1.04430	-0.15098
0.25000	0.40000	0.83470	0.91000	0.59000	0.72000	1.52589	-0.54649	-0.12746
0.25000	0.50000	0.82775	0.95000	0.40000	0.56000	2.03601	0.24454	-0.10954
0.25000	0.60000	0.80156	0.98000	0.25000	0.39000	2.69528	1.28763	-0.09600
0.25000	0.70000	0.78638	0.98000	0.11000	0.21000	3.68321	2.79407	-0.08767
0.25000	0.80000	0.73245	1.00000	0.05000	0.09000	7.15041	7.86989	-0.08097
0.25000	0.90000	0.63678	1.00000	0.00000	0.01000	26.54588	35.59491	-0.07881
0.30000	0.10000	0.83765	0.41000	0.97000	0.58000	0.74598	-2.10008	-0.26589
0.30000	0.20000	0.84316	0.69000	0.90000	0.78000	1.05763	-1.46606	-0.18301
0.30000	0.30000	0.84172	0.85000	0.78000	0.81000	1.25907	-1.05294	-0.15102

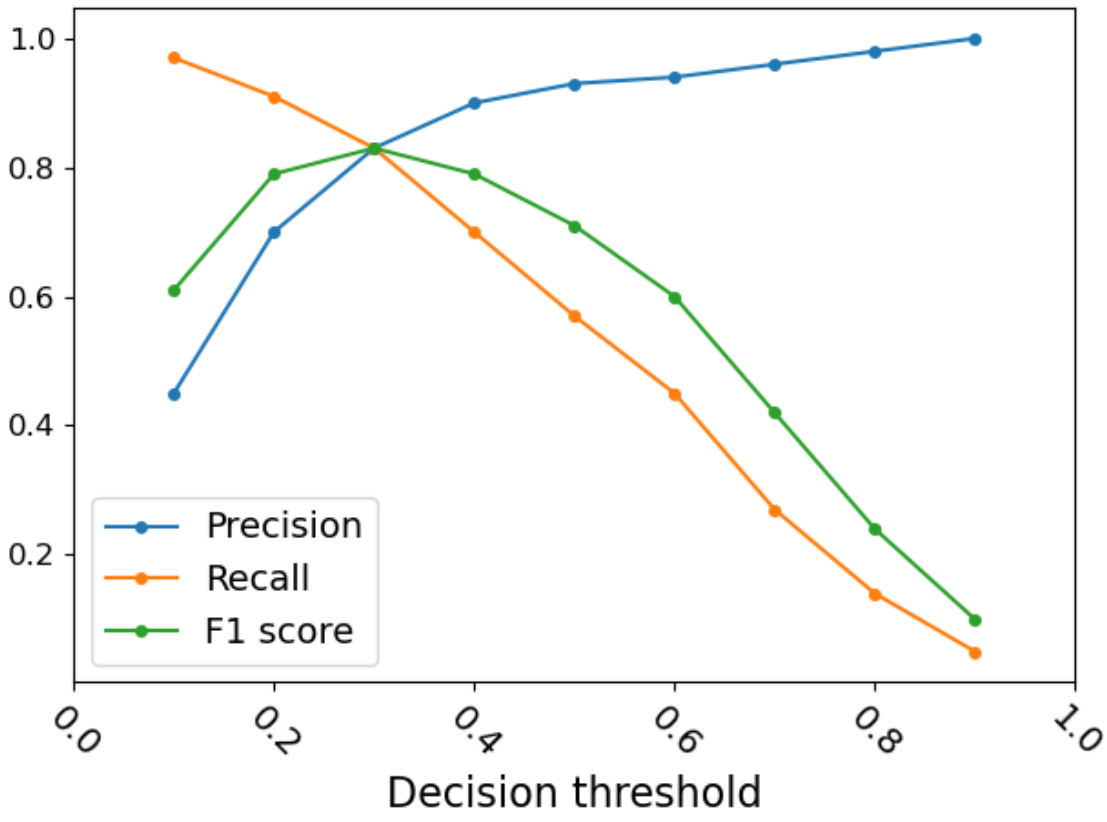
0.30000	0.40000	0.83411	0.92000	0.61000	0.73000	1.50047	-0.58391	-0.12845
0.30000	0.50000	0.82770	0.95000	0.44000	0.60000	1.92227	0.06202	-0.11305
0.30000	0.60000	0.81125	0.97000	0.28000	0.43000	2.55771	1.03979	-0.09953
0.30000	0.70000	0.79175	0.98000	0.14000	0.24000	3.38751	2.32055	-0.09032
0.30000	0.80000	0.74853	1.00000	0.06000	0.11000	5.57642	5.60367	-0.08233
0.30000	0.90000	0.72276	1.00000	0.01000	0.03000	16.67751	21.45208	-0.07926
0.35000	0.10000	0.83617	0.43000	0.96000	0.59000	0.73768	-2.09029	-0.26426
0.35000	0.20000	0.84465	0.69000	0.90000	0.78000	1.03155	-1.50100	-0.18512
0.35000	0.30000	0.84353	0.84000	0.80000	0.82000	1.23790	-1.10014	-0.15350
0.35000	0.40000	0.83625	0.90000	0.65000	0.75000	1.43439	-0.69440	-0.13233
0.35000	0.50000	0.83110	0.93000	0.49000	0.64000	1.82066	-0.12234	-0.11783
0.35000	0.60000	0.81964	0.95000	0.35000	0.51000	2.22540	0.55415	-0.10413
0.35000	0.70000	0.79432	0.97000	0.19000	0.32000	2.99485	1.73425	-0.09315
0.35000	0.80000	0.77654	0.97000	0.10000	0.18000	4.78539	4.36189	-0.08493
0.35000	0.90000	0.74224	1.00000	0.02000	0.03000	14.44527	18.24455	-0.07946
0.40000	0.10000	0.83869	0.43000	0.96000	0.60000	0.75388	-2.06621	-0.25990
0.40000	0.20000	0.84569	0.70000	0.91000	0.79000	1.07348	-1.46802	-0.18393
0.40000	0.30000	0.84403	0.84000	0.81000	0.82000	1.24807	-1.09224	-0.15380
0.40000	0.40000	0.83742	0.90000	0.68000	0.77000	1.46065	-0.68530	-0.13363
0.40000	0.50000	0.83345	0.92000	0.54000	0.68000	1.75238	-0.22192	-0.11977
0.40000	0.60000	0.81996	0.95000	0.39000	0.56000	2.19373	0.48250	-0.10576
0.40000	0.70000	0.80222	0.97000	0.23000	0.37000	2.87705	1.53899	-0.09482
0.40000	0.80000	0.78707	0.98000	0.11000	0.21000	4.23719	3.55979	-0.08650
0.40000	0.90000	0.73468	1.00000	0.03000	0.05000	9.17748	10.78716	-0.07997
0.45000	0.10000	0.83843	0.44000	0.97000	0.60000	0.76809	-2.04707	-0.25749
0.45000	0.20000	0.84533	0.70000	0.91000	0.79000	1.07899	-1.46447	-0.18408
0.45000	0.30000	0.84384	0.83000	0.81000	0.82000	1.24379	-1.10661	-0.15494
0.45000	0.40000	0.83747	0.90000	0.70000	0.79000	1.41697	-0.74906	-0.13561
0.45000	0.50000	0.83239	0.93000	0.56000	0.70000	1.72177	-0.27252	-0.12139
0.45000	0.60000	0.82334	0.94000	0.42000	0.58000	2.05566	0.27778	-0.10895
0.45000	0.70000	0.80584	0.97000	0.25000	0.39000	2.85914	1.48311	-0.09601
0.45000	0.80000	0.78563	0.98000	0.13000	0.22000	3.77512	2.90005	-0.08798
0.45000	0.90000	0.73706	1.00000	0.04000	0.07000	9.48519	11.20224	-0.08012
0.50000	0.10000	0.83982	0.45000	0.97000	0.61000	0.78354	-2.03424	-0.25434
0.50000	0.20000	0.84487	0.70000	0.91000	0.79000	1.07016	-1.46962	-0.18399
0.50000	0.30000	0.84438	0.83000	0.83000	0.83000	1.25297	-1.09486	-0.15451
0.50000	0.40000	0.83834	0.90000	0.70000	0.79000	1.43559	-0.74181	-0.13638
0.50000	0.50000	0.83385	0.93000	0.57000	0.71000	1.65963	-0.37357	-0.12387
0.50000	0.60000	0.82899	0.94000	0.45000	0.60000	2.02214	0.20261	-0.11086
0.50000	0.70000	0.81122	0.96000	0.27000	0.42000	2.78496	1.36402	-0.09717
0.50000	0.80000	0.79005	0.98000	0.14000	0.24000	3.54759	2.56251	-0.08915
0.50000	0.90000	0.74099	1.00000	0.05000	0.10000	8.12504	9.24961	-0.08055

