Universidade Estadual de Campinas
Instituto de Computação

Tito Barbosa Rezende

Interpretable Deep Neural Networks with Rejection Option on Tabular Data

Redes Neurais Profundas Interpretáveis com Opção de Rejeição em Dados Tabulares

CAMPINAS
2022

Tito Barbosa Rezende

Interpretable Deep Neural Networks with Rejection Option on Tabular Data

Redes Neurais Profundas Interpretáveis com Opção de Rejeição em Dados Tabulares

Dissertação apresentada ao Instituto de Computação da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Dissertation presented to the Institute of Computing of the University of Campinas in partial fulfillment of the requirements for the degree of Master in Computer Science.

Supervisor/Orientadora: Profa. Dra. Sandra Eliza Fontes de Avila
Co-supervisor/Coorientador: Dr. Luiz Sérgio Fernandes de Carvalho

Este exemplar corresponde à versão final da Dissertação defendida por Tito Barbosa Rezende e orientada pela Profa. Dra. Sandra Eliza Fontes de Avila.

CAMPINAS
2022

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

R339i

Rezende, Tito Barbosa, 1985-
Interpretable deep neural networks with rejection option on tabular data / Tito Barbosa Rezende. – Campinas, SP : [s.n.], 2022.

Orientador: Sandra Eliza Fontes de Avila.
Coorientador: Luiz Sérgio Fernandes de Carvalho.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Computação.

1. Aprendizado de máquina. 2. Redes neurais (Computação). I. Avila, Sandra Eliza Fontes de, 1982-. II. Carvalho, Luiz Sérgio Fernandes de, 1986-. III. Universidade Estadual de Campinas. Instituto de Computação. IV. Título.

Informações Complementares

Título em outro idioma: Redes neurais profundas interpretáveis com opção de rejeição em dados tabulares
Palavras-chave em inglês:
Machine learning
Neural networks (Computer science)
Área de concentração: Ciência da Computação
Titulação: Mestre em Ciência da Computação
Banca examinadora:
Sandra Eliza Fontes de Avila [Orientador]
Simone Nascimento dos Santos
Paula Dornhofer Paro Costa
Data de defesa: 01-09-2022
Programa de Pós-Graduação: Ciência da Computação

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0003-1324-2511
- Currículo Lattes do autor: http://lattes.cnpq.br/6443560284800794

Universidade Estadual de Campinas
Instituto de Computação

Tito Barbosa Rezende

Interpretable Deep Neural Networks with Rejection Option on Tabular Data

Redes Neurais Profundas Interpretáveis com Opção de Rejeição em Dados Tabulares

Comissão Examinadora:

- Profa. Dra. Sandra Eliza Fontes de Avila (Orientadora)
  Universidade Estadual de Campinas
- Profa. Dra. Paula Dornhofer Paro Costa
  Universidade Estadual de Campinas
- Dra. Simone Nascimento dos Santos
  CTCV/Hospital Brasília

A ata da defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria do Programa da Unidade.

Campinas, 01 de setembro de 2022

Acknowledgements

I would like to start my acknowledgments by mentioning my gratitude to Unicamp. My story with Unicamp began in the year 2000 in Cotuca. Later in 2003, I joined the Electrical Engineering faculty at FEEC, graduating in 2007. After 13 years away from the university, I decided to become a student again and pursue this Master's degree, mostly because of the encouragement and advice from Silvio Leite, who is a brother, a friend, and a co-worker. After learning what machine learning was and its impact on him, I joined the machine learning classes about it and met Sandra Avila, a wonderful professor who eventually became my Master's advisor. I am grateful that she guided me through this entire process. I still remember her telling me that what mattered the most was the process and not the result. She also introduced me to Dr. Luiz Sérgio Fernandes de Carvalho, who became my co-advisor. I am thankful for his time, patience, and all guidance given in this work.

I would like to extend my sincere thanks to the professors I had: Esther Colombini, André Santanchè, Paula Paro Costa, and Eduardo Valle; and to the colleagues who helped me get here: Marta Fernandes, Alceu Bissoto, Levy Gurgel, Edson Bollis, Pedro Valois, Danielle Lazarini, and Jampierre Rocha, to name a few. I am also obliged to my employers, Instituto Eldorado and Motorola, for the time off to attend classes and research.

I am profoundly thankful to my mother, Sulamita, an electrical engineer who prioritized her then toddlers, my brother and me, and gave up her Master's degree in its final stage in Unicamp. I dedicate this Master's thesis to her and to my father, Virgílio, who taught me how to program when I was only a 12-year-old kid and gave me my first paid software developer job in the family business. Lastly, I am extremely grateful to my wife, Ana, who always keeps me focused on what must be done; and to my children, Ester, João, Sara, and Maria, for having sacrificed part of the time they could have had with me.

Finally, I acknowledge how graceful God has been to me, arranging the right time, people, opportunities, and challenges to make me grow and become whom He wants me to be. This is definitely the work of His providence, as written by Paul the apostle: "For from Him and through Him and for Him are all things".

Resumo

Redes neurais profundas são uma ferramenta poderosa para o diagnóstico assistido por computador, especialmente no campo das imagens médicas. Pesquisas recentes têm procurado este mesmo sucesso com dados tabulares, um campo dominado por algoritmos baseados em árvores de decisão, como o Random Forest e o XGBoost. A TabNet é uma arquitetura de rede neural profunda que obteve resultados melhores ou pelo menos similares a estes algoritmos. No entanto, a adoção destes algoritmos avançados na medicina tem sido tardia, devido a serem "caixas pretas". Como médicos irão confiar em uma predição se eles não sabem com base em que essa predição foi feita? E como eles saberão se os modelos não estão apenas escolhendo aleatoriamente uma predição, se os tais não tem nenhuma indicação disso? Aqui, demonstramos que médicos e cientistas de dados podem obter uma luz quanto à confiabilidade das predições pela combinação de interpretabilidade com opção de rejeição. Eles também poderão saber com base em quais atributos um modelo tomou uma dada decisão. Utilizando-se de dois dataset médicos, o conhecido Framingham Heart Study (FHS) e o Myocardial ISchemIa prognostic EValuation AngioCardio-Clarity (MI-SIEVE ACC) criado por este trabalho e que contém dados reais de clínicas e hospitais do Brasil, pudemos predizer eventos cardiovasculares adversos, da sigla em inglês MACE, com melhores resultados utilizando nossa abordagem TabNet+SAT (TabNet com opção de rejeição) quando comparado aos algoritmos tradicionais como Random Forest e XGBoost, e ganhos de até 5 pontos percentuais para o FHS e 2 pontos percentuais para o MI-SIEVE ACC, utilizando opção de rejeição. Nossa abordagem possibilita aos especialistas de domínio a interpretação tanto global quanto local da predição ou rejeição, feature a feature. A capacidade intrínseca de interpretação da TabNet é usada para se dizer porque um modelo está fazendo uma determinada predição, enquanto que a opção de rejeição integrada diz se uma predição deve ser considerada confiável ou rejeitada. Finalmente, demonstramos que nossa abordagem possibilita o emprego do aprendizado profundo para a substituição de calculadoras de risco cardiovascular tradicionais.

Abstract

Deep Neural Networks are a powerful tool for computer-aided diagnosis, especially in medical imaging. Recent research is also trying to succeed with tabular data, a field dominated by tree-based algorithms such as Decision Trees, Random Forests, and XGBoost. TabNet is a deep tabular data learning architecture that has achieved at least similar results to those algorithms. However, adopting such advanced algorithms in medicine is delayed because they are "black boxes". How will medical doctors trust a prediction they do not know the basis of, and how will they know when the models are guessing if the algorithms are not telling it? Here, we show that data scientists and physicians can have a great insight into whether the prediction is reliable by combining interpretability with the rejection option. They can also know based on which features the model decision was made. Using two medical datasets, the well-known Framingham Heart Study (FHS) and our Myocardial ISchemIa prognostic EValuation AngioCardio-Clarity (MI-SIEVE ACC) real-world dataset collected from hospitals and clinics in Brazil, major adverse cardiovascular events could be predicted with higher results using our approach (TabNet with rejection option) than traditional Random Forest and XGBoost, with gains up to 5% for FHS and 2% for the MI-SIEVE ACC dataset. Our approach enables domain specialists to interpret both global and instance-based model prediction or rejection, feature-wise. The interpretability capability of TabNet is used to give light on *why* the model is making such a prediction, while the integrated rejection option gives *when* a prediction should be considered or rejected. Finally, we demonstrate that this approach could be an enabler in using deep learning to replace legacy cardiovascular risk calculators.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

1.1 Motivation

Cardiovascular diseases are the first cause of death in the world [51, 70, 73]. In 2019, according to the European Society of Cardiology (ESC), within the countries that are a member of the ESC, there were 2,2 millions of deaths among women and 1,9 million deaths among men. In Brazil, in the year of 2017, the Brazilian Society of Cardiology (SBC) [51] reported 388 thousand deaths. This corresponds to 27% of deaths in the country in a year. The SBC even created a web portal[1] to increase the awareness of the Brazilian population. The portal states that cardiovascular disease causes 1,100 deaths per day, double the deaths caused by cancer, 2.3 times more than accidents and violence, and 6.5 times all infectious diseases. For comparison, in Brazil, the highest number of deaths on a single day (8th of April 2021) due to Covid-19 was 4,249[2].

To mitigate this problem and help on medical decisions, one powerful and widely used tool is the heart disease risk calculator. Its objective is to calculate the probability or risk of one or more Major Adverse Cardiovascular Events (MACE) such as heart failure, non-fatal re-infarction, recurrent angina pain, re-hospitalization, repeat percutaneous coronary intervention, coronary artery bypass grafting, and all-cause mortality [55]. When the risk is correctly assessed, preventive and therapeutic measures can be taken to avoid or delay such deaths. This is done by following medical society guidelines establishing thresholds of those risk scores above which there is a recommendation for intervention[3].

The mostly used calculators by the medical community are TIMI (Thrombolysis In Myocardial Infarction) [6], TIMI-50 [11], and GRACE 2.0 (Global Registry of Acute Coronary Events) [26]. They are the result of statistical analysis of clinical and observational trials. TIMI risk calculator was obtained by analyzing data of 7,081 patients using multivariate logistic regression [16]. TIMI-50 was obtained by analyzing the data of 26,449 patients using Cox's method [15]. GRACE 2.0 was obtained by analyzing the data of 11,389 patients using stepwise multiple logistic analysis [59].

Although clinical and observational trials present great success, there is still tremen-

[1] https://www.cardiometro.com.br

[2] https://infoms.saude.gov.br/extensions/covid-19_html/covid-19_html.html

[3] https://www.emergenciausp.com.br/novidades-da-diretriz-de-intervencao-coronaria-percutanea-sbhci-sbc-para-o-emergencista

dous unexplored potential in medical data due to the amount of information collected daily on hospitals and clinics for legal and administrative purposes. This data can contain valuable information and scientific knowledge not covered by the clinical trials and could be used to obtain newer and better risk calculators.

In recent years, with an ever-growing computational power, the advent of learning algorithms, and public access to a huge amount of data, several advances have been achieved in different areas, such as image classification, automatic text translation, search engines, and recommendation systems. Following that trend, medical doctors and researchers have strived to make medical data available and apply machine-learning techniques to obtain models that learn with data. An example of this trend is the several medical challenges hosted on the Kaggle platform [3–5].

The motivation of this Master's thesis is to contribute to science's advancement by extracting knowledge from real data collected from clinics and hospitals in Brazil. We propose a risk calculator, in other words, a binary classifier, using machine learning techniques that represent the state of the art in several areas, such as deep neural networks. The challenge, however, is that a deep neural network is described by up to millions of different values and does not give insight into its knowledge or the way decisions are made, making it unfeasible for medical solutions. Here is where an interpretable deep tabular data learning architecture, TabNet [8], takes place.

1.2 Challenges

Deep Neural networks have demonstrated great success in several areas like image classification for cancer diagnosis [71], audio transcription [30], and text comprehension [19], outperforming conventional machine learning techniques. Nevertheless, those deep neural networks are still struggling on tabular data [34, 67] (the subject of this Master's thesis) to outperform machine learning algorithms based on decision trees such as Random Forests [12], and XGBoost [13].