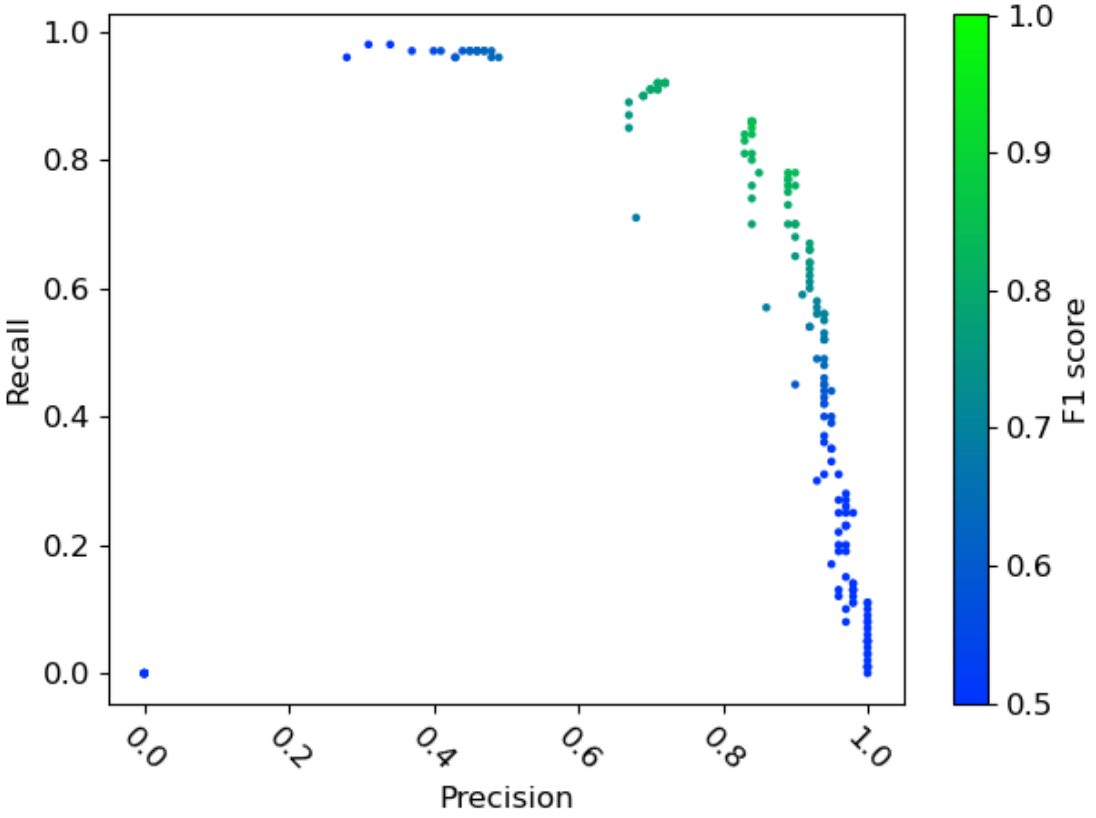
0.55000	0.10000	0.84012	0.45000	0.97000	0.62000	0.77893	-2.01740	-0.25242
0.55000	0.20000	0.84573	0.71000	0.91000	0.80000	1.07433	-1.47281	-0.18464
0.55000	0.30000	0.84426	0.83000	0.84000	0.83000	1.24893	-1.11441	-0.15604
0.55000	0.40000	0.84064	0.89000	0.70000	0.79000	1.41349	-0.78121	-0.13796
0.55000	0.50000	0.83504	0.93000	0.58000	0.71000	1.61786	-0.43850	-0.12552
0.55000	0.60000	0.83046	0.94000	0.46000	0.62000	1.96071	0.10375	-0.11279
0.55000	0.70000	0.80956	0.96000	0.31000	0.46000	2.58259	1.07231	-0.09928
0.55000	0.80000	0.79561	0.97000	0.15000	0.26000	3.44173	2.39930	-0.08994
0.55000	0.90000	0.74380	1.00000	0.07000	0.13000	6.55591	7.01438	-0.08134
0.60000	0.10000	0.83994	0.46000	0.97000	0.62000	0.77067	-2.01402	-0.25063
0.60000	0.20000	0.84676	0.71000	0.91000	0.80000	1.08154	-1.46653	-0.18440
0.60000	0.30000	0.84521	0.84000	0.84000	0.84000	1.22198	-1.13941	-0.15687
0.60000	0.40000	0.84040	0.89000	0.73000	0.80000	1.41695	-0.78421	-0.13837
0.60000	0.50000	0.83647	0.92000	0.60000	0.73000	1.59769	-0.47233	-0.12685
0.60000	0.60000	0.83123	0.94000	0.48000	0.64000	1.95578	0.08557	-0.11363
0.60000	0.70000	0.81304	0.95000	0.33000	0.49000	2.47978	0.91720	-0.10072
0.60000	0.80000	0.79575	0.95000	0.17000	0.28000	3.39240	2.32215	-0.09024
0.60000	0.90000	0.75244	1.00000	0.08000	0.14000	6.74903	7.25209	-0.08163
0.65000	0.10000	0.84141	0.46000	0.97000	0.62000	0.78409	-2.00465	-0.24911
0.65000	0.20000	0.84647	0.71000	0.92000	0.80000	1.08398	-1.46056	-0.18399
0.65000	0.30000	0.84535	0.84000	0.86000	0.85000	1.23350	-1.13364	-0.15705
0.65000	0.40000	0.84030	0.89000	0.75000	0.81000	1.41101	-0.79977	-0.13933
0.65000	0.50000	0.83761	0.92000	0.62000	0.74000	1.60680	-0.48136	-0.12801
0.65000	0.60000	0.83016	0.94000	0.49000	0.65000	1.85454	-0.06264	-0.11620
0.65000	0.70000	0.81391	0.94000	0.36000	0.52000	2.35939	0.73687	-0.10266
0.65000	0.80000	0.80244	0.96000	0.19000	0.32000	3.27611	2.12246	-0.09185
0.65000	0.90000	0.75300	1.00000	0.09000	0.17000	5.65196	5.67761	-0.08265
0.70000	0.10000	0.84091	0.46000	0.97000	0.63000	0.79054	-1.99726	-0.24756
0.70000	0.20000	0.84701	0.71000	0.92000	0.80000	1.08852	-1.46361	-0.18444
0.70000	0.30000	0.84479	0.84000	0.85000	0.85000	1.23509	-1.14171	-0.15802
0.70000	0.40000	0.84060	0.90000	0.76000	0.82000	1.40389	-0.81910	-0.14051
0.70000	0.50000	0.83829	0.92000	0.63000	0.75000	1.59448	-0.50749	-0.12899
0.70000	0.60000	0.82896	0.94000	0.52000	0.67000	1.84283	-0.09040	-0.11694
0.70000	0.70000	0.81600	0.94000	0.37000	0.53000	2.34890	0.71265	-0.10311
0.70000	0.80000	0.80336	0.96000	0.20000	0.33000	3.31458	2.16918	-0.09182
0.70000	0.90000	0.74998	1.00000	0.10000	0.18000	5.61907	5.62234	-0.08287
0.75000	0.10000	0.84104	0.47000	0.97000	0.63000	0.80772	-1.98278	-0.24531
0.75000	0.20000	0.84735	0.71000	0.92000	0.80000	1.09682	-1.45373	-0.18354
0.75000	0.30000	0.84479	0.84000	0.86000	0.85000	1.24041	-1.14343	-0.15863
0.75000	0.40000	0.84140	0.89000	0.76000	0.82000	1.40853	-0.82773	-0.14154
0.75000	0.50000	0.83950	0.92000	0.64000	0.76000	1.58324	-0.52981	-0.12994
0.75000	0.60000	0.82988	0.94000	0.52000	0.67000	1.79817	-0.15383	-0.11813

0.75000	0.70000	0.81968	0.94000	0.40000	0.56000	2.22000	0.53281	-0.10474
0.75000	0.80000	0.80740	0.96000	0.22000	0.35000	3.11512	1.87569	-0.09331
0.75000	0.90000	0.76239	1.00000	0.11000	0.19000	5.50060	5.41704	-0.08347
0.80000	0.10000	0.84116	0.47000	0.97000	0.63000	0.80617	-1.98047	-0.24490
0.80000	0.20000	0.84704	0.72000	0.92000	0.80000	1.07982	-1.46680	-0.18405
0.80000	0.30000	0.84398	0.84000	0.86000	0.85000	1.22881	-1.15309	-0.15864
0.80000	0.40000	0.84083	0.89000	0.77000	0.82000	1.38761	-0.85304	-0.14213
0.80000	0.50000	0.83949	0.92000	0.64000	0.76000	1.57350	-0.54788	-0.13063
0.80000	0.60000	0.83082	0.94000	0.53000	0.68000	1.74489	-0.23087	-0.11950
0.80000	0.70000	0.82181	0.94000	0.42000	0.58000	2.19251	0.47263	-0.10607
0.80000	0.80000	0.80554	0.97000	0.23000	0.37000	3.11270	1.86624	-0.09363
0.80000	0.90000	0.77297	1.00000	0.11000	0.19000	5.23518	5.01198	-0.08414
0.85000	0.10000	0.84075	0.48000	0.97000	0.64000	0.79637	-1.96740	-0.24231
0.85000	0.20000	0.84727	0.72000	0.92000	0.81000	1.07918	-1.46978	-0.18433
0.85000	0.30000	0.84473	0.84000	0.86000	0.85000	1.22707	-1.15862	-0.15899
0.85000	0.40000	0.84061	0.89000	0.77000	0.83000	1.38970	-0.85660	-0.14263
0.85000	0.50000	0.83993	0.92000	0.66000	0.77000	1.56609	-0.57018	-0.13185
0.85000	0.60000	0.83150	0.94000	0.55000	0.70000	1.68428	-0.31201	-0.12104
0.85000	0.70000	0.82539	0.94000	0.43000	0.59000	2.09440	0.32905	-0.10803
0.85000	0.80000	0.80793	0.97000	0.23000	0.38000	2.88115	1.52867	-0.09551
0.85000	0.90000	0.78789	0.98000	0.12000	0.21000	4.90129	4.49639	-0.08506
0.90000	0.10000	0.84045	0.48000	0.96000	0.64000	0.79605	-1.95844	-0.24028
0.90000	0.20000	0.84716	0.72000	0.92000	0.81000	1.09031	-1.45676	-0.18354
0.90000	0.30000	0.84466	0.84000	0.86000	0.85000	1.22620	-1.16068	-0.15918
0.90000	0.40000	0.84178	0.89000	0.78000	0.83000	1.37508	-0.87757	-0.14345
0.90000	0.50000	0.84115	0.92000	0.66000	0.77000	1.54420	-0.59850	-0.13239
0.90000	0.60000	0.83307	0.94000	0.56000	0.70000	1.70645	-0.29868	-0.12170
0.90000	0.70000	0.82750	0.94000	0.44000	0.60000	2.08709	0.30823	-0.10880
0.90000	0.80000	0.80607	0.97000	0.26000	0.41000	2.76429	1.35170	-0.09670
0.90000	0.90000	0.78573	0.96000	0.12000	0.21000	4.59161	4.06020	-0.08557
0.95000	0.10000	0.84147	0.49000	0.96000	0.65000	0.79696	-1.96115	-0.24013
0.95000	0.20000	0.84656	0.72000	0.92000	0.81000	1.09302	-1.45563	-0.18383
0.95000	0.30000	0.84475	0.84000	0.86000	0.85000	1.22962	-1.16369	-0.15969
0.95000	0.40000	0.84309	0.90000	0.78000	0.84000	1.38594	-0.87692	-0.14390
0.95000	0.50000	0.84083	0.92000	0.67000	0.78000	1.53558	-0.61127	-0.13287
0.95000	0.60000	0.83341	0.94000	0.56000	0.70000	1.68588	-0.33014	-0.12245
0.95000	0.70000	0.82806	0.94000	0.45000	0.61000	2.05680	0.26207	-0.10950
0.95000	0.80000	0.80875	0.97000	0.27000	0.43000	2.74231	1.30589	-0.09751
0.95000	0.90000	0.78673	0.96000	0.13000	0.22000	4.55737	3.99919	-0.08589

A.4 PLOTS IN FULL SIZE






Recall
Precision
F1 score
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
1.0
0.9
0.8
0.7
0.6
0.5