In addition to the type of data, there are at least three other challenges deep neural networks need to overcome to be used with tabular data, especially in the medical context. The first is **interpretability**. To the medical community, although prediction accuracy is critical, it is paramount to interpret the outputs of those models. By interpretation, we mean that a human being that is a domain specialist can understand the cause of a decision [49]. A un interpretable algorithm is called a "black-box", alluding that human beings cannot look and understand what is going on inside the "box". A recent deep learning architecture, TabNet [8], came with the promise of providing interpretation on tabular data while delivering the same or superior performance as traditional machine learning tree ensemble algorithms.

The second challenge is treating the samples the model cannot classify properly. By samples, we mean the instances of the data, for example, a patient. The literature calls this technique as a **rejection option** [66] or **selective classification** [32]. This robustness is necessary to indicate when the model predictions present low confidence. TabNet did not have this capability, and this implementation/adaptation is challenging.

The third challenge is related to the **amount of data**. Usually, medical datasets are small, and that is also the case for the dataset used in this work: MI-SIEVE ACC (Myocardial ISchemIa prognostic EValuation AngioCardio-Clarity) (Section 4.1.1), a dataset we created for this work that consists of 9,635 samples and 207 features. It is well known that deep learning benefits from a massive amount of data [29] but struggles in small datasets.

1.3 Objectives

The objectives of this Master's thesis are:

O1. To create a risk prediction calculator for major adverse cardiovascular events (MACE) for patients that will go through a percutaneous cardiac intervention, using real medical data from the MI-SIEVE ACC dataset;

O2. To evaluate the TabNet architecture in the MI-SIEVE ACC and other datasets, identify problems, and propose improvements;

O3. To adapt TabNet to consider the rejection option and evaluate the performance on various datasets;

O4. To compare TabNet results against decision tree ensemble algorithms such as Random Forest [12] and XGBoost [13];

O5. To evaluate and compare the results of the TabNet model against the baseline calculators, TIMI and Grace 2.0, as well as its interpretability.

1.4 Research Questions

Q1. What is the impact of using the deep neural network architecture TabNet on the risk prediction of cardiovascular diseases compared with the conventional risk scores?

Q2. How to integrate a selective classifier/rejection option on TabNet? What is the impact on the results?

1.5 Contributions

The most important contributions of this Master's thesis are:

C1. A pipeline with algorithms capable of training risk calculators on tabular data (Chapter 5 Methodology);

C2. Leveraging the use of the rejection option for tabular data with deep neural networks, called TabNet+SAT Deep Learning (Chapter 5 Methodology);

C3. Introduction of the MI-SIEVE ACC dataset for its use in Machine Learning pipelines (Section 4.1.1 MI-SIEVE ACC and Section 6.3.2 MI-SIEVE ACC Experiment: All Features).

1.6 Outline

We organized the text as follows.

In Chapter **Background**, we present the most important concepts from both the medical and machine learning perspective necessary for the comprehension of the text, such as a stent, myocardial infarction, sequential attention, and TabNet.

In Chapter **Related Work**, we present the literature review, which was focused on the machine learning techniques applied to the medical field.

In Chapter **Datasets**, we describe the **MI-SIEVE ACC dataset**, its introduction and the struggles with data cleaning, preparation, and handling of missing or incorrect data, guided by the domain specialist. We also describe other datasets used to confirm our results.

In Chapter **Methodology**, we introduce the methodology, following Fayyad et al.'s [21] proposal, which includes data selections, pre-processing, transformation, modeling, and interpretation/evaluation.

In Chapter **Experiments and Results**, we show the experimental results for each dataset. Using Random Forest and XGBoost as baselines, we compare the AUC of TabNet and TabNet+SAT

In Chapter **Conclusion**, we conclude that our proposed architecture, TabNet+SAT, can be used for the creation of a cardiovascular risk prediction calculator and that its use would be beneficial for the medical community given its rejection option and interpretation capabilities.

Chapter 2

Background

In this chapter, considering the multidisciplinary nature of this Master's thesis, we seek to clarify concepts from cardiology and machine learning perspectives, highlighting the vocabulary and techniques essential for the reader's comprehension.

2.1 Cardiology

Regarding cardiology, we present the concepts of coronary arterial disease, myocardial infarction, balloon, stent, and TIMI flow.

Coronary Artery Disease (CAD) is caused by plaque buildup in the walls of the arteries that supply blood to the heart (called coronary arteries) and other parts of the body. Plaque is made up of cholesterol deposits and other substances in the artery. Plaque buildup causes the inside of the arteries to narrow over time, which can partially or block the blood flow. This process, illustrated in Figure 2.1, is called atherosclerosis.

The symptoms are chest pain (angina), discomfort in arms or shoulders, shortness of breath, weakness, or nausea. If not treated, it can cause the weakening of heart muscle or, eventually, a heart attack.

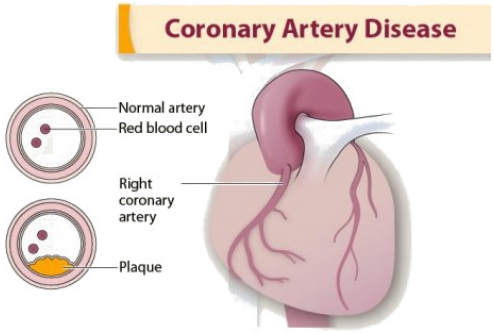
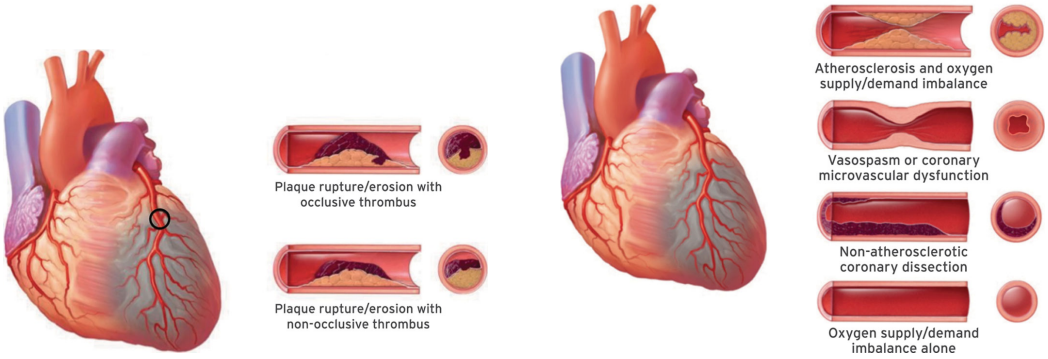


Figure 2.1: Coronary artery disease is caused by plaque buildup in the wall of the arteries that supply blood to the heart (called coronary arteries). Figure reproduced from `https://www.cdc.gov/heartdisease/coronary_ad.htm`.

Myocardial Infarction (MI) is defined as a myocardial (heart muscle) cell death due to prolonged ischemia (reduction or absence of blood irrigation) [69]. It can be classified into Type 1 and Type 2, as represented in Figure 2.2. MI Type 1 is caused by atherothrombotic CAD and is usually precipitated by atherosclerotic plaque disruption (rupture or erosion). MI Type 2 occurs when there is cell death caused by ischemia that is not related to plaque disruption, such as any imbalance between oxygen supply and demand.



(a) Myocardial Infarction Type 1 (b) Myocardial Infarction Type 2

Figure 2.2: Myocardial Infarction Types 1 and 2 illustrations. Figure reproduced from `https://www.ahajournals.org/doi/10.1161/CIR.0000000000000617`.

Balloon catheter is a special catheter used on percutaneous coronary intervention (PCI), which, through its inflation, re-establishes the blood flow of a blocked coronary artery. When necessary, the balloon is also used to deploy a stent, as illustrated in Figure 2.3.

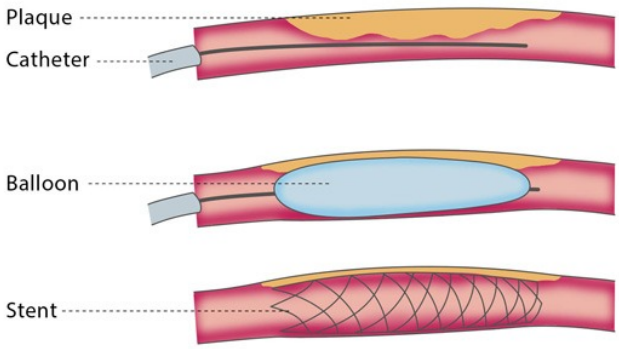


Figure 2.3: Angioplasty illustration showing a plaque restricting the blood flow, a balloon catheter, and the stent. Figure reproduced from `https://www.cirse.org/patients/ir-procedures/angioplasty-and-stenting/`.

Stent is an expansible metallic tube inserted in a coronary artery at the blocked location to prevent its collapse and blood flow restriction. It is widely used in PCI as a CAD or heart attack treatment. Figure 2.3 illustrates the deployment of a stent.

Thrombolysis in Myocardial Infarction (TIMI) Flow Perfusion of the myocardium can be categorized using the TIMI myocardial perfusion (TMP) classification system [7]. In TMP grade 3, there is the normal diffuse ground glass appearance of myocardial blush. The dye is only mildly persistent or gone at the end of the washout phase. The washout phase is the time after the end of dye injection, during which dye would typically be expected to clear from the epicardial vessels during opacification of the myocardium, followed by clearing from the myocardium. In TMP grade 2, dye enters the myocardium but accumulates and exits more slowly so that at the end of the washout phase dye in the myocardium is strongly persistent; however, dye clears by the next injection. In TMP grade 1, the dye does not leave the myocardium, and there is a stain on the next injection. In TMP grade 0, the dye does not enter the myocardium, and minimal or no blush apparent during the injection and washout phases. Static pictures of the TMP grades are shown in Figure 2.4.

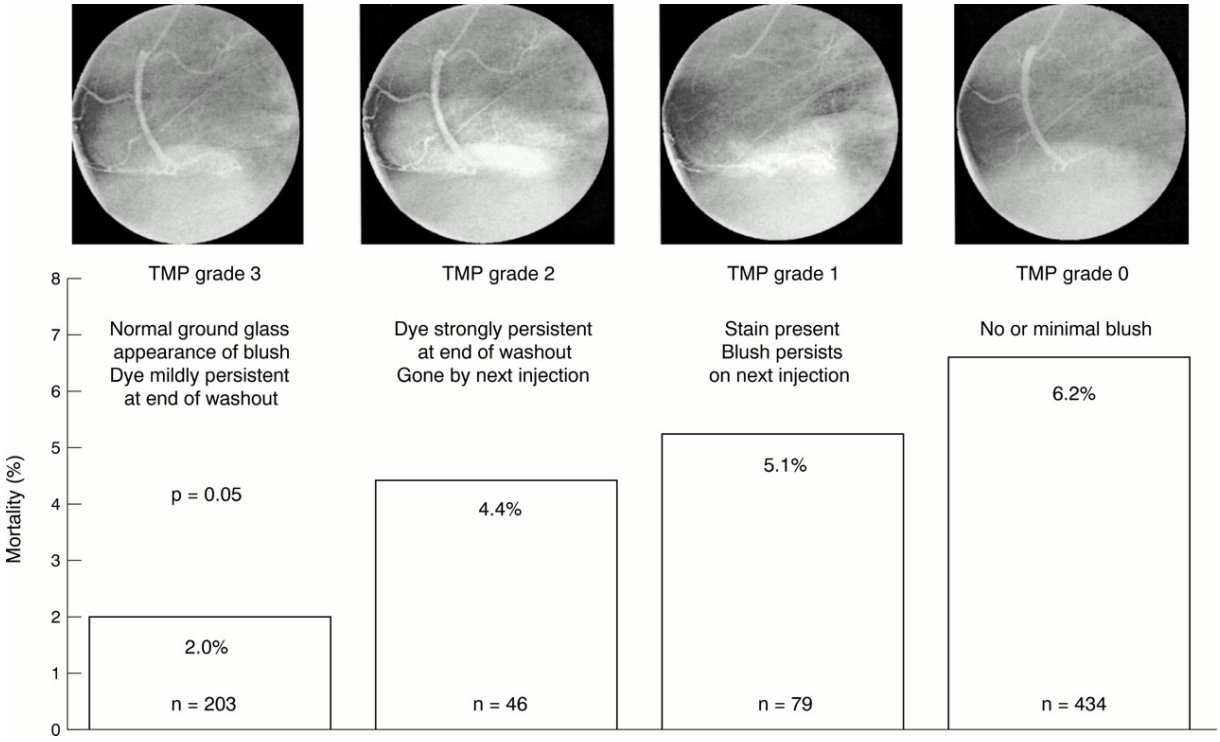


Figure 2.4: TIMI myocardial perfusion (TMP) grades. Figure reproduced from `https://heart.bmj.com/content/86/5/485`.