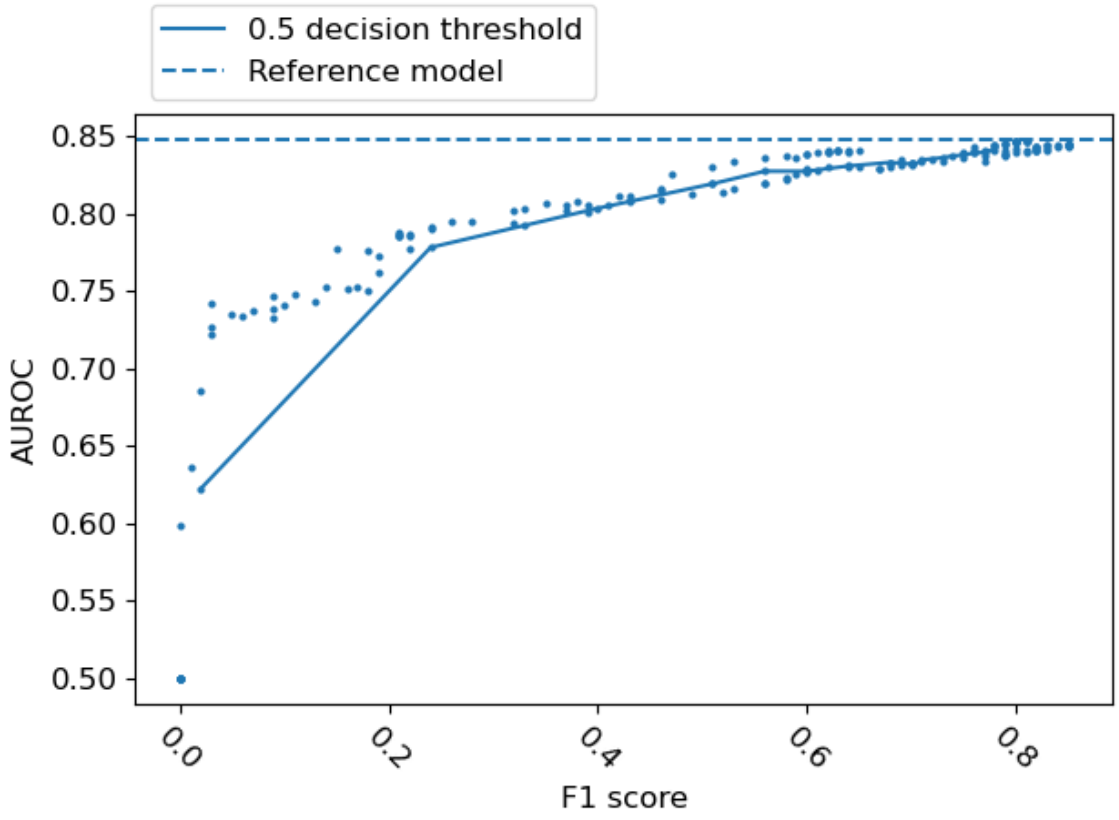


Figure A.3: Logistic regression

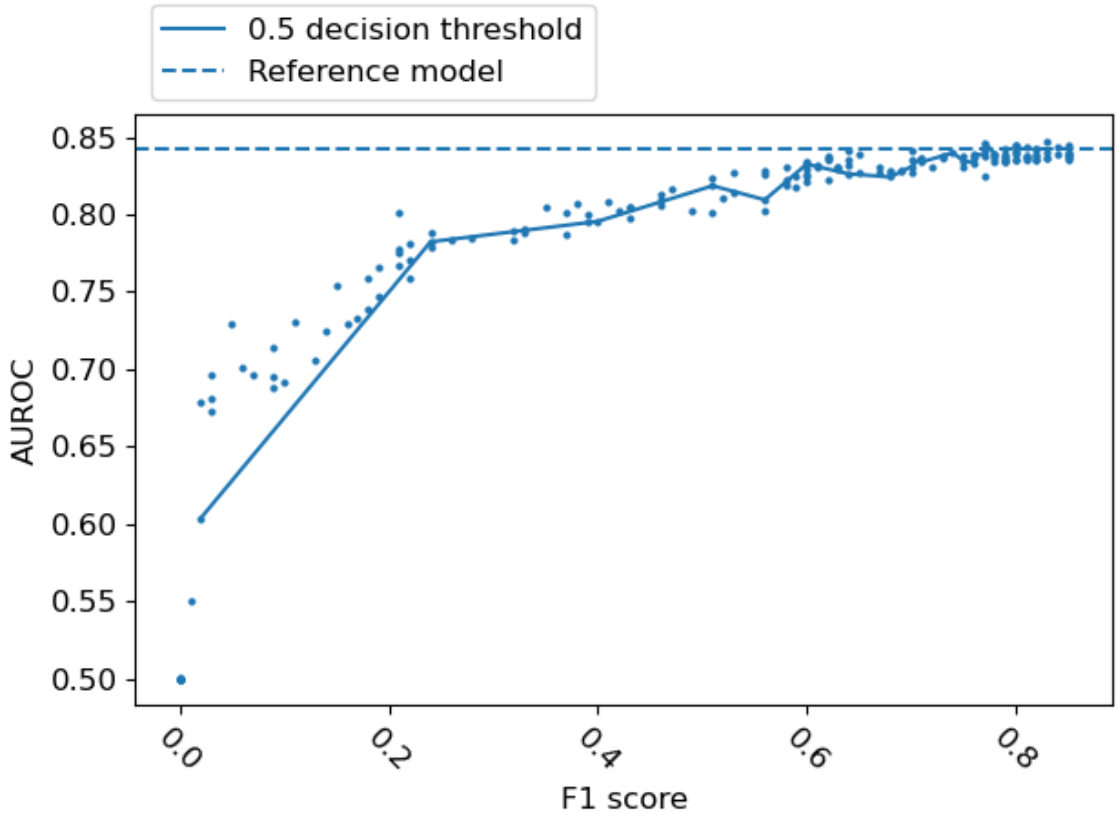


Figure A.4: FFNN

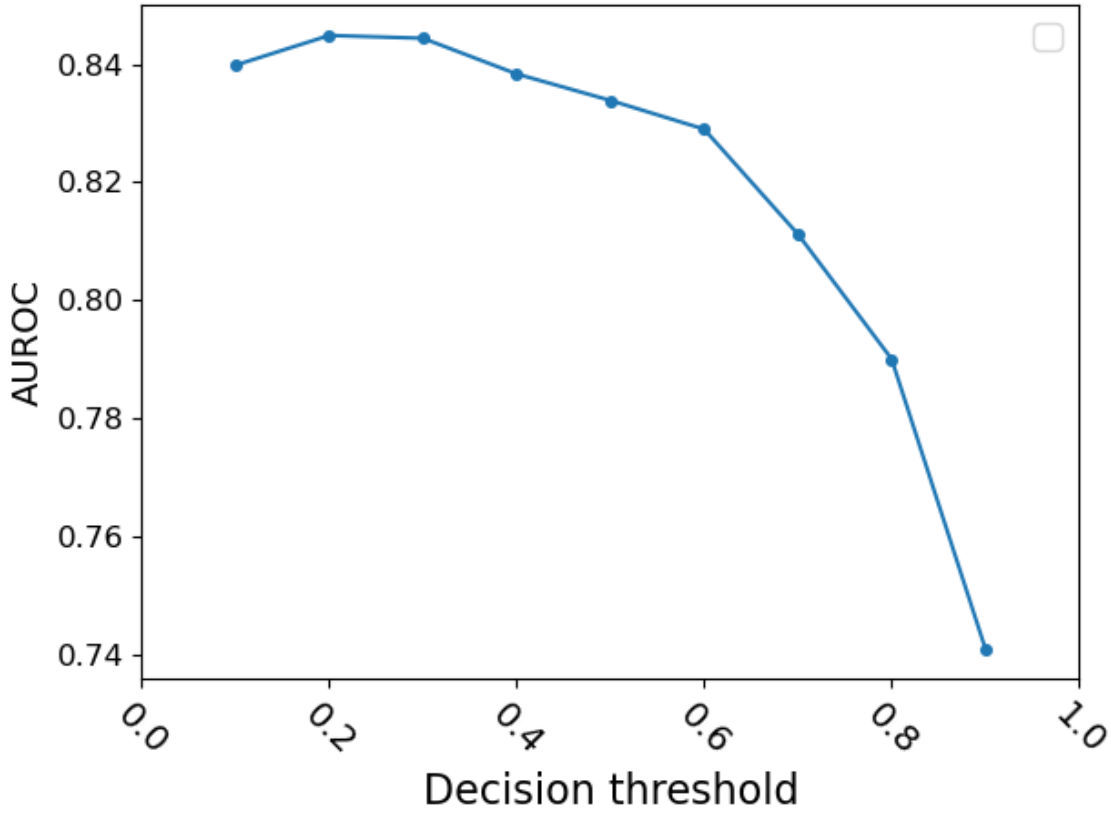


Figure A.5: Logistic regression

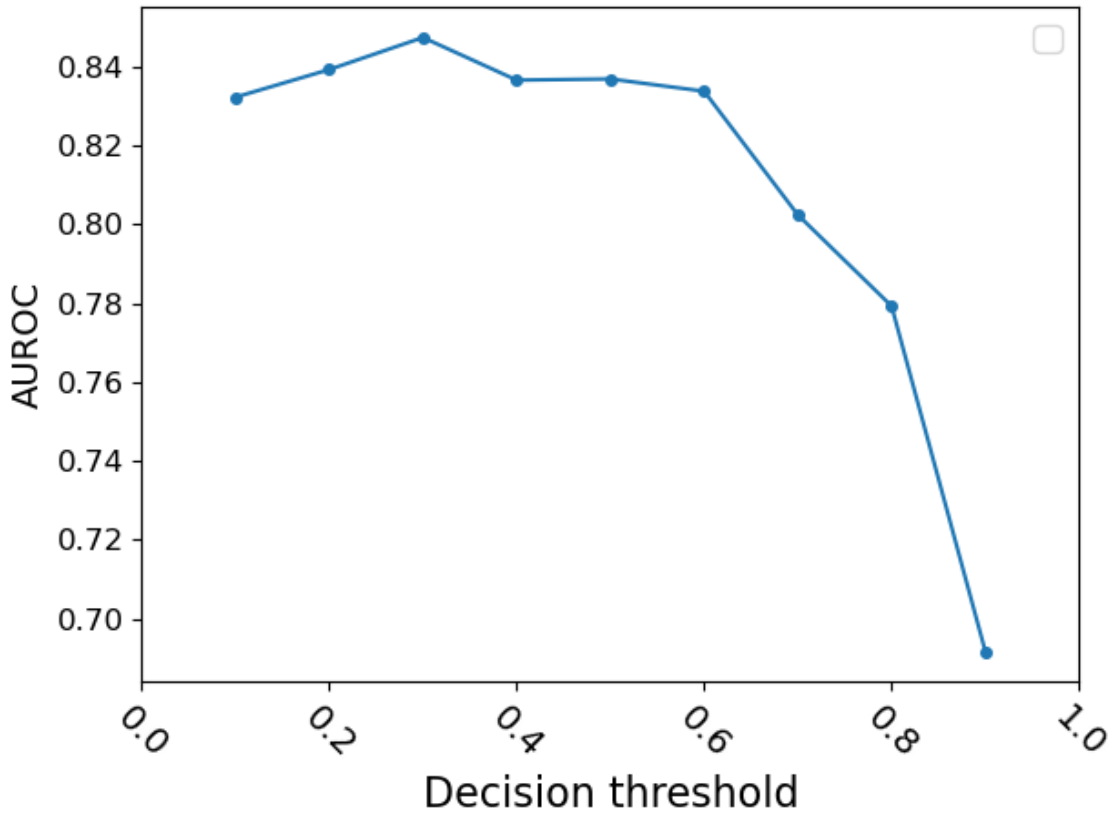


Figure A.6: FFNN