2.2 Machine Learning

Regarding machine learning concepts, we present here the TabNet architecture and the concepts related to the rejection option.

2.2.1 TabNet Architecture

TabNet is a high-performance interpretable canonical deep tabular data learning architecture [8]. It can be used as a binary classifier, predicting whether patience is likely to suffer

a MACE event or not. As a canonical deep neural network, it also outputs the probability of the predicted classes, thus working as a risk calculator when trained to predict the same endpoints as cardiovascular risk calculators do, in our case, MACE events. Its main contributions are local and global built-in interpretability, performance comparable with the best ML algorithm, support for unsupervised pre-training, and built-in feature selection and engineering. This combination of built-in interpretability and built-in feature selection allied with performance comparable with state-of-the-art ML algorithms such as XGBoost is why we chose TabNet.

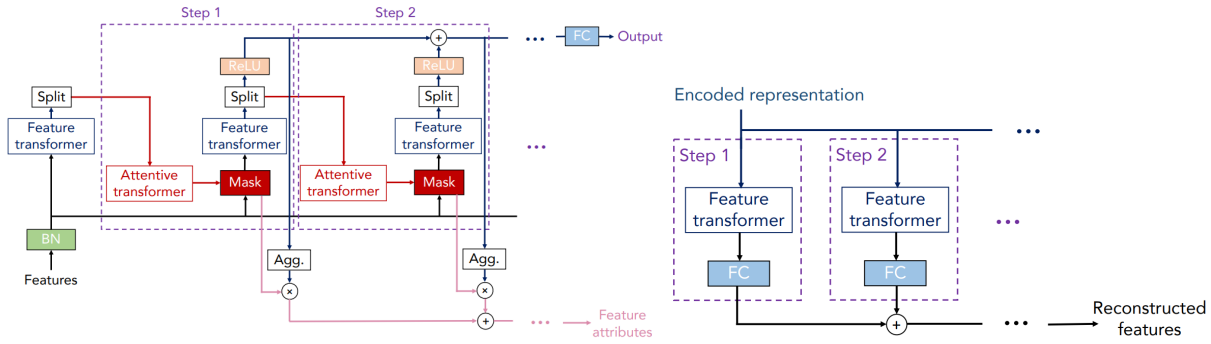
ML models can be interpreted post-hoc, with advanced techniques like LIME [60] and SHAP [45]. Those techniques consist of showing which features contributed to the predictions. Although this is quite helpful, the model is already trained, and the interpretability can reveal that the model is considering many irrelevant features, which are not the most important, but certainly influence the prediction. That's when feature selection comes into place to eliminate irrelevant features that are antagonists of the model's generalization capability. But the problem is that feature selection must be made before training, and those techniques are applied after training. An engineer needs to do this for several rounds to fine-tune the model. By achieving both selection and interpretation during training, TabNet can by itself learn which features to select and also explain the prediction by showing how much attention is given to a feature during the decision process. We will now see how this is achieved by inspecting TabNet's architecture.

Figure 2.5 shows the main components of the network. A custom number of "steps" forms the encoder. Each decision step receives the same D dimensional features $f \in^{B \times D}$, where B is the batch size. The encoding is based on sequential multi-step processing with N_{steps} decision step. The i^{th} step inputs the processed information from the $(i-1)^{th}$ step to decide which feature to use and outputs the processed feature representation to be aggregated into the overall decision. The inner blocks of a decision step are the Attentive transformer, the Mask, and the Feature transformer. The decoder comprises a feature transformer and a fully-connected layer at each step. Its output is aggregated to form the reconstructed features. The decoder is only used in the self-supervised mode.

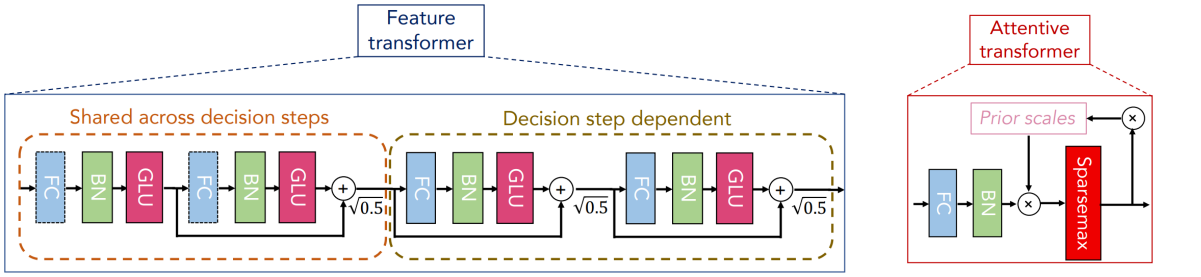
The Feature transformer is a four-layer network, where the first two layers are shared among all decision steps, and the other two belong exclusively to their step. A fully-connected layer forms each layer, with batch normalization and the gated linear units (GLU) activation function [17]. In between layers, there are skip connections resembling ResNets [31]. The Attentive transformer block is formed by only one layer. This block is responsible for TabNet's attention mechanisms. Finally, the output of the Feature transformer feeds the network output and also the next layer Attentive transformer. The classification occurs on a fully-connected layer from the aggregation of each step output.

TabNet does not require feature engineering. It is learned from the data by the network itself in the Feature transformer. The network on the Attentive transformer also learns the feature selection using sequential attention on each decision step. Learning happens sequentially, contrary to many DNN architectures with numerous hidden layers.

Figure 2.6 exemplifies the feature selection process on the well-known **Adult Census** [20] dataset, that was chosen by the authors to explain TabNet's interpretability. All features are fed to all steps, but only a few are selected. This also helps with interpretation



(a) Encoder architecture. (b) Decoder architecture.



(c) Feature transformer. (d) Attention components.

Figure 2.5: TabNet architecture. Figure reproduced from Arik et al. [8].

because we can now know what is being learned at each step. Here, for illustration, professional occupation-related features were selected on the first step, and investment related on the second. This eliminates the burden of feature selection from the data scientists, and the network learns how to make its own feature selection. Irrelevant features are excluded and not taken into consideration.

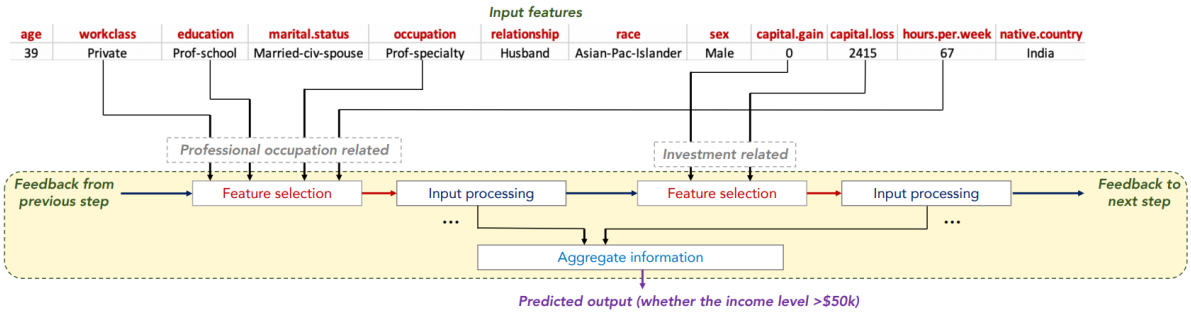


Figure 2.6: Sparse feature selection exemplified for the Adult Census dataset. Figure reproduced from Arik et al. [8].

TabNet also has a self-supervised mode through its encoder-decoder architecture. The learning task employed is an arbitrary mask that hides feature values that the network from the learned representations must reconstruct. Considering a binary mask $S \in \{0,1\}^{B \times D}$ with B batch samples and D feature dimension, the TabNet encoder inputs $(1 - S) \cdot f$ and the decoder outputs the reconstructed features $S \cdot f$, being f the original feature vector and $\hat{f}$ the reconstructed feature vector. The reconstruction loss in

the self-supervised phase is:

$$\mathcal{L} = \sum_{b=1}^{B}\sum_{j=1}^{D}\left[\frac{(\hat{f}_{b,j} - f_{b,j}) \cdot S_{b,j}}{\sqrt{\sum_{b=1}^{B}(f_{b,j} - \frac{1}{B}\sum_{b=1}^{B} f_{b,j})^2}}\right]^2 . \tag{2.1}$$

The self-supervised learning can be used as a pre-trained step to improve TabNet's performance on a *posteriori* supervised step or to aid in training with data sets with reduced samples. Self-supervised mode is illustrated in Figure 2.7. A random mask is applied to the input table for the training pass, marked with the symbol "?". After passing through the encoder and then the decoder, the masked values are predicted. A loss is computed because we know the original values, and the error is back-propagated through the network, adjusting both the encoder and decoder weights. The pre-training repeats this process until the defined number of epochs is reached. When the supervised phase starts, only the encoder is used, and it already has its weights values from the pre-trained phase, so it just needs fine-tuning. The label column, "Income > \$50k" in this case, is only used in the supervised phase.

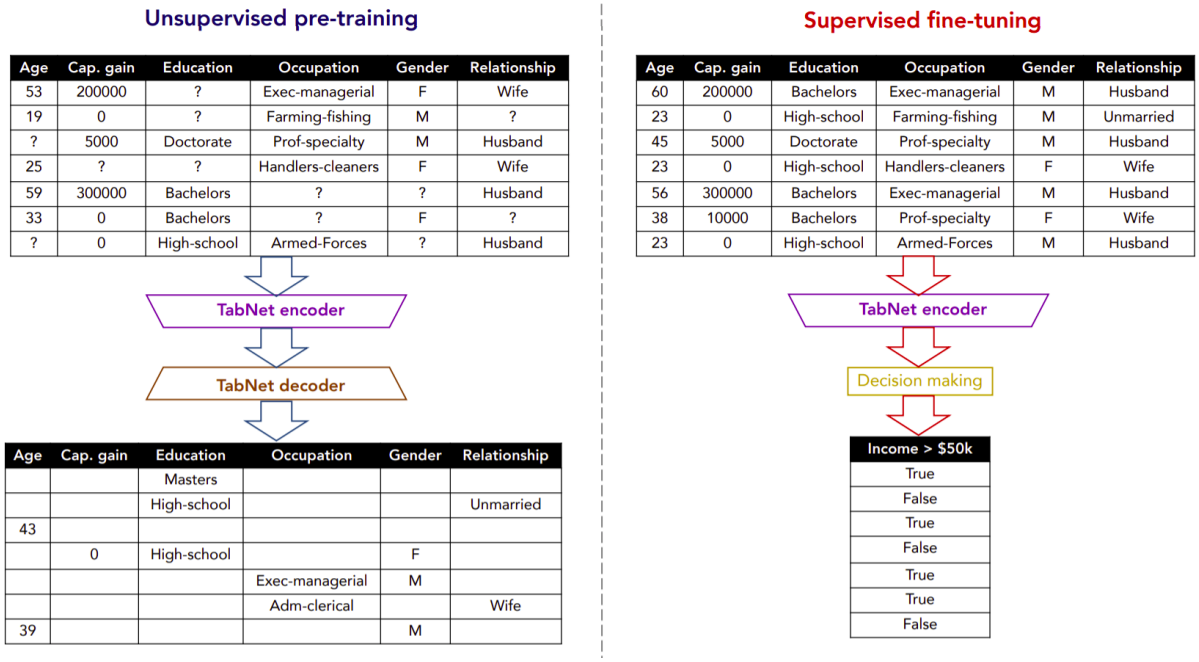


Figure 2.7: Self-supervised pre-training. Figure reproduced from Arik et al. [8].

Sequential Attention Mechanism

TabNet uses a sequential attention mechanism to select features on each decision step [72]. The mask $M[i]$ of Equation 2.2 is obtained learning the function h_i on step i processing the attributes $a[i-1]$. The attributes $a[i-1]$ are originated at the output of the feature transformers of the previous step after a split between $d[i]$ (attributes selected for decision) and $a[i]$ (attributes selected for attention on the following step). This split is controlled by two hyper-parameters, N_d, and N_a. In this way, TabNet produces variability of the

attributes utilized at each decision step. The *prior* $P[i-1]$ is calculated by Equation 2.3 and controls how many times a determined attribute has already been used on previous decision steps through hyper-parameter γ.

$$M[i] = sparsemax(P[i-1]h_i(a[i-1])), \tag{2.2}$$

$$P[i] = \prod_{j=1}^{i}(\gamma - M[j]). \tag{2.3}$$

In order to enable probabilities equal to zero and select features sparsely, instead of a softmax function, the Equation 2.2 uses a sparsemax function [47].

Finally, interpretability is achieved by combining the feature selection mask learned by the attention mechanisms on each decision step, as shown in Figure 2.8. Masks 1 to 4 were learned on decision steps 1 to 4. Each column in the figure represents one feature, ranging from X_1 to X_11. Each row represents one test sample. The figure is reproduced from the TabNet paper, and for clarity, they used synthetic datasets. On synthetic dataset 2 (Syn2), only one feature was selected for each decision step. This can be seen by the white pixels columns, representing the highest attention given, in contrast to the black pixels. M_{agg} is the normalized aggregation of the activation map of each step. We can see gray pixels of different intensities, representing how much attention was given for each feature in each sample. By doing this, we can see which feature was important for the prediction for each sample (row). Synthetic dataset 4 (Sync4) is a more challenging dataset with more than one feature selected in each step. However, data scientists can still verify which features the model is selecting and the relative importance of those features, shredding light inside the "black box" model.

2.2.2 Rejection Option

Following interpretability, a second challenge is a reliability. For critical applications, a model must indicate when its prediction is unreliable. This task is called classification with rejection option [24, 48]. Selective classification trades classifier coverage off against accuracy [33]. The classifier is allowed to output "unknown" for certain samples. As some samples are not classified, *coverage* is defined as the fraction of classified samples in the dataset. This is achieved by transforming the C class problem into a $C+1$ class; the last class is the unknown class. The selective classification or rejection option is now formally defined as follows. Given an input x, a selective classifier outputs

$$(f, g)(x) = \begin{cases} Abstain, & g(x) \geq \tau \\ f(x), & \text{otherwise,} \end{cases} \tag{2.4}$$

where τ is a threshold that controls the coverage trade-off.

The rejection option transforms a binary classification problem into a three-class problem: positive, negative, and "don't know". The neural network learns the functions f and g, with f the conventional output of a classifier corresponding to C classes and g an additional class representing the "don't know" class.

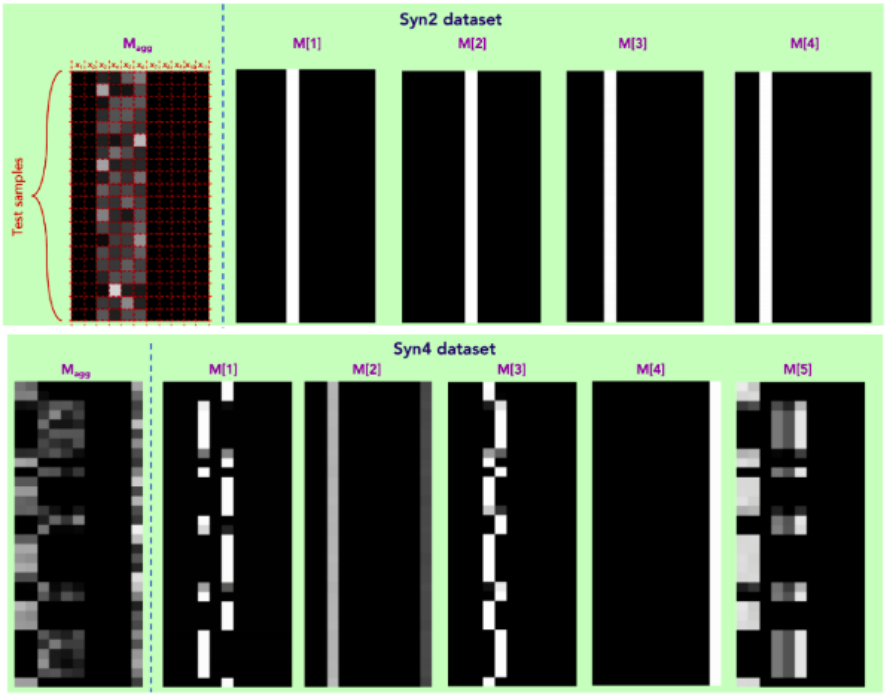


Figure 2.8: Attention maps of each decision block and the combined map. The lighter square means greater attention is given to that sample feature on a grayscale. A black square means that the following feature is irrelevant for that sample. Figure reproduced from Arik et al. [8].

Although studied since 1957, there has not been any well-established, effective method to assess prediction uncertainty, especially for deep learning models [14]. Recent research has proposed viable solutions, such as SelectiveNet, a DNN architecture that can wrap other DNNs and add a new prediction head, called "selective head" [23], and Deep Glamber, proposing a new loss which is based in portfolio theory [44]. Those works focus on image classification with deep neural networks. This Master's thesis uses the state-of-the-art Self Adaptive Training (SAT) technique [33]. We call our network TabNet+SAT, and to the best of our knowledge, those selective classification methods have not yet been tried with tabular data. Also, as referred by Geng et al. [24], the interpretability of the rejection option seems to have not been discussed yet.

Self Adaptive Training

Self Adaptive Training (SAT) [33] is a training methodology to implement a selective classifier with a rejection option. The last layer of an SAT network has $C + 1$ neurons, and their output is normalized with a softmax function. The selective function s indicates whether a sample was selected based on the value of g and τ. Coverage ϕ represents the percentage of classified samples of a given labeled set S_m, defined by the sum of selected samples from S_m divided by the number of elements.

$$s(x_i) = \begin{cases} 1, & g(x) \geq \tau \\ 0, & \text{otherwise.} \end{cases} \tag{2.5}$$

$$\phi(s|S_m) = \frac{1}{m}\sum_{i=1}^{m} s(x_i). \tag{2.6}$$

The self-adaptive approach consists of training the model for E_s number of epochs just like any other regular training. After E_s epochs, SAT uses the model's predictions to correct training data using a moving average t_i, called *targets*. t_i is initialized with the ground truth labels. Equation 2.7 defines how t_i is updated.

$$t_i = \alpha \times t_i + (1 - \alpha) \times p_i. \tag{2.7}$$

On every training step after E_s epochs, t_i is updated based on α hyper-parameter and is step by step corrected towards model's prediction p_i, which contains $C+1$ classes. This extra class provides the model with the ability to abstain. In other words, in the presence of label ambiguity, it is expected that the target t_i will differ from the original label value y_i. If the shift of t_i is towards the correct prediction, the difference between targets and predictions will diminish, and the prediction and target will converge. The classifier is trained end-to-end with a special loss function. Given a mini-batch of m samples and data pairs $\{(x_i, y_i)\}_m$, model predictions p_i and its exponential moving average t_i for each sample, the classifier f is optimized by minimizing:

$$\mathcal{L}(f) = -\frac{1}{m}\sum[\mathbf{t}_{i,y_i}\log\mathbf{p}_{i,y_i} + (1 - \mathbf{t}_{i,y_i})\log\mathbf{p}_{i,c}], \tag{2.8}$$

where p_{y_i} is the log probability for the class of the ith index of the one hot-encoded label vector y and $p_{i,c}$ is the log probability of the last column of the logit array, which is the unknown class because every SAT NN outputs a $c+1$ logit vector for every sample. Thus, the first term of Equation 2.8 is the cross-entropy loss used in standard classifiers. The second term is the rejection option loss, identifying uncertain samples in the dataset. If t_{i,y_i} is too small, the sample is deemed uncertain, and the second term will cause the classifier to reject. If t_{i,y_i} is close to one, the loss acts as a cross-entropy loss, and the classifier will correct the prediction.

Chapter 3

Related Work

The medical and data science community is actively researching better disease risk prediction algorithms. To gather a representative sample of works using machine learning for disease risk prediction, we took inspiration from Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) [50]. However, this review does not intend to be a formal meta-analysis.

We used Google Scholar with the query "AI machine learning medical health disease risk prevention predication imbalanced "supervised learning"', from 2018, which gave 4,970 results. To refine the results, we excluded from the search query the works related to image, ethics, and natural language processing (NLP), resulting in 99 articles. Table 3.1 summarizes the works.

Table 3.1: Summary of ML-based risk prediction works.

	Summary
Disease/Problem	acute kidney injury, epidemiology, patient no show, sleep apnea, diabetes, heart failure, infant health, hyperchloremia, intensive care unit (ICU) mortality risk, neck pain
Algorithm	deep neural networks (DNN), decision trees (DT), random forests (RF), extreme gradient boosting (XGBoost), logistic regression (LR), support vector machines (SVM), k-nearest neighbors (kNN)
Dataset Size	191 to 33,329 samples, 10 to 1225 attributes
Metric	area under the curve (AUC), accuracy, F1, specificity (true negative rate), sensitivity (recall, true positive rate), precision, kappa score

We can see that the disease/problem range is wide. Every medical health issue could be an application candidate for risk prediction using tabular data, as this kind of data is regularly collected from patients during hospital or clinics' daily routine. Regarding metrics, most of the traditional machine learning metrics are used, which demonstrates that there is no consensus over which metric to use. This also makes it difficult to compare the results among works.

By evaluating those works, we found that state of the art regarding risk prediction using tabular data in the medical context is the application of traditional machine learning techniques, such as logistic regression (LR), random forest (RF), support vector machine (SVM), and extreme gradient boosting machine (XGBoost). Some works also use deep neural networks (DNN), but they are still overwhelmed by decision tree-based algorithms. The dataset ranged from a few hundred to 30,000 samples and 10 to 1,000 attributes. Considering that most of those datasets are imbalanced, where the number of positive labels is a small fraction of the dataset, this particular application is challenging for DNN. There is also a trend in the literature to use several algorithms and compare the results, focusing on performance. Only a few works aim at the model interpretability [41, 62], and none deal with the rejection option.

3.1 Cardiovascular Risk Calculators Works

In addition to this more general search, we also searched for works proposing ML algorithms to replace the most commonly used risk calculators, TIMI [6, 11] and GRACE [26]. The attempt to replace traditional risk calculators with ML and DL is not new. There are dozens of works comparing ML algorithms with risk calculators [18, 35, 37, 57, 63, 64]. Despite that, we could not find any work that shared the dataset or the source code, making their results reproduction impossible. Here, we present some interesting works that provide valuable insights for further research.

Kasim et al. have several publications on this topic [9, 38, 39]. Interestingly, although DL had achieved superior performance over the traditional TIMI calculator, the authors point out the "black box" problem of DL algorithms. With a slightly inferior performance, a logistic regression algorithm can provide a risk prediction while being "transparent", i.e., which features/weights were selected/learned. Panchavati et al. [52] used an XGBoost algorithm with SHAP (SHapley Additive exPlanations) [45]. They achieved better results with XGBoost over TIMI and GRACE and found that the patient history for myocardial infarction feature contributed the most to the models' output. However, although SHAP is widely used and provides both global and local interpretability, Panchavati's article does not explore how medical doctors could benefit from interpretability. Inspired by such works, we aimed to find state-of-the-art algorithms that could work with tabular data, are interpretable, and could evaluate their prediction reliability.

3.2 Deep Learning with Tabular Data Works

Once we identified this opportunity, we searched for works related to tabular data and deep learning. Mikael Huss's blogpost [34] greatly summarizes the subject, bringing three approaches: multi-layer DNN, NODE [54], and TabNet [8]. The multi-layer DNN approach is exemplified by Airbnb's experience with tabular data [28]. Typically, big companies, such as Airbnb, have a huge amount of data, and DNN significantly benefits from it. Despite that, Airbnb faced several implementation problems in dealing with tabular data. Eventually, their DNN model outperformed the gradient boost decision tree (GBDT) but

lacked interpretability.

Neural oblivious decision ensemble (NODE) [54] is a DNN architecture inspired by decision trees. Its implementation is based on a differentiable decision tree implemented with neurons. It uses *entmax* [53], a non-linear function similar to *softmax* that effectively performs a "soft" splitting, as it forces weights to zero or one, instead of the asymptotic behavior of softmax. Entmax is a generalization of both softmax, and sparsemax [47], used by TabNet (see Section 2.2.1). The smoothness of such "splitting" is controlled by the parameter α, as shown in Figure 3.1.