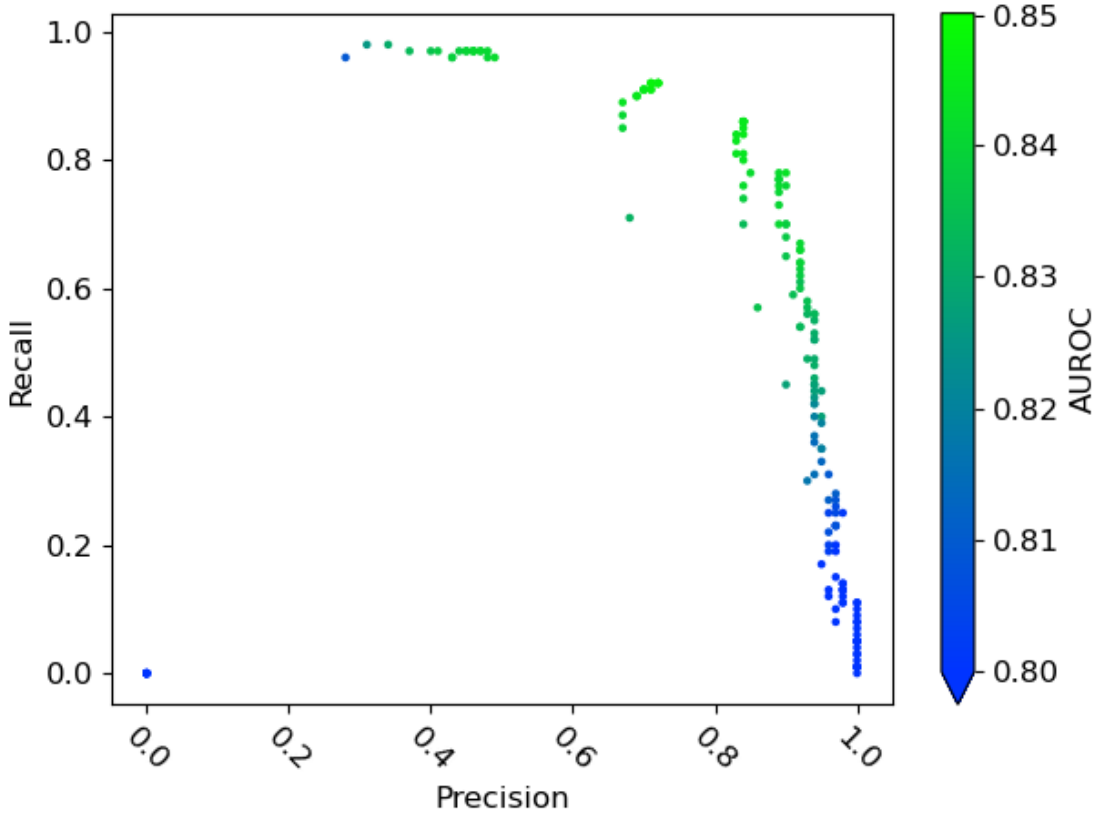


Figure A.7: Logistic regression

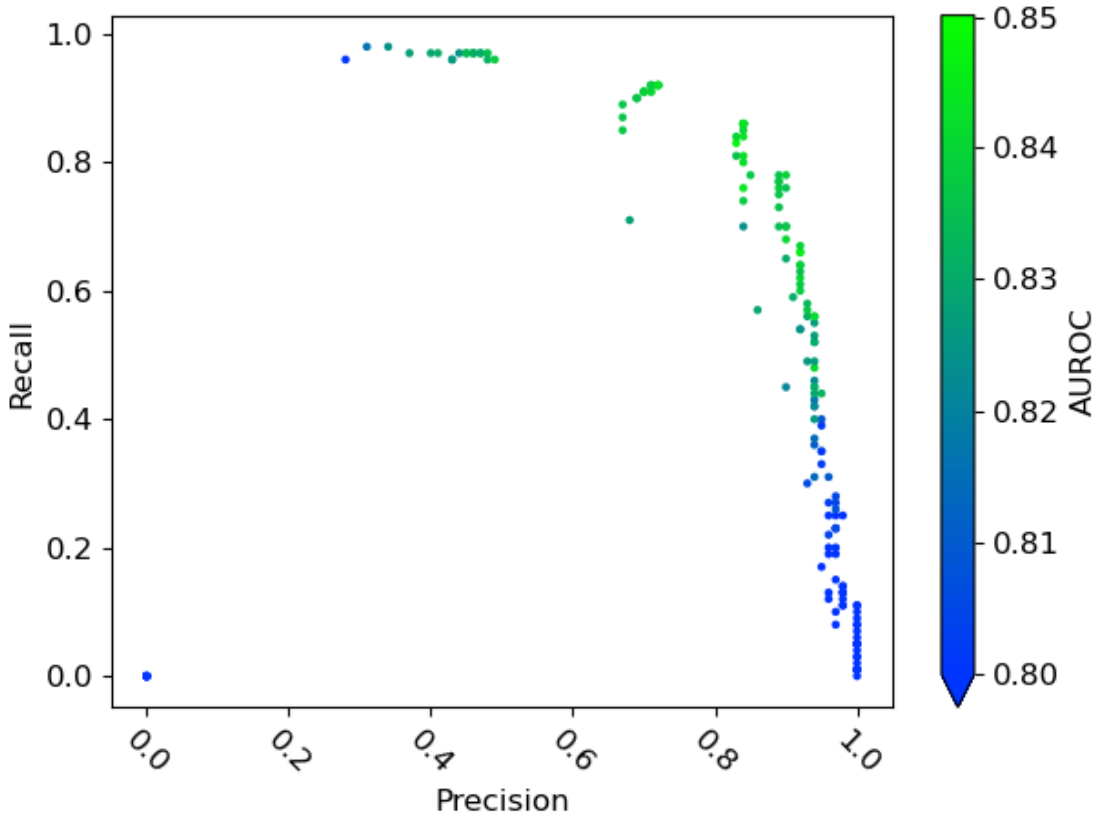


Figure A.8: FFNN

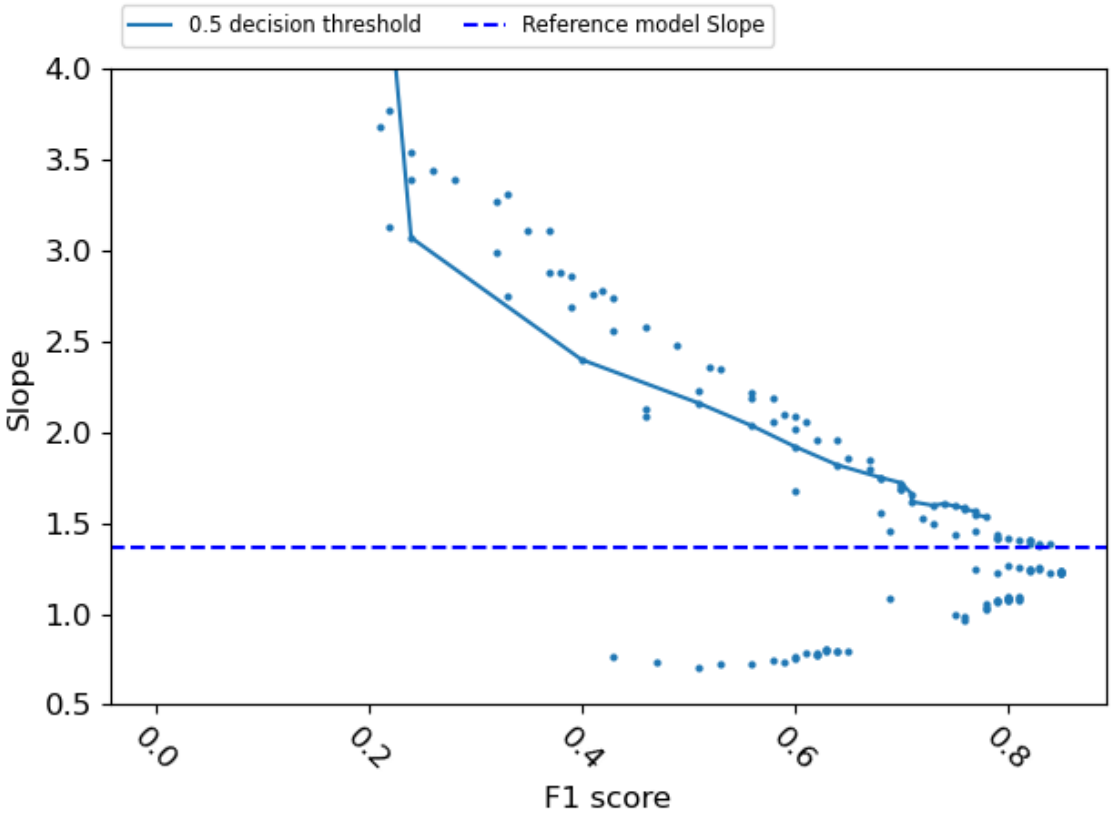


Figure A.9: Logistic regression

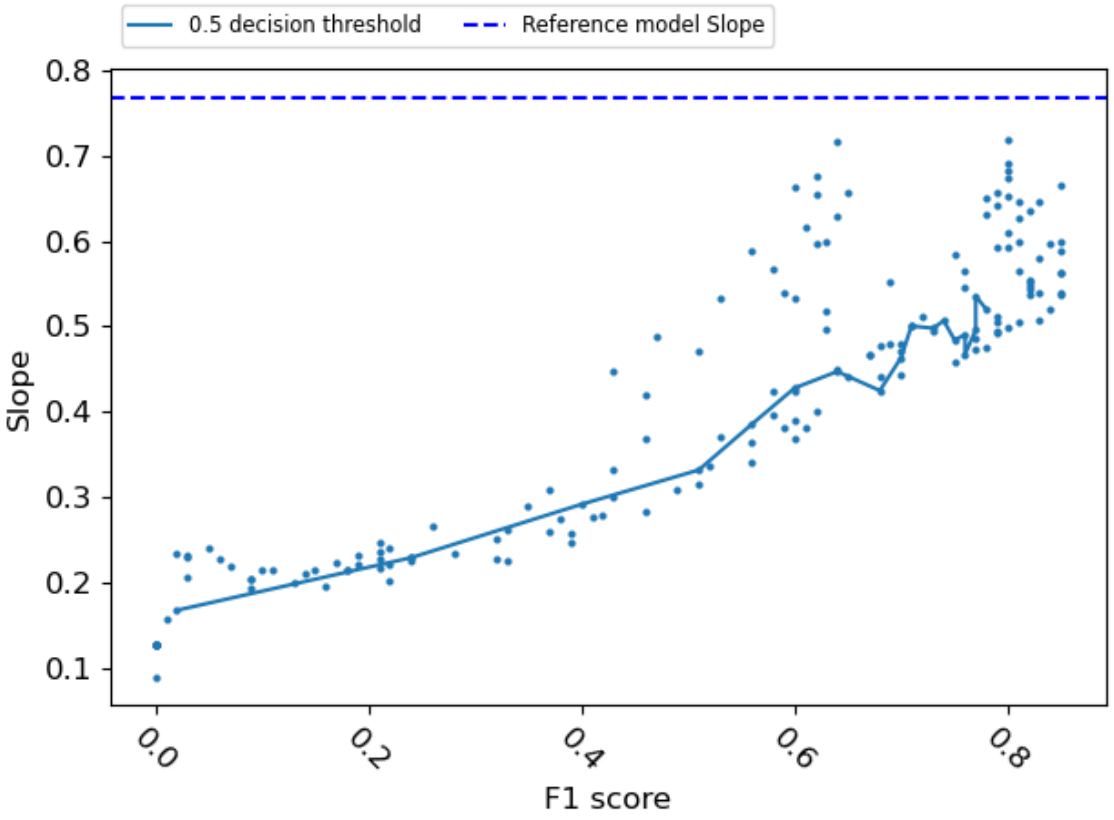


Figure A.10: FFNN