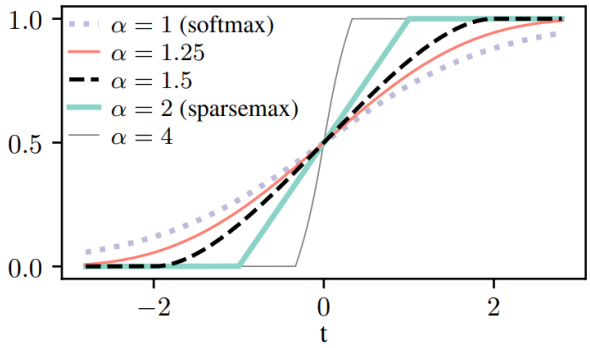


Figure 3.1: Entmax function, a generalization of softmax and sparsemax, enables the soft splitting behavior of decision trees by neural networks. Figure reproduced from Peters et al. [53].

Very recent works are presenting alternatives to TabNet, such as GATE (Gated Additive Tree Ensemble) [36], NBM (Neural Basis Model) [58], Hopular (Modern Hopfield Networks for Tabular Data) [65], FT-Transformer (Feature Tokenizer Transformer) [25], and SAINT (Self-Attention and Intersample Attention Transformer) [68]. Those neural networks have the drawback of being a "black box"; they lack interpretability. For this reason, we choose TabNet, an architecture designed for tabular data and built-in interpretability.

Chapter 4

Datasets

One of the most remarkable contributions of this Master's thesis is a real-world dataset called *Myocardial ISchemIa prognostic EValuation AngioCardio-Clarity*. This dataset was the result of the cleaning, pre-processing, and aggregation of data received from hospitals and clinics in Brazil and is used to obtain a risk calculator from real-world data. The well-known Framingham Heart Study [1] dataset was also used to compare our results with literature results in the same domain.

We also evaluated and confirmed our results in the same datasets from TabNet's article, covering a wide variety of domains and sizes, namely Mushroom, KDD Cup 2009, HIGGS, and Adult Census. Table 4.1 contains a summary of the datasets used.

Table 4.1: Summary of all datasets used in this Master's thesis.

Dataset	**#Features**	**#Samples**	**#Classes**
MI-SIEVE ACC	207	9635	2
Framingham Heart Study	20	11,627	2
Mushroom	22	8124	2
KDD Cup 2009	230	50,000	3
HIGGS	28	11,000,000	2
Adult Census	14	48,842	2

4.1 Medical Datasets

4.1.1 MI-SIEVE ACC

The Myocardial ISchemIa prognostic EValuation AngioCardio-Clarity (MI-SIEVE ACC) dataset was created from patient electronic health records who have undergone percutaneous coronary intervention. Data was originally divided into tables: *Patient*, *Intervention*, *Balloon*, *Stent*, *Vessel*, *Complication*, and *Outcome*. Those tables comprise clinical data such as age, weight, hypertension, diabetes, and inter-procedural information like duration, the number of vessels treated, and drugs used. Table 4.2 summarizes each table's features.

Table 4.2: MI-SIEVE ACC (Myocardial ISchemIa prognostic EValuation AngioCardio-Clarity) dataset. #Feat. stands for the number of features.

Table	**Summary**	**#Feat.**	**Type of Features**
Patient	patient personal information	46	age, sex, race, smoker, number of cigarettes per day, family antecedents, comorbidities
Intervention	intervention information	42	intervention date, time, type, drugs administered
Medical Complication	information of the medical complications during the intervention	9	degree of complication, final destination, time of the complication
Balloon	information regarding the balloons used during the intervention	41	angiography result, degree of stenosis, TIMI flow after, lesion classification (ACC/AHA), and drugs taken as adenosine, adrenaline, papaverine
Stent	information regarding the stent installed	48	coronary artery, length and diameter of the stent and angiography result
Vessel	information regarding the treated vessel	21	vessel identification, degree of lesion, degree of stenosis, type of lesion and calcification
Outcome	information of the intervention outcomes	17	clinical success, infarction, death, occlusion, vascular complication, renal failure, stroke, bleeding

Although it is very common to collect patient data, it was not collected for training machine learning models. Thus, it imposes several challenges, such as:

- The lack of unique identification (IDs) keys for patients and interventions (there are IDs collisions due to the identical IDs for different clinics and hospitals);
- Missing data for essential features like blood pressure, cardiac frequency, breath frequency, and body temperature;
- Data is exported from relational tables, so it needs to be structured before feeding it to ML algorithms;
- Tables have a hierarchical relationship: one line of table *Patient* is related to I interventions that treated V vessels using B balloons and S stents;
- Multi-outcome problems: bleeding, kidney complications, infarction, and death;
- An imbalanced dataset with positive labels representing less than 10% of samples.

Extensive work of cleaning, structuring, pre-processing, aggregation, and feature engineering was done to transform raw data into a dataset useful for ML. The domain specialist, Luiz Sérgio Fernandes de Carvalho, guided all this process, analyzing feature by feature. All data identified as corrupted, inconsistent, or with a significant number of missing values was excluded.

A preliminary statistical analysis of the data also revealed biases. To avoid those biases, we excluded some features to obtain a fair and responsible model.

Of 9639 patients, 69.1% are male, and 30.9% are female. Although it is evident that this data is imbalanced, it is also known that men and women have a different probability of cardiovascular diseases [51].

Regarding ethnicity, there is a great disproportion of the "white" ethnicity, composing 93.5% of samples. According to the Brazilian Institute of Geography and Statistics (IBGE, *Instituto Brasileiro de Geografia e Estatística*) [2], only 42.7% of the population self-identifies with the "white" ethnicity, meanwhile 46.8% with "brown/black" ones. As the Brazilian Society of Cardiology (SBC, *Sociedade Brasileira de Cardiologia*) [51] does not report any relevant cardiovascular risk related to ethnicity, we opt to exclude this feature from the dataset. The same reasoning was applied to the degree of education.

Concerning the other features in the MI-SIEVE ACC dataset, we found no other biases, as those are patient clinic information such as sex, age, blood pressure, diabetes, and hypertension, or intervention information such as vessel condition, drugs administered, complications, and stent information.

The use of patient data in the MI-SIEVE ACC database was approved by the Research Ethics Committee (CEP/IGESDF, *Comitê de Ética em Pesquisa do Instituto de Gestão Estratégica de Saúde do Distrito Federal*) according to the approval number 3.854.051 on February 21st, 2020, and approval 4.263.940 on September 8th, 2020.

We provide here the datasheet of this dataset [22], which is a methodological approach to document a dataset's origin, composition, and uses.

4.1.2 Framingham Heart Study

The Framingham Heart Study (FHS) dataset [1] is a longitudinal investigation of constitutional and environmental factors influencing the development of cardiovascular disease in men and women. Examination of participants has occurred every two years, and the cohort has been followed for morbidity and mortality over that period. It began in 1948, sampling patients in Framingham, Massachusetts. The objectives are to study the incidence and prevalence of cardiovascular disease and its risk factors. The cardiovascular disease conditions under investigation include coronary heart disease (angina pectoris, myocardial infarction, coronary insufficiency, and sudden and non-sudden death), stroke, hypertension, peripheral arterial disease, and congestive heart failure.

The data used in this Master's thesis is the "teaching" dataset, and although the dataset can be used to validate the machine learning techniques, the findings and conclusions cannot be used for medical publications. The National Heart, Lung, and Blood Institute (NHLBI), which made this dataset available, explicitly ask to make all users aware of such limitation. The dataset consists of 11,627 samples and 39 features, of which three are informative (randid, time, and period), 20 are features, and 16 are outcomes.

Table 4.3 shows the baseline characteristics of the FHS individuals grouped by any cardiovascular heart disease (ANYCHD) outcome, the outcome we chose for risk prediction. We omit from this table the columns RANDID, TIME, and PERIOD. We also provide the total missing data counts. HDLC and LDLC have most of their data missing as the study started collecting only total cholesterol, and only recently did HDLC and LDLC start to be collected. Table 4.4 gives a detailed description of each feature and outcome.

Table 4.3: Baseline characteristics of the Framingham Heart Study dataset individuals with or without any cardiovascular heart disease (ANYCHD). SD stands for standard deviation. Please refer to Table 4.4 for understanding Variable and Units.

Variable	Units	Missing	Overall	Without ANYCHD	With ANYCHD
n (total)			11,627	8469	3158
SEX, n (%)	1	0	5022 (43.2)	3255 (38.4)	1767 (56.0)
	2		6605 (56.8)	5214 (61.6)	1391 (44.0)
AGE, mean (SD)		0	54.8 (9.6)	53.8 (9.4)	57.4 (9.5)
SYSBP, mean (SD)		0	136.3 (22.8)	133.6 (21.6)	143.5 (24.4)
DIABP, mean (SD)		0	83.0 (11.7)	82.1 (11.2)	85.6 (12.5)
BPMEDS, n (%)	0	593	10,090 (91.4)	7492 (93.2)	2598 (86.8)
	1		944 (8.6)	548 (6.8)	396 (13.2)
CURSMOKE, n (%)	0	0	6598 (56.7)	4804 (56.7)	1794 (56.8)
	1		5029 (43.3)	3665 (43.3)	1364 (43.2)
CIGPDAY, mean (SD)		79	8.3 (12.2)	8.1 (12.0)	8.6 (12.6)
EDUC, n (%)	1	295	4690 (41.4)	3206 (38.8)	1484 (48.3)
	2		3410 (30.1)	2610 (31.6)	800 (26.1)
	3		1885 (16.6)	1480 (17.9)	405 (13.2)
	4		1347 (11.9)	965 (11.7)	382 (12.4)
TOTCHOL, mean (SD)		409	241.2 (45.4)	238.1 (44.2)	249.4 (47.5)
HDLC, mean (SD)		8600	49.4 (15.6)	50.7 (15.6)	45.6 (15.0)
LDLC, mean (SD)		8601	176.5 (46.9)	174.0 (45.8)	183.5 (49.2)
BMI, mean (SD)		52	25.9 (4.1)	25.6 (4.0)	26.7 (4.3)
GLUCOSE, mean (SD)		1440	84.1 (25.0)	82.8 (21.1)	87.7 (33.0)
DIABETES, n (%)	0	0	11,097 (95.4)	8219 (97.0)	2878 (91.1)
	1		530 (4.6)	250 (3.0)	280 (8.9)
HEARTRTE, mean (SD)		6	76.8 (12.5)	76.8 (12.3)	76.8 (12.8)
PREVAP, n (%)	0	0	11,000 (94.6)	8469 (100.0)	2531 (80.1)
	1		627 (5.4)		627 (19.9)
PREVCHD, n (%)	0	0	10,785 (92.8)	8469 (100.0)	2316 (73.3)
	1		842 (7.2)		842 (26.7)
PREVMI, n (%)	0	0	11,253 (96.8)	8469 (100.0)	2784 (88.2)
	1		374 (3.2)		374 (11.8)
PREVSTRK, n (%)	0	0	11,475 (98.7)	8388 (99.0)	3087 (97.8)
	1		152 (1.3)	81 (1.0)	71 (2.2)
PREVHYP, n (%)	0	0	6283 (54.0)	5017 (59.2)	1266 (40.1)
	1		5344 (46.0)	3452 (40.8)	1892 (59.9)

Table 4.4: Framingham Heart Study detailed data description.

Variable	**Description**	**Units**	**Range/Count**	**Data Type**
RANDID	unique identification number for each participant		2448-9999312	id
PERIOD	examination cycle	1=period 1	n=4434	categorical
		2=period 2	n=3930	
		3=period 3	n=3263	
TIME	number of days since baseline exam		0-4854	numeric
SEX	participant sex	1=men	n=5022	categorical
		2=women	n=6605	
AGE	age at exam (years)		32-81	numeric
SYSBP	systolic blood pressure (mean of last two of three measurements) (mmHg)		83.5-295	numeric
DIABP	diastolic blood pressure (mean of last two of three measurements) (mmHg)		30-150	numeric
BPMEDS	use of anti-hypertensive medication at exam	0=not currently used	n=10,090	categorical
		1=current use	n=944	

CURSMOKE	current cigarette smoking at exam	0=not current smoker 1=current smoker	n=6598 n=5029	categorical
CIGPDAY	number of cigarettes smoked each day	0=not current smoker 1-90 cigarettes per day		numeric
EDUC	attained education	1=0-11 years 2=high school diploma, GED 3=some college, vocational school 4=college (BS, BA) degree or more		categorical
TOTCHOL	serum total cholesterol (mg/dL)		107-696	numeric
HDLC	high density lipoprotein cholesterol (mg/dL)	available for period 3 only	10-189	numeric
LDLC	low density lipoprotein cholesterol (mg/dL)	available for period 3 only	20-565	numeric
BMI	body mass index, weight in kilograms/height meters squared		14.43-56.8	numeric
GLUCOSE	casual serum glucose (mg/dL)		39-478	numeric
DIABETES	diabetic according to criteria of the first exam treated or the first exam with casual glucose of 200 mg/dL or more	0=Not a diabetic 1=Diabetic	n=11,097 n=530	categorical
HEARTRTE	heart rate (ventricular rate) in beats/min		37-220	numeric
PREVAP	prevalent angina pectoris at exam	0=free of disease 1=prevalent disease	n=11,000 n=627	categorical
PREVCHD	prevalent coronary heart disease defined as pre-existing angina pectoris, myocardial infarction (hospitalized, silent or unrecognized), or coronary insufficiency (unstable angina)	0=free of disease 1=prevalent disease	n=10,785 n=842	categorical
PREVMI	prevalent myocardial infarction	0=free of disease 1=prevalent disease	n=11,253 n=374	categorical
PREVSTRK	prevalent stroke	0=free of disease 1=prevalent disease	n=11,475 n=152	categorical
PREVHYP	prevalent hypertensive. The subject was defined as hypertensive if treated or if the second exam at which mean systolic was >=140 mmHg or mean diastolic >=90 mmHg	0=free of disease 1=prevalent disease	n=6283 n=5344	categorical

The outcomes provided have both the event type (e.g., stroke, death) and the time for more complex time-dependent analysis. Event time is counted in the number of days since the study began. The events are registered until the end of the study (still actively following patients), participant dies or cannot be contacted to ascertain his (her) status. If the outcome is present, the time of the event reflects the time of the outcome; if it is not present, the time of the event reflects the time of the followup. Table 4.5 shows each outcome's total count and percentages, the average and standard deviation of event time. As usual in a medical dataset, the outcomes are imbalanced. Table 4.6 gives a detailed description of each outcome.

4.2 TabNet Datasets

4.2.1 Mushroom

The Mushroom dataset is publicly available on the UCI Machine Learning Repository [20], obtained initially from the Audubon Society Field Guide to North American Mushrooms [43]. This dataset includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota families. Each species is

Table 4.5: Clinical outcomes of the Framingham Heart Study dataset. SD stands for standard deviation.

Variable	**Units**	**Missing**	**Overall**
n (total)			11,627
ANGINA, n (%)	0	0	9725 (83.6)
	1		1902 (16.4)
HOSPMI, n (%)	0	0	10,473 (90.1)
	1		1154 (9.9)
MI_FCHD, n (%)	0	0	9839 (84.6)
	1		1788 (15.4)
ANYCHD, n (%)	0	0	8469 (72.8)
	1		3158 (27.2)
STROKE, n (%)	0	0	10,566 (90.9)
	1		1061 (9.1)
CVD, n (%)	0	0	8728 (75.1)
	1		2899 (24.9)
HYPERTEN, n (%)	0	0	2985 (25.7)
	1		8642 (74.3)
DEATH, n (%)	0	0	8100 (69.7)
	1		3527 (30.3)
TIMEAP, mean (SD)		0	7241.6 (2477.8)
TIMEMI, mean (SD)		0	7593.8 (2136.7)
TIMEMIFC, mean (SD)		0	7543.0 (2192.1)
TIMECHD, mean (SD)		0	7008.2 (2641.3)
TIMESTRK, mean (SD)		0	7660.9 (2011.1)
TIMECVD, mean (SD)		0	7166.1 (2541.7)
TIMEHYP, mean (SD)		0	3599.0 (3464.2)
TIMEDTH, mean (SD)		0	7854.1 (1788.4)

identified as edible, poisonous, or of unknown edibility and is not recommended. This latter class was combined with the poisonous one. The guide clearly states that there is no simple rule for determining the edibility of a mushroom. There are 8124 instances with 22 features and one label. The label class is either e (edible) or p (poisonous). This dataset is balanced, having 4208 edible instances and 3916 poisonous ones. The only feature that has missing values is the stalk-root.

4.2.2 KDD Cup 2009

The KDD Cup 2009 dataset is from the Customer Relationship Management (CRM) domain, provided by the French telecom company *Orange*. The challenge is to predict customer behavior. The kinds of behavior are churn (propensity of a customer to switch to another telecom provider), appetency (likelihood of the customer to purchase new products or services), and upselling (likelihood of the customer to purchase upgrades, in other words, to increase the average "ticket" value).

We used the smaller dataset version containing 230 features and 50,000 samples. This is the same version used by TabNet benchmarks. The classification task is to predict the probability of each outcome independently, so there are three binary classification tasks with the same set of features for different labels.

KDD Cup 2009 is a challenging dataset, with most of its data composed of null values.

Table 4.6: Framingham Heart Study detailed outcome description.

Variable	Description
ANGINA	Angina Pectoris
HOSPMI	Hospitalized Myocardial Infarction
MI_FCHD	Hospitalized Myocardial Infarction or Fatal Coronary Heart Disease
ANYCHD	Angina Pectoris, Myocardial infarction (Hospitalized and silent or unrecognized), Coronary Insufficiency (Unstable Angina), or Fatal Coronary Heart Disease
STROKE	Atherothrombotic infarction, Cerebral Embolism, Intracerebral Hemorrhage, or Subarachnoid Hemorrhage or Fatal Cerebrovascular Disease
CVD	Myocardial infarction (Hospitalized and silent or unrecognized), Fatal Coronary Heart Disease, Atherothrombotic infarction, Cerebral Embolism, Intracerebral Hemorrhage, or Subarachnoid Hemorrhage or Fatal Cerebrovascular Disease
HYPERTEN	Hypertensive. Defined as the first exam treated for high blood pressure or the second exam in which either Systolic is $140 mmHg or Diastolic $90mmHg
DEATH	Death from any cause
TIMEAP	Number of days from Baseline exam to first Angina during the followup or Number of days from Baseline to censor date. Censor date may be the end of followup, death, or last known contact date if the subject is lost to followup
TIMEMI	Defined as above for the first HOSPMI event during followup
TIMEMIFC	Defined as above for the first MI_FCHD event during followup
TIMECHD	Defined as above for the first ANYCHD event during followup
TIMESTRK	Defined as above for the first STROKE event during followup
TIMECVD	Defined as above for the first CVD event during followup
TIMEHYP	Defined as above for the first HYPERTEN event during followup
TIMEDTH	Number of days from Baseline exam to death if occurring during followup or Number of days from Baseline to censor date. Censor date may be the end of followup, or the last known contact date if the subject is lost to followup

Data is also anonymized, making domain interpretation impossible. Categorical data is also challenging because one single feature has hundreds of categories. Finally, the dataset is highly imbalanced. Positive labels are only 7.3% for appetency, 1.8% for churn, and 7.3% for upselling.

4.2.3 HIGGS

HIGGS [10] is a dataset obtained from particle accelerator data. The classification problem is distinguishing between a signal process that produces Higgs bosons and a background process that does not. It is available on the UCI Machine Learning Repository [20]. The dataset is enormous compared with the other datasets used by TabNet, comprising 11,000,000 samples. This is an attractive characteristic because we can compare the performance of deep learning algorithms with more traditional ones. Deep learning is expected to benefit from large amounts of data.

There are a total of 28 features, 21 are kinematic properties measured by the particle detectors, and the last 7 are functions of the first 21, prepared by physicists to help discriminate between the two classes. Contrary to the Adult Census dataset (next section), this dataset requires some domain knowledge to interpret. Thus, we will not provide each feature description here, as they are available in Baldi et al. [10].

4.2.4 Adult Census

The Adult Census dataset is also publicly available in the UCI Machine Learning Repository [20]. It was originally extracted from the 1994 Census bureau database [40]. It has one label class indicating if one individual's yearly income is greater than U$ 50,000. The dataset consists of 48,842 samples and 14 features. Out of 14 features, 6 are continuous, and the others are categorical. Table 4.7 provides a detailed description of categorical values or an indication if the feature is continuous.

Table 4.7: Adult Census dataset features description, providing either the existent categorical values or the word "continuous".

	Feature	**Description**
1	age	continuous
2	workclass	private, self-emp-not-inc, self-emp-inc, federal-gov, local-gov, state-gov, without-pay, never-worked
3	fnlwgt	continuous
4	education	bachelors, some-college, 11th, hs-grad, prof-school, assoc-acdm, assoc-voc, 9th, 7th-8th, 12th, masters, 1st-4th, 10th, doctorate, 5th-6th, preschool
5	education-num	continuous
6	marital-status	married-civ-spouse, divorced, never-married, separated, widowed, married-spouse-absent, married-af-spouse
7	occupation	tech-support, craft-repair, other-service, sales, exec-managerial, prof-specialty, handlers-cleaners, machine-op-inspct, adm-clerical, farming-fishing, transport-moving, priv-house-serv, protective-serv, armed-forces
8	relationship	wife, own-child, husband, not-in-family, other-relative, unmarried
9	race	white, asian-pac-islander, amer-indian-eskimo, other, black
10	sex	female, male
11	capital-gain	continuous
12	capital-loss	continuous
13	hours-per-week	continuous
14	native-country	United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands

The baseline characteristics of the individuals grouped by income class are shown in Table 4.8. This dataset is highly imbalanced; as expected, there are much more people with a lower income than a higher one. It also contains features with many missing values, such as workclass, occupation, and native-country. From the baseline characteristics, we can already notice that age is a determinant factor for income. Other obvious determinants like the workclass values "without-pay" and "never-worked", which will fall on the lower income class. TabNet paper uses this dataset as the interpretability benchmark, as the domain is familiar to anyone. As most of the socio-economical datasets, the Adult Census is also biased. Women correspond only 33.1% of the individuals and "black" people only 9.6%. As our goal is only to evaluate how the algorithm learns from the data and how dubious samples are rejected, biased datasets are not a problem to this Master's thesis. They are even more "interesting" to evaluate since we can further investigate and try to interpret the algorithm decisions.

Table 4.8: Adult Census dataset baseline characteristics description grouped by the income class. SD stands for standard deviation.