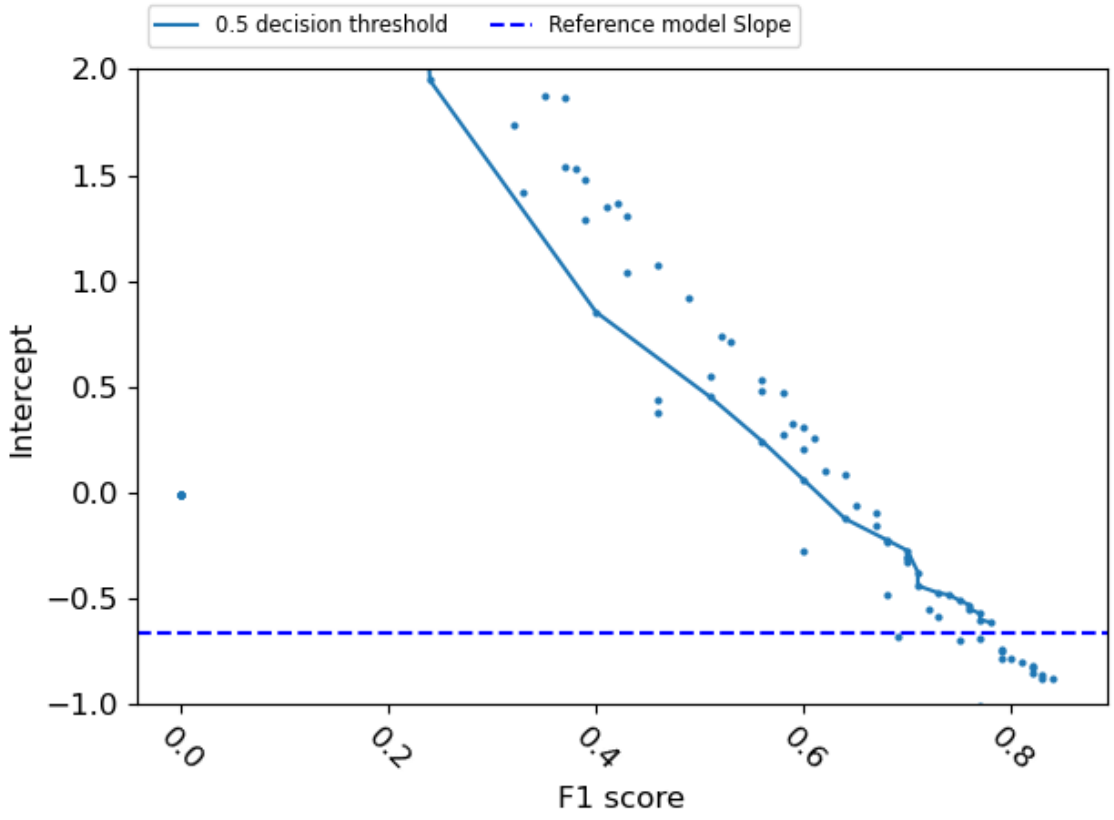


Figure A.11: Logistic regression

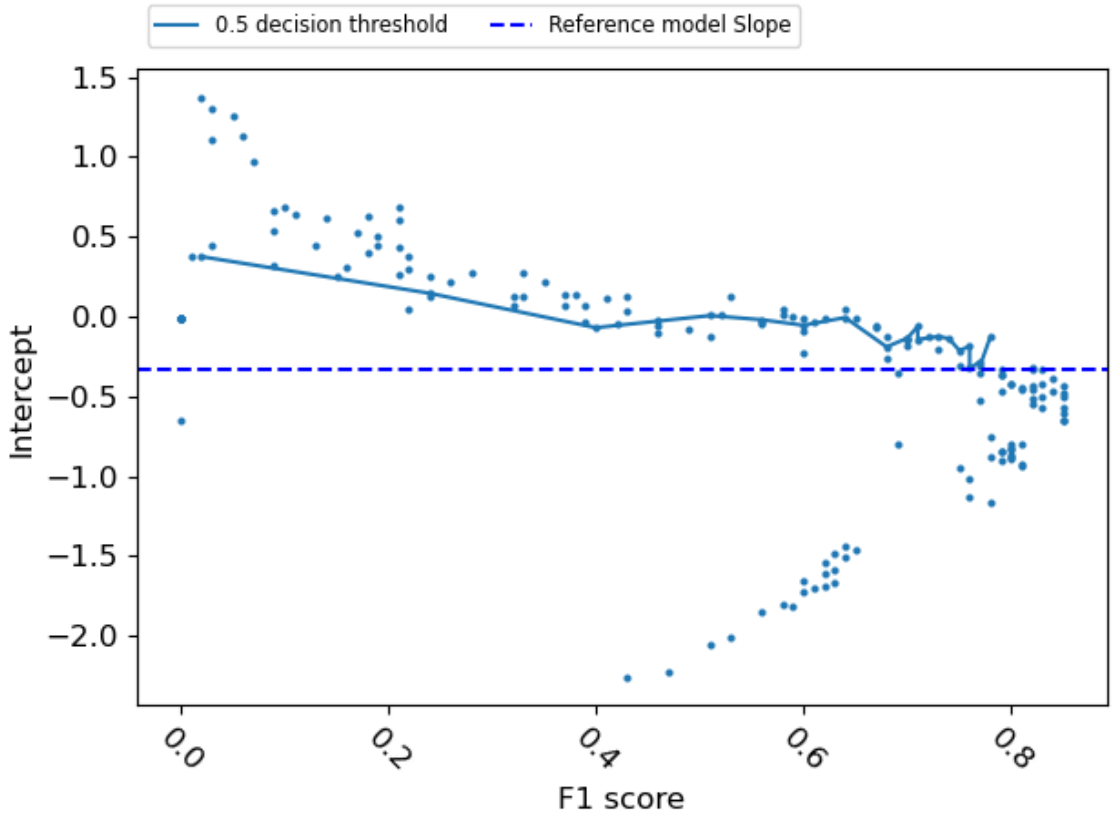


Figure A.12: FFNN

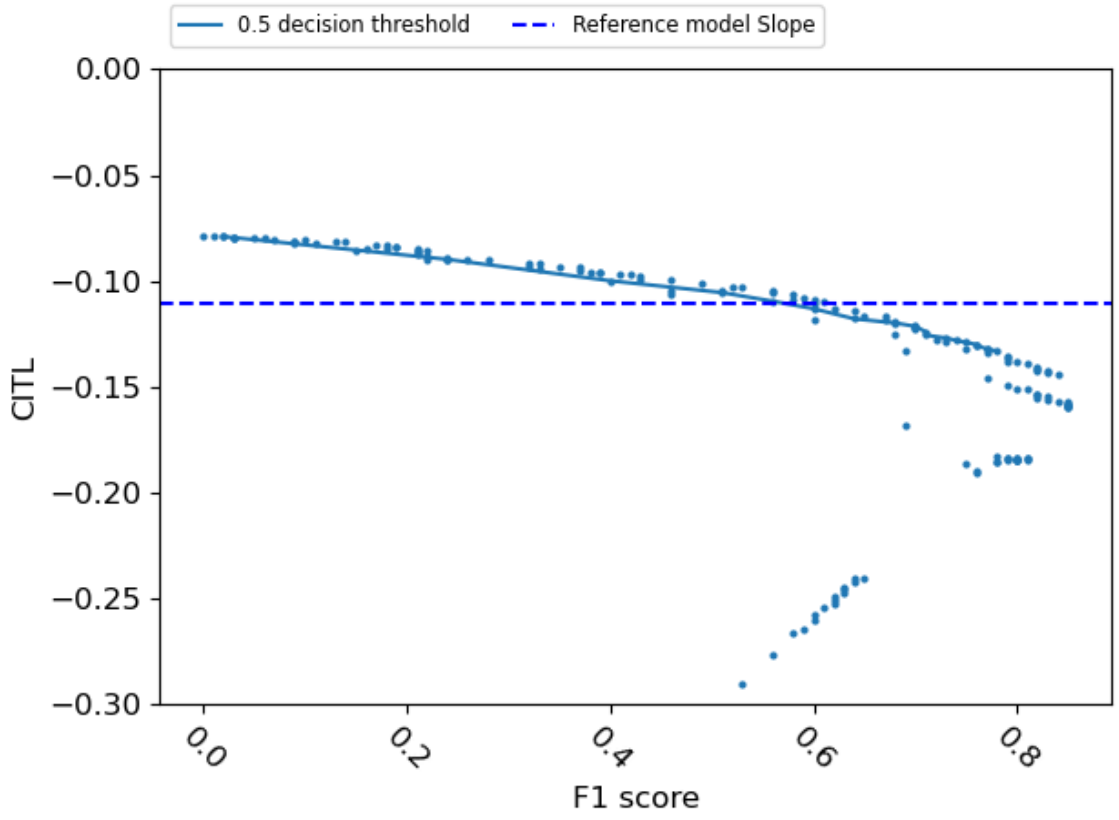


Figure A.13: Logistic regression

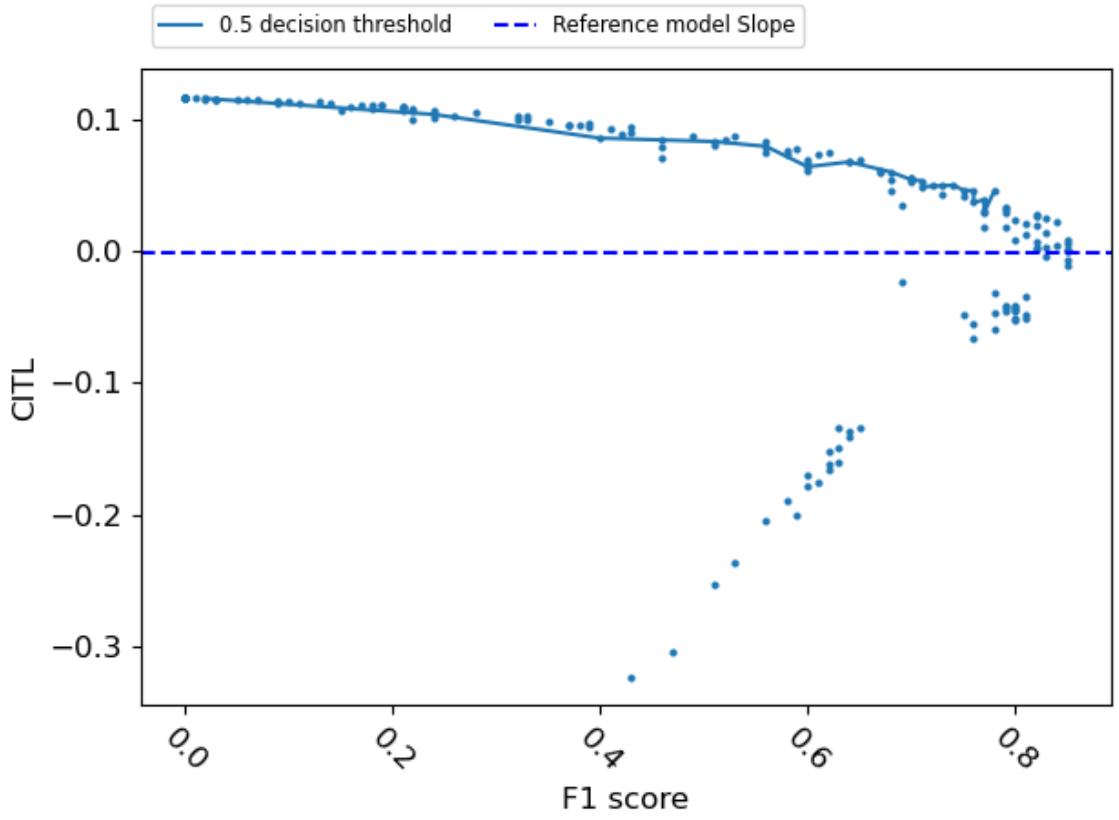


Figure A.14: FFNN