				Income	
Feature		**Missing**	**Overall**	**<=50K**	**>50K**
n			32,561	24,720	7841
age, mean (SD)		0	38.6 (13.6)	36.8 (14.0)	44.2 (10.5)
workclass, n (%)	federal-gov	1836	960 (3.1)	589 (2.6)	371 (4.8)
	local-gov		2093 (6.8)	1476 (6.4)	617 (8.1)
	never-worked		7 (0.0)	7 (0.0)	
	private		22,696 (73.9)	17,733 (76.8)	4963 (64.9)
	self-emp-inc		1116 (3.6)	494 (2.1)	622 (8.1)
	self-emp-not-inc		2541 (8.3)	1817 (7.9)	724 (9.5)
	state-gov		1298 (4.2)	945 (4.1)	353 (4.6)
	without-pay		14 (0.0)	14 (0.1)	
fnlwgt, mean (SD)		0	189,778.4	190,340.9	188,005.0
			(105,550.0)	(106,482.3)	(102,541.8)
education, n (%)	10th	0	933 (2.9)	871 (3.5)	62 (0.8)
	11th		1175 (3.6)	1115 (4.5)	60 (0.8)
	12th		433 (1.3)	400 (1.6)	33 (0.4)
	1st-4th		168 (0.5)	162 (0.7)	6 (0.1)
	5th-6th		333 (1.0)	317 (1.3)	16 (0.2)
	7th-8th		646 (2.0)	606 (2.5)	40 (0.5)
	9th		514 (1.6)	487 (2.0)	27 (0.3)
	assoc-acdm		1067 (3.3)	802 (3.2)	265 (3.4)
	assoc-voc		1382 (4.2)	1021 (4.1)	361 (4.6)
	bachelors		5355 (16.4)	3134 (12.7)	2221 (28.3)
	doctorate		413 (1.3)	107 (0.4)	306 (3.9)
	hs-grad		10,501 (32.3)	8826 (35.7)	1675 (21.4)
	masters		1723 (5.3)	764 (3.1)	959 (12.2)
	preschool		51 (0.2)	51 (0.2)	
	prof-school		576 (1.8)	153 (0.6)	423 (5.4)
	some-college		7291 (22.4)	5904 (23.9)	1387 (17.7)
education-num, mean (SD)		0	10.1 (2.6)	9.6 (2.4)	11.6 (2.4)
marital-status, n (%)	divorced	0	4443 (13.6)	3980 (16.1)	463 (5.9)
	married-af-spouse		23 (0.1)	13 (0.1)	10 (0.1)
	married-civ-spouse		14,976 (46.0)	8284 (33.5)	6692 (85.3)
	married-spouse-absent		418 (1.3)	384 (1.6)	34 (0.4)
	never-married		10,683 (32.8)	10192 (41.2)	491 (6.3)
	separated		1025 (3.1)	959 (3.9)	66 (0.8)
	widowed		993 (3.0)	908 (3.7)	85 (1.1)
occupation, n (%)	adm-clerical	1843	3770 (12.3)	3263 (14.1)	507 (6.6)
	armed-Forces		9 (0.0)	8 (0.0)	1 (0.0)
	craft-repair		4099 (13.3)	3170 (13.7)	929 (12.1)
	exec-managerial		4066 (13.2)	2098 (9.1)	1968 (25.7)
	farming-fishing		994 (3.2)	879 (3.8)	115 (1.5)
	handlers-cleaners		1370 (4.5)	1284 (5.6)	86 (1.1)
	machine-op-inspct		2002 (6.5)	1752 (7.6)	250 (3.3)
	other-service		3295 (10.7)	3158 (13.7)	137 (1.8)
	priv-house-serv		149 (0.5)	148 (0.6)	1 (0.0)
	prof-specialty		4140 (13.5)	2281 (9.9)	1859 (24.3)
	protective-serv		649 (2.1)	438 (1.9)	211 (2.8)
	sales		3650 (11.9)	2667 (11.6)	983 (12.8)
	tech-support		928 (3.0)	645 (2.8)	283 (3.7)
	transport-moving		1597 (5.2)	1277 (5.5)	320 (4.2)
relationship, n (%)	husband	0	13,193 (40.5)	7275 (29.4)	5918 (75.5)
	not-in-family		8305 (25.5)	7449 (30.1)	856 (10.9)
	other-relative		981 (3.0)	944 (3.8)	37 (0.5)
	own-child		5068 (15.6)	5001 (20.2)	67 (0.9)
	unmarried		3446 (10.6)	3228 (13.1)	218 (2.8)
	wife		1568 (4.8)	823 (3.3)	745 (9.5)
race, n (%)	amer-indian-eskimo	0	311 (1.0)	275 (1.1)	36 (0.5)
	asian-pac-islander		1039 (3.2)	763 (3.1)	276 (3.5)
	black		3124 (9.6)	2737 (11.1)	387 (4.9)
	other		271 (0.8)	246 (1.0)	25 (0.3)

. . . continued

. . . continued

Feature		Missing	Overall	Income <=50K	Income >50K
	white		27,816 (85.4)	20,699 (83.7)	7117 (90.8)
sex, n (%)	female	0	10,771 (33.1)	9592 (38.8)	1179 (15.0)
	male		21,790 (66.9)	15,128 (61.2)	6662 (85.0)
capital-gain, mean (SD)		0	1077.6 (7385.3)	148.8 (963.1)	4006.1 (14,570.4)
capital-loss, mean (SD)		0	87.3 (403.0)	53.1 (310.8)	195.0 (595.5)
hours-per-week, mean (SD)		0	40.4 (12.3)	38.8 (12.3)	45.5 (11.0)
native-country, n (%)	Cambodia	583	19 (0.1)	12 (0.0)	7 (0.1)
	Canada		121 (0.4)	82 (0.3)	39 (0.5)
	China		75 (0.2)	55 (0.2)	20 (0.3)
	Columbia		59 (0.2)	57 (0.2)	2 (0.0)
	Cuba		95 (0.3)	70 (0.3)	25 (0.3)
	Dominican-Republic		70 (0.2)	68 (0.3)	2 (0.0)
	Ecuador		28 (0.1)	24 (0.1)	4 (0.1)
	El-Salvador		106 (0.3)	97 (0.4)	9 (0.1)
	England		90 (0.3)	60 (0.2)	30 (0.4)
	France		29 (0.1)	17 (0.1)	12 (0.2)
	Germany		137 (0.4)	93 (0.4)	44 (0.6)
	Greece		29 (0.1)	21 (0.1)	8 (0.1)
	Guatemala		64 (0.2)	61 (0.3)	3 (0.0)
	Haiti		44 (0.1)	40 (0.2)	4 (0.1)
	Holand-Netherlands		1 (0.0)	1 (0.0)	
	Honduras		13 (0.0)	12 (0.0)	1 (0.0)
	Hong		20 (0.1)	14 (0.1)	6 (0.1)
	Hungary		13 (0.0)	10 (0.0)	3 (0.0)
	India		100 (0.3)	60 (0.2)	40 (0.5)
	Iran		43 (0.1)	25 (0.1)	18 (0.2)
	Ireland		24 (0.1)	19 (0.1)	5 (0.1)
	Italy		73 (0.2)	48 (0.2)	25 (0.3)
	Jamaica		81 (0.3)	71 (0.3)	10 (0.1)
	Japan		62 (0.2)	38 (0.2)	24 (0.3)
	Laos		18 (0.1)	16 (0.1)	2 (0.0)
	Mexico		643 (2.0)	610 (2.5)	33 (0.4)
	Nicaragua		34 (0.1)	32 (0.1)	2 (0.0)
	Outlying-US (Guam-USVI-etc)		14 (0.0)	14 (0.1)	
	Peru		31 (0.1)	29 (0.1)	2 (0.0)
	Philippines		198 (0.6)	137 (0.6)	61 (0.8)
	Poland		60 (0.2)	48 (0.2)	12 (0.2)
	Portugal		37 (0.1)	33 (0.1)	4 (0.1)
	Puerto-Rico		114 (0.4)	102 (0.4)	12 (0.2)
	Scotland		12 (0.0)	9 (0.0)	3 (0.0)
	South		80 (0.3)	64 (0.3)	16 (0.2)
	Taiwan		51 (0.2)	31 (0.1)	20 (0.3)
	Thailand		18 (0.1)	15 (0.1)	3 (0.0)
	Trinadad&Tobago		19 (0.1)	17 (0.1)	2 (0.0)
	United-States		29,170 (91.2)	21,999 (90.6)	7171 (93.2)
	Vietnam		67 (0.2)	62 (0.3)	5 (0.1)
	Yugoslavia		16 (0.1)	10 (0.0)	6 (0.1)

Chapter 5

Methodology

Figure 5.1 illustrates our methodology inspired by *Knowledge Discovery in Databases* (KDD) [21], the process of discovering useful knowledge from data. The training pipeline consists of the following steps: **selection**, **pre-processing**, **k-folding**, **modeling**, **evaluation and interpretation**.

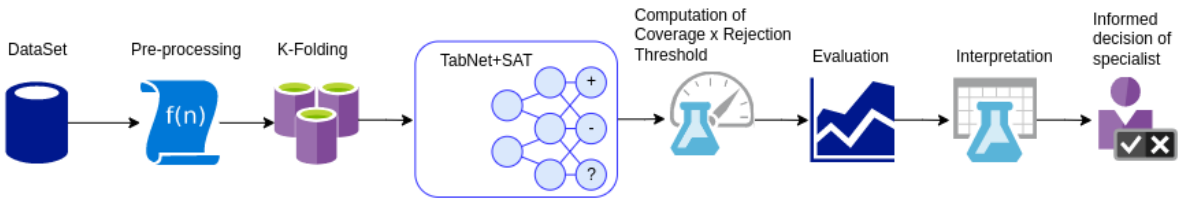


Figure 5.1: Training pipeline composed of the steps: pre-processing, k-folding splits, TabNet+SAT model as well the baseline models and interpretation/evaluation.

In the **selection** step, two types of tabular datasets suitable for a binary classification task are chosen. The first type is benchmark datasets, especially the ones used in the TabNet paper [8]. We aim to confirm TabNet results and also evaluate the rejection option's performance on those datasets. The second type is medical datasets, focusing on cardiovascular disease risk prediction. We selected the target variables for the medical datasets according to the cardiologist Dr. Luiz Sérgio Fernandes de Carvalho.

In the **k-folding** step, we split data into train/validation stratified k-folds. The same samples are fed into all models for a fair evaluation. This is done n times randomly, resulting in $k \times n$ experiment runs to assert results stability.

In the **pre-processing** step, we first clean data of any unnecessary features and labels, and then missing values are removed or imputed. Our approach to missing values follows the method used for each dataset when the source code of the experiments is available. If not, we impute the missing values with zeroes. We also use the method provided by reference experiments source code for categorical features. When the authors do not provide this information, categorical features are encoded, and the list of categorical features is later passed to TabNet as an input parameter. Our goal is to compare with the reference experiment from the TabNet paper and not to improve data pre-processing.

In the **modeling** step, we apply our modified version of TabNet, called TabNet+SAT,

which includes the Self Adaptive Training (SAT) approach according to Huang et al. [33]. SAT technique is explained in Section 2.2.2.

The goal of the TabNet+SAT model is to achieve superior performance with the trade-off of coverage, thus rejecting the samples that the model prediction is uncertain. Mixing selective classification with the interpretation capability of TabNet gives medical doctors and data scientists insights to understand why some samples are correctly classified and others are not. Also, in a production environment, this kind of classifier could help physicians make informed decisions about when the prediction is reliable — and not.

Finally, in the **evaluation and interpretation** step, we use the AUC metric (Area Under the ROC Curve), as accuracy is tricky due to the imbalance of the dataset. The rejection is controlled by choosing the level of the rejection threshold based on the third class, the "unknown" class, that will give the desired coverage. The evaluation metric is applied only to the selected samples. The interpretation is made by selecting the best TabNet+SAT model and using the entire dataset to generate predictions. The result is an activation map matrix of rows as individual samples and the aggregated feature contribution from each TabNet step as columns. This activation map matrix is normalized on the sample axis to assess the relative importance of the feature for each sample, giving an activation map plot. Additionally, the raw probability of each class (negative, positive, "don't know") and the predicted class, the true label, and the rejection label are used for interpretation.

A simplified pipeline is used for **inference**. We apply the same pre-processing to prepare the data that is fed to the trained TabNet+SAT model, which outputs either a positive, negative, or rejected label based on the threshold hyper-parameter *tau*. The network also outputs the feature importance map that enables the domain specialist to interpret the prediction, knowing which features were considered for the prediction and their relative importance. The **inference** pipeline is shown in Figure 5.2.

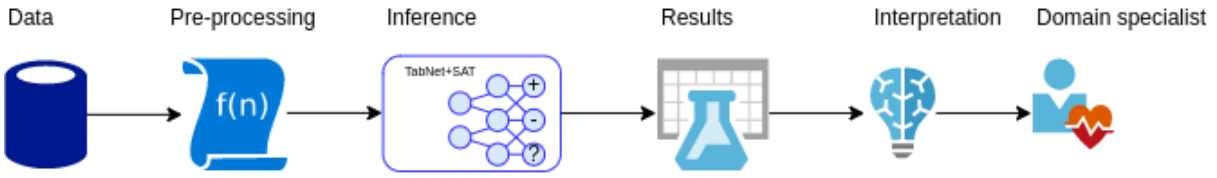


Figure 5.2: The inference pipeline comprises pre-processing, inference, results, interpretation, and decision by the domain specialist.

Chapter 6

Experiments and Results

In this chapter, we apply our methodology to several datasets to evaluate TabNet+SAT performance and its interpretation capabilities. We split the results into two groups according to the datasets used: TabNet's paper (Section 6.2) and medical datasets (Section 6.3). We spent considerable effort running experiments on all datasets where TabNet has reported initial results. We chose to demonstrate our results' reproducibility and show the performance gained by applying SAT. We also interpreted the results of the Adult Census dataset, following TabNet's paper, which uses this dataset as an interpretation benchmark. The Adult Census dataset is, in fact, a good benchmark for interpretation as it does not require any specific domain knowledge. The second group consists of two medical datasets not present in TabNet's paper. The first, the Framingham Heart Study, is a well-known dataset used to predict the risk of heart diseases. The second is the MI-SIEVE ACC dataset, a challenging real-world dataset introduced in this Master's thesis. This dataset imposes many difficulties for any algorithm and is used to evaluate TabNet+SAT in the real world.