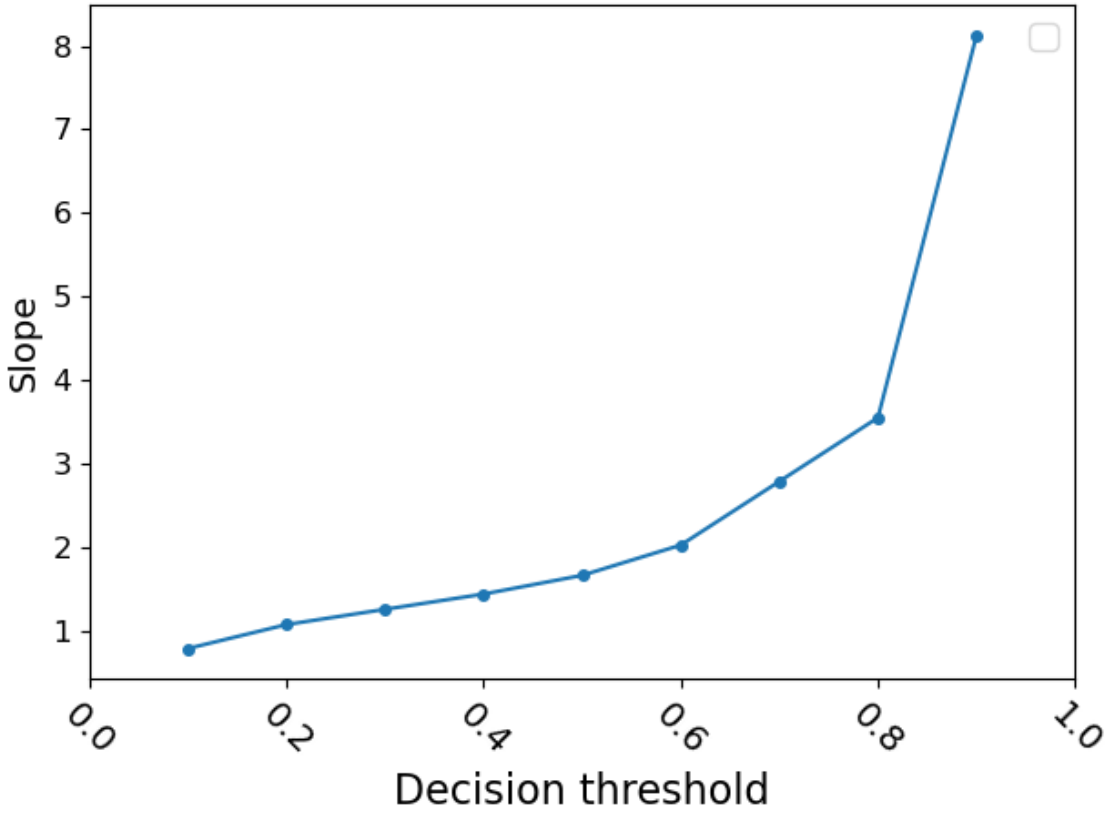


Figure A.15: Logistic regression

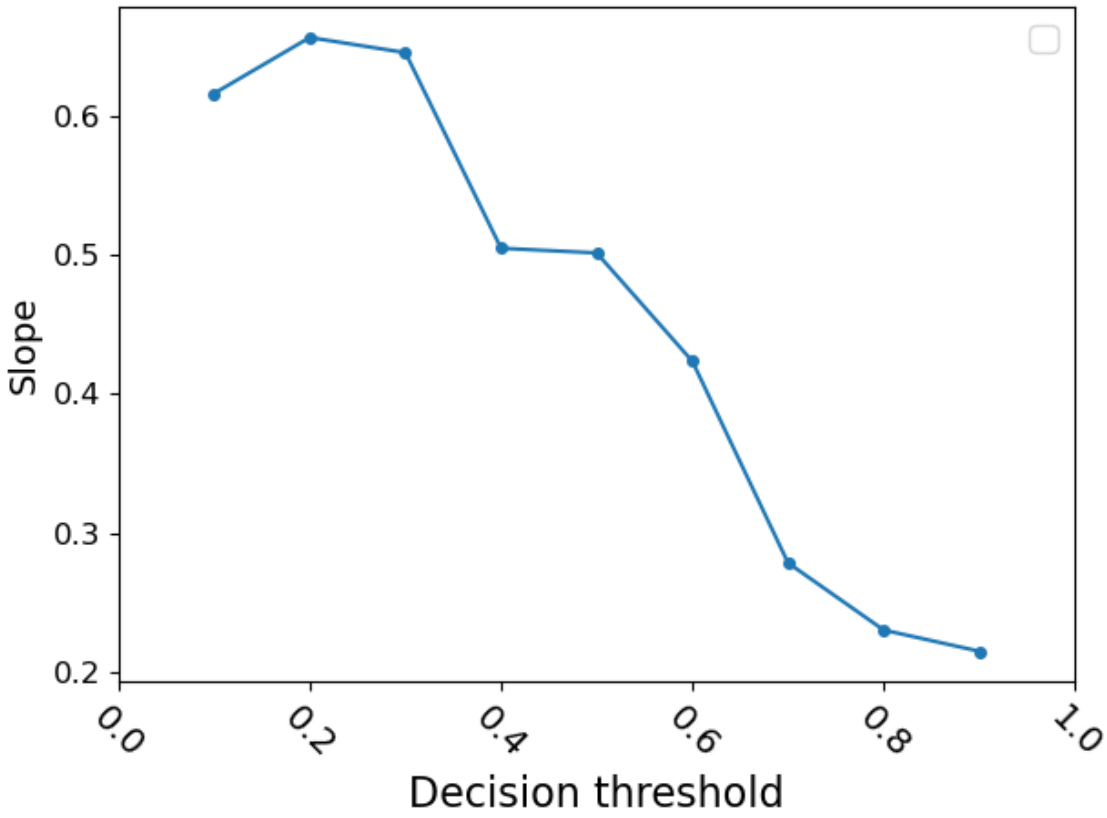


Figure A.16: FFNN

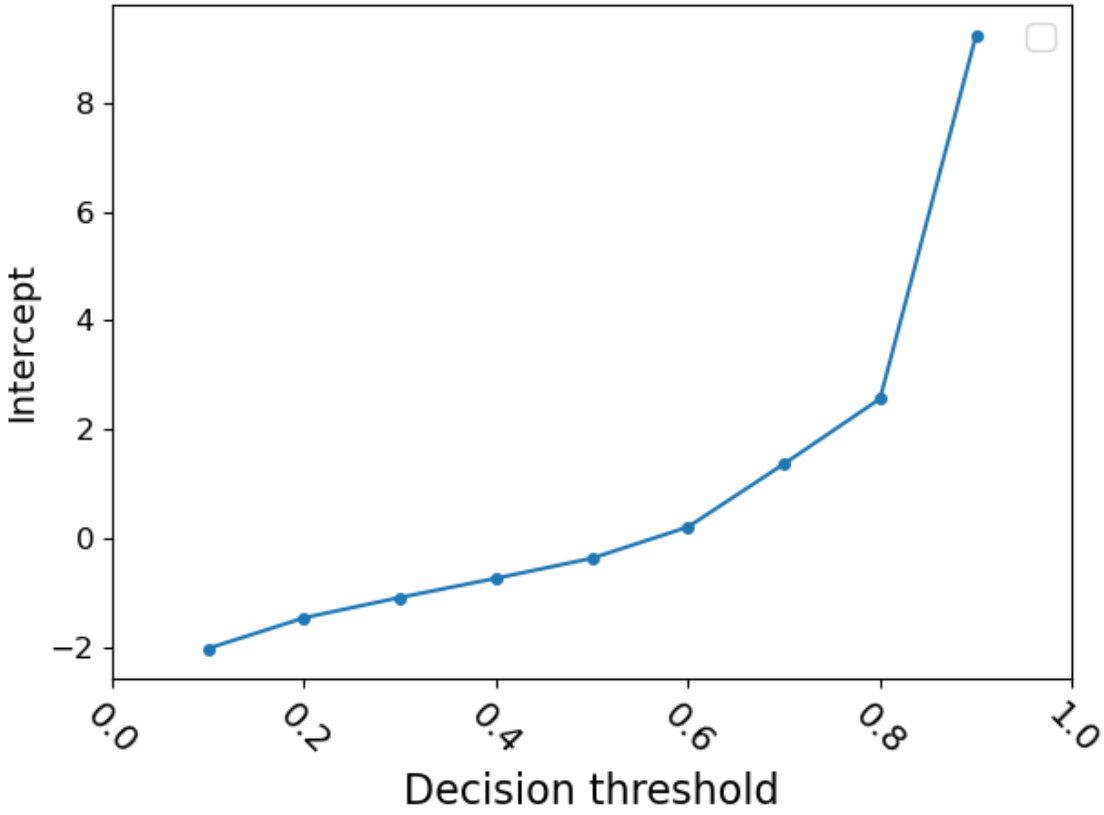


Figure A.17: Logistic regression

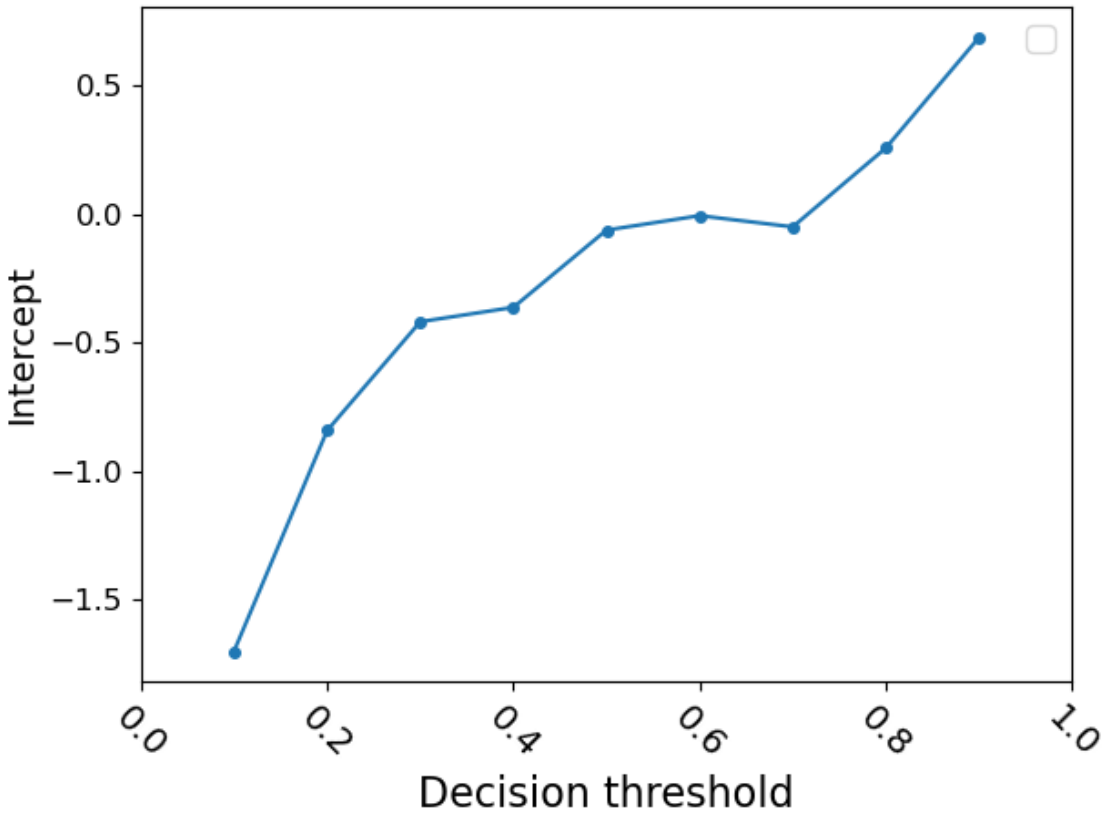


Figure A.18: FFNN

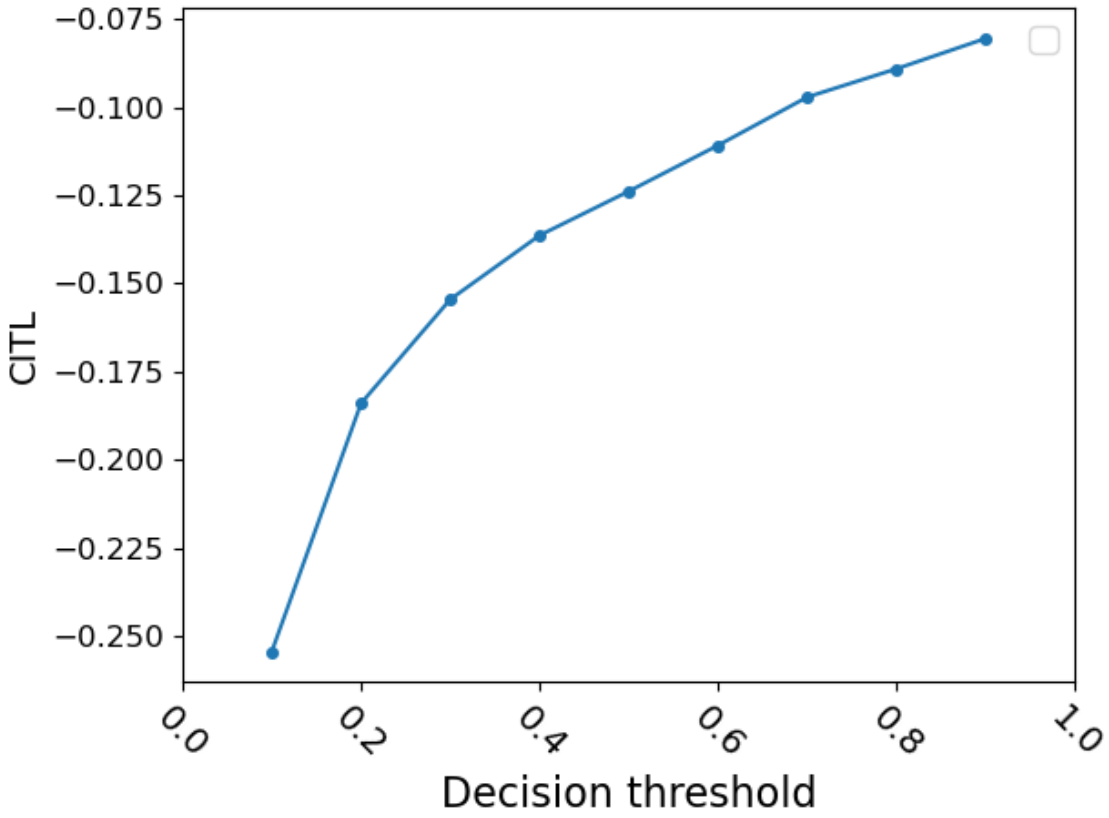


Figure A.19: Logistic regression

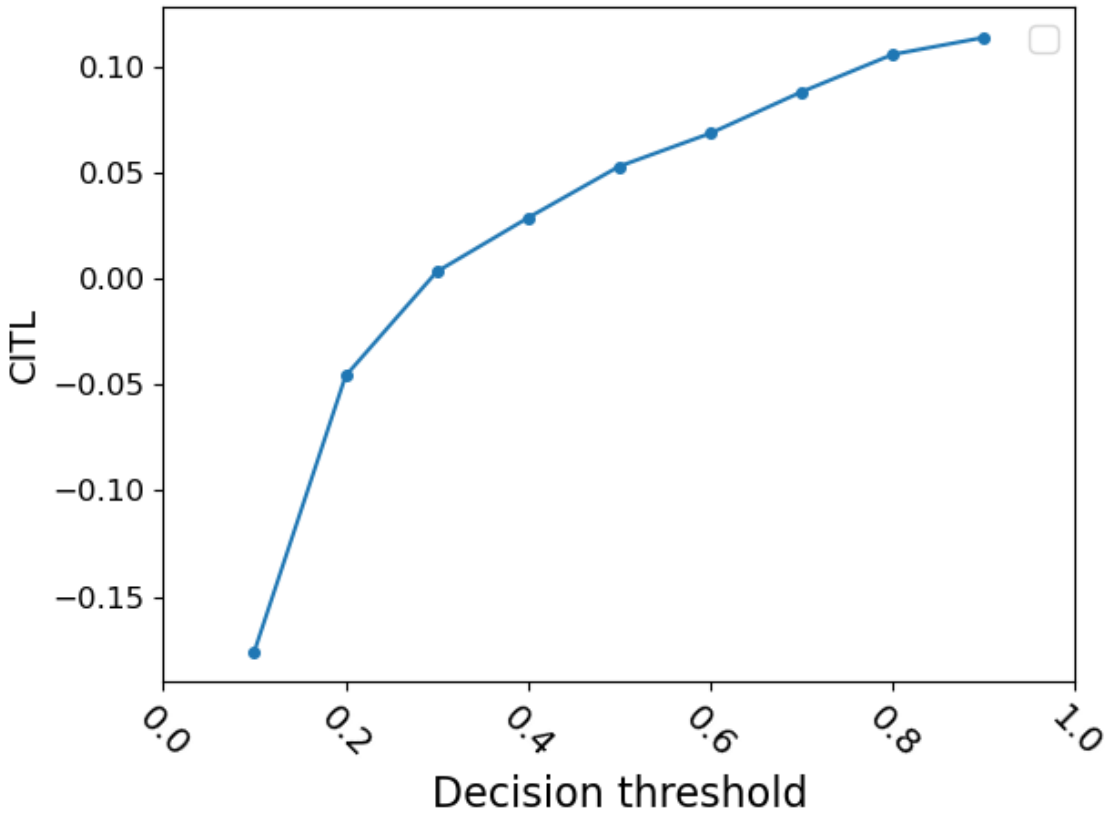


Figure A.20: FFNN

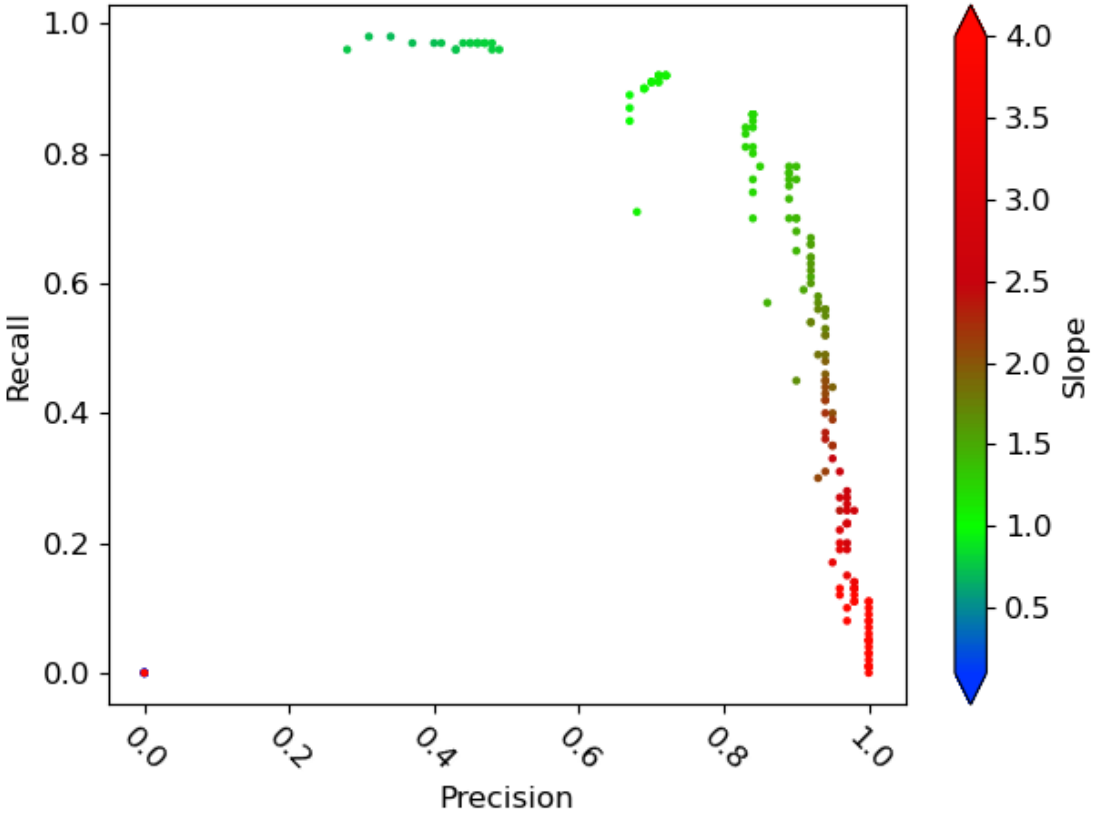


Figure A.21: Logistic regression

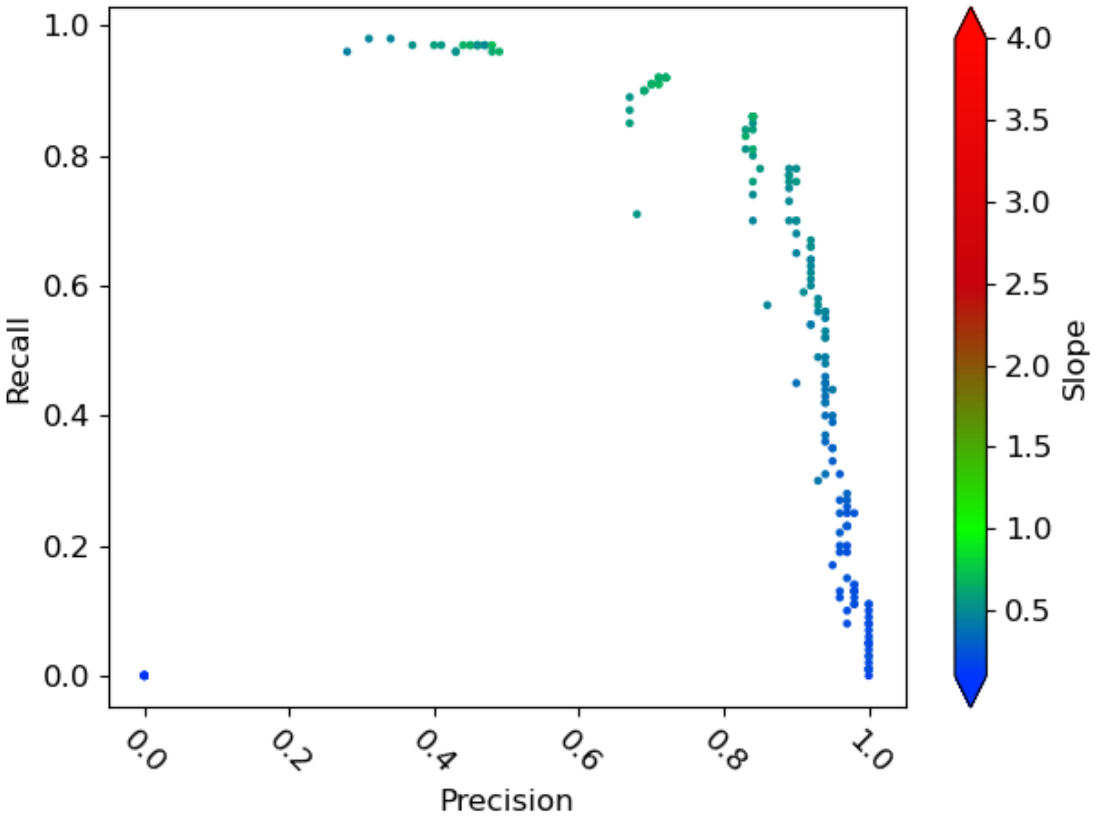


Figure A.22: FFNN

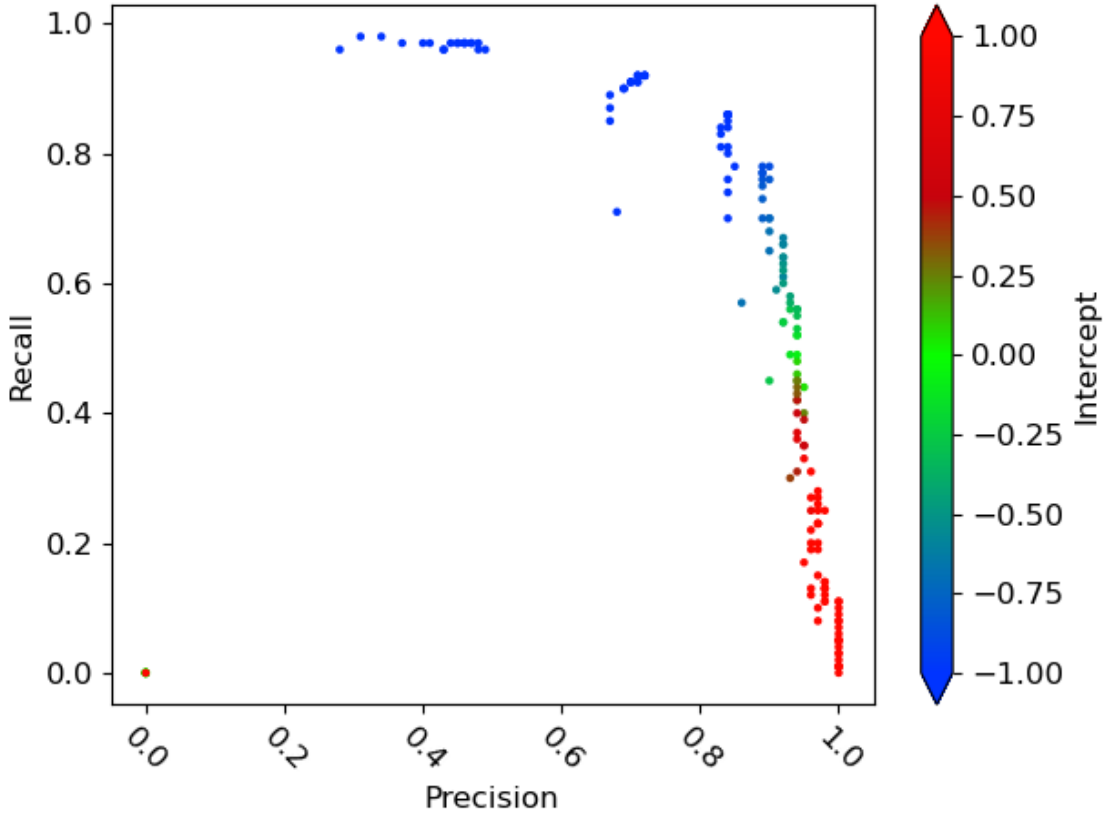


Figure A.23: Logistic regression

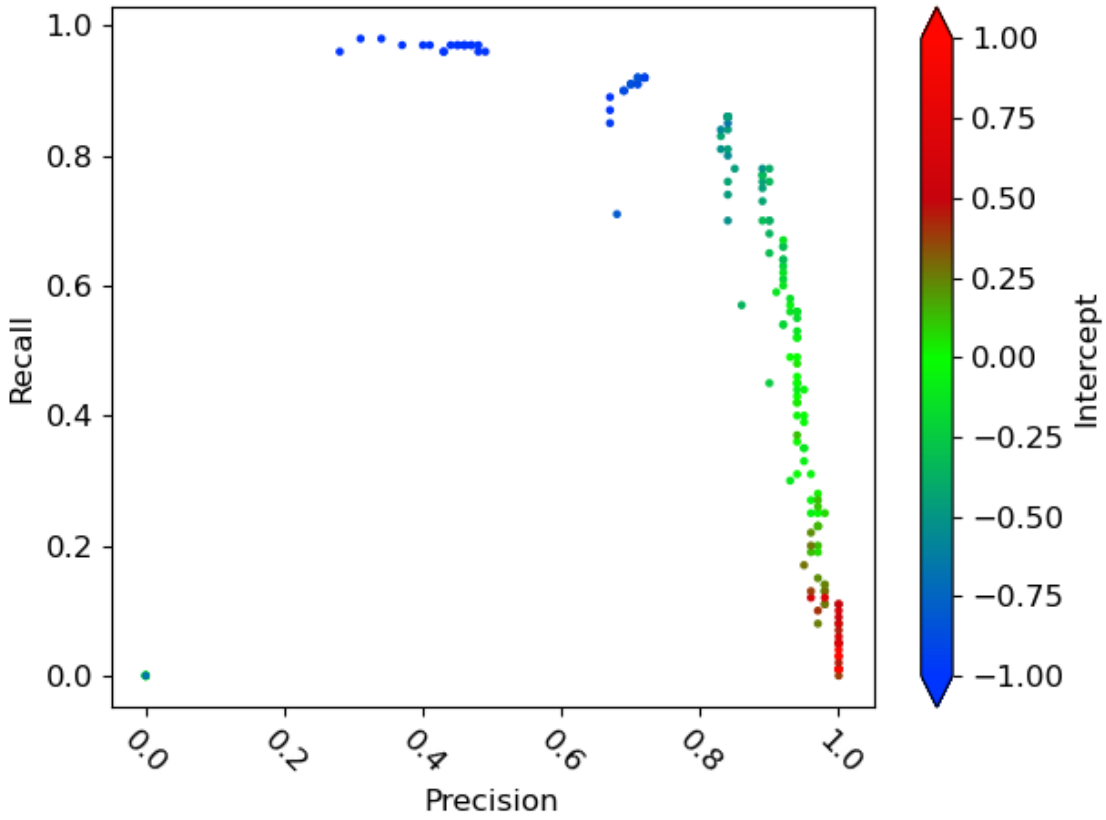


Figure A.24: FFNN

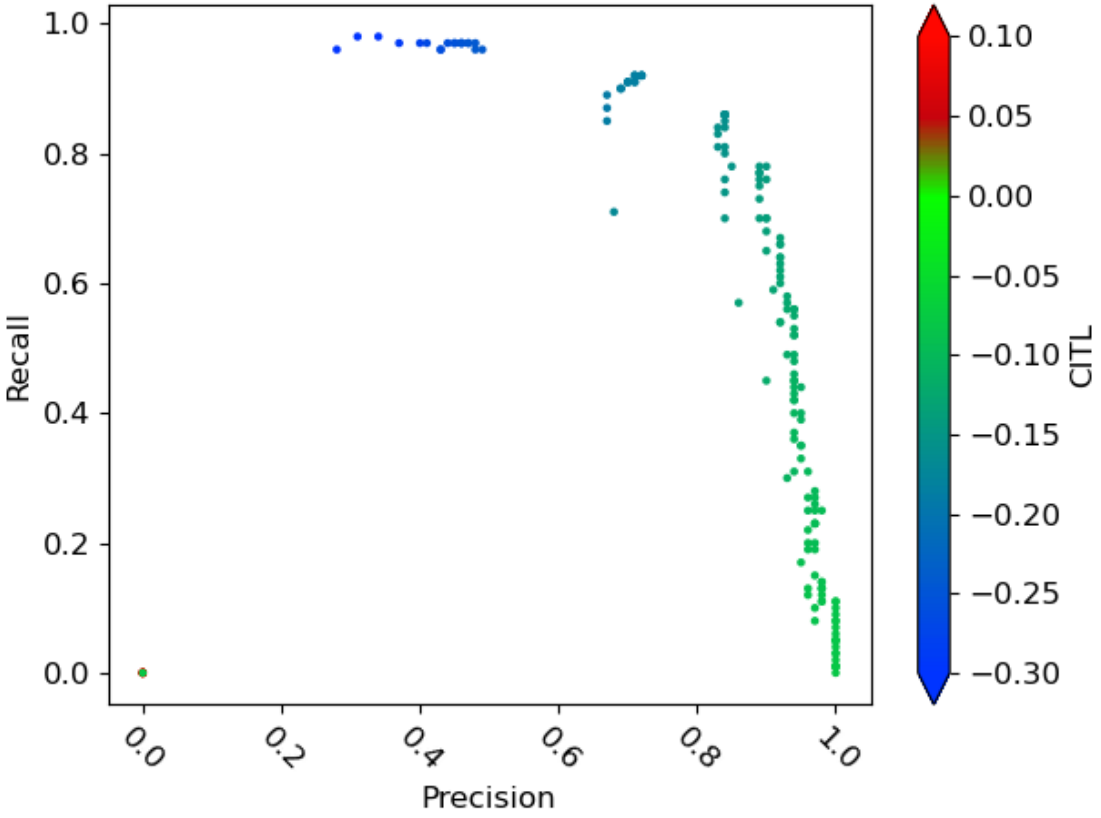


Figure A.25: Logistic regression

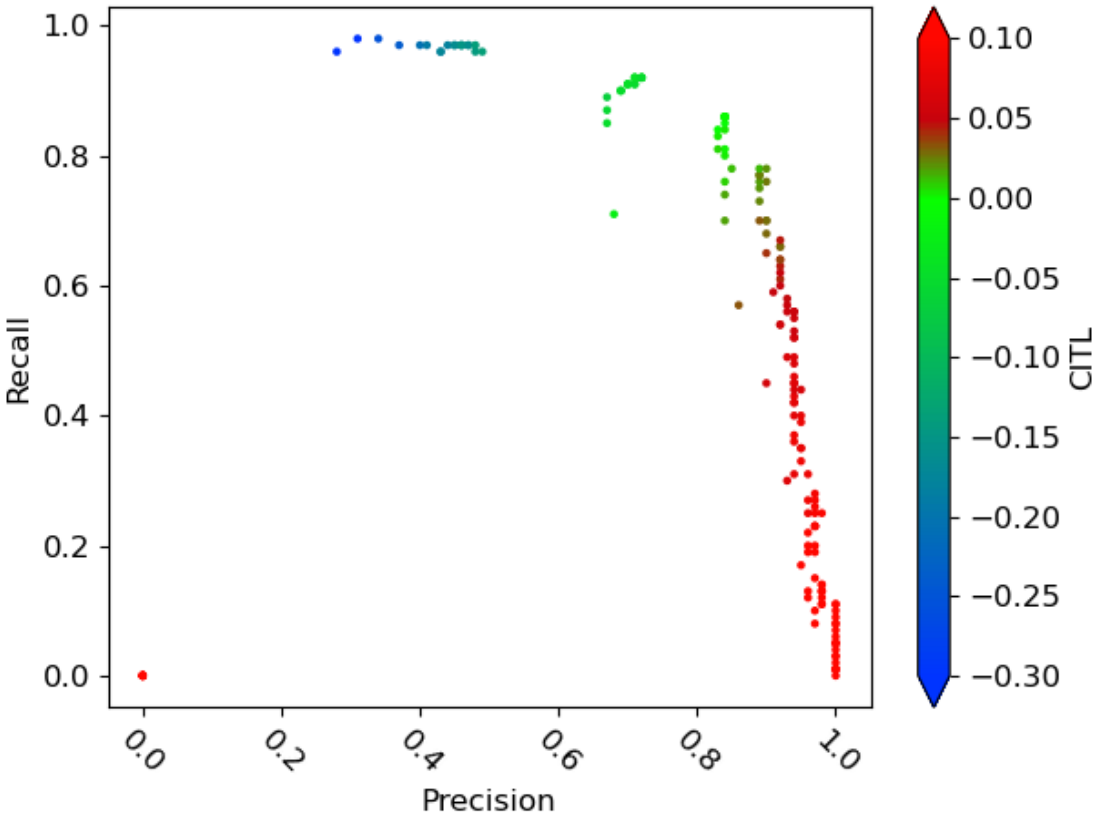


Figure A.26: FFNN