The experiments are presented in a logical order. Chronologically, this Master's thesis starts with experiments on the MI-SIEVE dataset. During such experiments, we found out how difficult it was to measure the quality of our results and how to interpret them. Most of the literature focuses on performance, comparing ML algorithms without considering the challenges of the reliability and interpretability of those predictions. Our results show that the TabNet+SAT combination addresses the challenges of our research. The failed results are omitted, but we include several attempts with different selective classification approaches that were not reproducible, even when using the authors' source code.

6.1 Experimental Setup

In order to reproduce the TabNet results [8], the fair scenario is to use the same hyperparameters and the same data split (train/test) so that the reported accuracy and feature importance can be confirmed in TabNet and then on TabNet+SAT. Although TabNet source code is available by the authors[1], it does not contain the source code of all experiments. The original implementation was in TensorFlow 1.0 and has not been updated since

[1] `https://github.com/google-research/google-research/tree/master/tabnet`

publication. The paper describes the hyper-parameters, but the dataset pre-processing is not. There is another implementation available using PyTorch[2]. In our experiments, we chose this implementation as it is better documented, up-to-date, object-oriented, and modular. Also, the eager execution of PyTorch favors debugging.

Here, to better understand each experiment, we explain the hyperparameters used by TabNet as follows:

- N_d is the width of the decision prediction layer. Bigger values give more capacity to the model with the risk of overfitting. Values typically range from 8 to 64.
- N_a is the width of the attention embedding for each mask.
- λ_{sparse} is the extra sparsity loss coefficient. The bigger this coefficient, the sparser the model is in feature selection.
- B is the number of examples per batch.
- B_v is the virtual batch size, the size of the mini-batches used for "Ghost Batch Normalization".
- m_B is the momentum for batch normalization.
- N_{steps} is the number of steps in the architecture.
- γ is the coefficient for feature reusage in the masks. A value close to 1 makes mask selection the least correlated between layers.

We compare TabNet+SAT to Random Forest and XGBoost, using the scikit-learn[3] implementation with default hyper-parameters.

All experiments were conducted on a 64-bit Debian Linux machine powered by Intel Xeon CPU 6230 @ 2.10 GHz with 80 cores and 1.0 TB RAM, equipped with 8 NVIDIA Quadro RTX 8000 GPUs with 48GB dedicated memory. Our source code is in Python.

6.2 TabNet Datasets

6.2.1 Mushroom

The Mushroom dataset is the simplest of all datasets used by TabNet. The authors reported 100% binary accuracy in the test set, which consists of only 8,124 samples. We include it here for completeness, as we aim to reproduce our experiments in all TabNet datasets. The paper does not inform which pre-processing is done. As all features are categorical, we use dummy encoding to transform categorical features into numerical features. This increases the number of features from 22 to 117. The hyper-parameters used are the same as reported in TabNet's paper: $N_d = N_a = 8$, $\lambda_{sparse} = 0.001$, $B = 2048$, $B_v = 128$, $m_B = 0.9$, $N_{steps} = 3$, and $\gamma = 1.5$.

In our results, we also achieved 100% in all three algorithms: Random Forest, XGBoost, and TabNet, thus confirming TabNet results.

[2] `https://github.com/dreamquark-ai/tabnet/blob/develop/census_example.ipynb`
[3] `https://scikit-learn.org`

6.2.2 KDD Cup 2009

The KDD Cup 2009 dataset consists of three independent binary classification tasks with the same label features: appetency, churn, and upselling. We treated the three tasks as experiments with only the data preparation in common.

TabNet's paper does not provide the source code or details of data pre-processing, but it mentions that they used the pre-processing and split of Prohorenkoa et al. [56]. We consider that this refers to their code available on GitHub[4]. To compare the results, we use the same pre-processing, but we do not use the same split provided by Prokhorenkova et al. [56]. We split the data twice using a stratified 5-fold cross-validation, resulting in 10 experiments.

The pre-processing is done by imputing missing data with zeroes. The categorical data is encoded by casting the string values to the "category" data type and replacing them with the categorical codes. The hyper-parameters used were: $N_d = N_a = 32$, $\lambda_{sparse} = 0.001$, $B = 8,192$, $B_v = 256$, $m_B = 0.9$, $N_{steps} = 7$, and $\gamma = 1.2$. All those hyper-parameters were obtained from TabNet's paper.

TabNet [8] reported an accuracy of 98.2% for appetency, 92.7% for churn, and 95.0% for upselling in the test set. Although the results seem high, the accuracy metric is very misleading and should not be used on imbalanced datasets. Out of 50,000 labels, appetency has 890 positive labels (1.8%), churn has 3,672 positive labels (7.3%), and upselling has 3,682 positive labels (7.4%). A better metric is AUC or balanced accuracy. We chose AUC as our default metric, but for all experiments, we have also calculated for other metrics. A classifier that predicts the negative class will reach such accuracy. Moreover, in our experiments, this is exactly what happens for Random Forest, XGBoost, and TabNet. All reach the same reported accuracy by just predicting the negative class.

To achieve meaningful results, we decided to automatically balance samples using inverse frequencies, and then TabNet did not overfit in the negative class. Its accuracy was much lower, but it achieved a higher AUC and balanced accuracy. Table 6.1 shows an AUC comparison between algorithms. What is also interesting and confirms our research is that the TabNet+SAT model could also improve its results when trading-off coverage. This also indicates that TabNet+SAT could not only classify the samples but also identify, to some extent, the uncertain ones.

In our experiment, we achieved 0.792 AUC in appetency, 0.699 AUC in churn, and 0.836 AUC in upselling. For all three datasets, TabNet+SAT could increase the AUC by sacrificing coverage. At 70% coverage, we achieved 0.836 AUC in appetency (+0.044 gain), 0.740 AUC in churn (+0.041 gain), and 0.900 AUC in upselling (+0.064 gain). TabNet and TabNet+SAT at 100% coverage results are very close to each other but are different because they are different executions of the experiment. Thus, the network was trained from scratch again. We cannot interpret the results because those three datasets have anonymized features. We discuss interpretability for other datasets. For KDD, it is enough to highlight TabNet+SAT's capability of increasing performance on all three KDD datasets when sacrificing coverage. Figure 6.1 shows all algorithms' mean AUC of

[4] https://github.com/catboost/benchmarks/blob/master/quality_benchmarks/prepare_appetency_churn_upselling/prepare_appetency_churn_upselling.ipynb

Table 6.1: KDD Cup 2009 AUC results averaged of all 10 runs for Random Forest, XGBoost, TabNet, and TabNet+SAT for various coverage rates. TabNet+SAT-c% stands for TabNet+SAT under c coverage rate.

	AUC		
	appetency	churn	upselling
Random Forest	0.736 ± 0.009	0.679 ± 0.008	0.823 ± 0.008
XGBoost	0.787 ± 0.012	0.703 ± 0.008	0.850 ± 0.006
TabNet	0.792 ± 0.017	0.699 ± 0.016	0.836 ± 0.016
TabNet+SAT-100%	0.789 ± 0.021	0.706 ± 0.008	0.835 ± 0.018
TabNet+SAT-90%	0.806 ± 0.023	0.717 ± 0.009	0.854 ± 0.019
TabNet+SAT-80%	0.822 ± 0.024	0.728 ± 0.010	0.876 ± 0.020
TabNet+SAT-70%	0.836 ± 0.025	0.740 ± 0.009	0.900 ± 0.021

appetency, churn, and upselling. We can see how the metric increases when we apply rejection by sacrificing coverage.

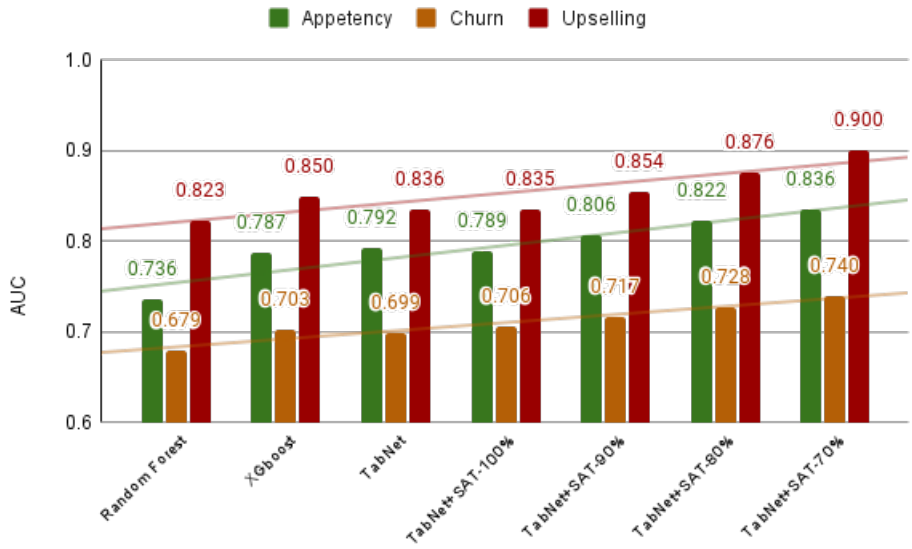


Figure 6.1: KDD AUC results of appetency, churn, and upselling for all algorithms.

6.2.3 HIGGS

The HIGGS dataset is the biggest of all TabNet datasets. It comprises 11,000,000 samples and 28 features. The original TabNet's paper intention with this dataset was to compare TabNet accuracy with different dataset sizes. Because TabNet is a DNN, it was expected to benefit from a huge sample size. TabNet's paper reports experiments with two TabNet versions: one small, 81,000 parameters, and one medium, 660,000 parameters. As this dataset takes considerable time to train, we evaluated the small version: TabNet-S.

TabNet-S hyper-parameters are $N_d = 24$, $N_a = 26$, $\lambda_{sparse} = 0.000001$, $B = 163384$, $B_v = 512$, $m_B = 0.6$, $N_{steps} = 5$, and $\gamma = 1.5$. The learning rate initial value is 0.02 and

decreases with a 0.9 ratio every 50 epochs. TabNet-S [8] achieved 78.8% accuracy in the test set.

In order to compare the results, we used the same hyper-parameters reported. The TabNet paper reported no pre-processing. The original train/test split was not made available by the authors. To reduce experiment time, we used a stratified 5-fold cross-validation without repeating it twice (as done for other experiments).

Experiment results confirmed that the dataset size was beneficial to TabNet. As shown in Table 6.2, TabNet exceeded Random Forest and XGBoost results. XGBoost achieved an AUC of 0.823, Random Forest 0.840, and TabNet 0.863. TabNet paper does not report the AUC but the accuracy, which can be misleading, especially in unbalanced datasets. Nevertheless, we also computed the accuracy to compare with the TabNet paper. The original paper reported a 78.8% accuracy, without informing if pre-training[5] was used or dataset fraction, while we achieved 77.0% without pre-training and 77.6% with pre-training. The difference is up to 1.8 percentage points. Even with this difference, the numbers demonstrate that the results are reproducible.

Table 6.2: HIGGS AUC results averaged of all 5 runs for Random Forest, XGBoost, TabNet, and TabNet+SAT for various coverage rates. TabNet+SAT-c% stands for TabNet+SAT under c coverage rate.

	AUC
Random Forest	0.840 ± 0.001
XGBoost	0.823 ± 0.001
TabNet	0.856 ± 0.005
TabNet+SAT-100%	0.847 ± 0.003
TabNet+SAT-90%	0.864 ± 0.003
TabNet+SAT-80%	0.881 ± 0.003
TabNet+SAT-70%	0.897 ± 0.003

The performance gain from unsupervised learning was 0.007. Although not negligible, the impact is low when all the dataset is labeled. TabNet paper reports one experiment with 100,000 samples in which test accuracy increased from 72.9% to 73.2% (+0.27%). The gain is much higher with 1,000 samples, from 57.5% to 61.4% (+3.9%) with pre-training. In real-world scenarios, pre-training could be very beneficial if many samples are unlabeled, and TabNet could be only fine-tuned in the small labeled fraction. Experiments with other datasets have not demonstrated significant performance gain with pre-training.

Finally, in Table 6.2, we also observe that TabNet+SAT AUC increased from 0.847 at 100% coverage to 0.897 at 70% coverage, a five percentage points. This confirms that SAT can learn how to reject samples for different dataset sizes. Figure 6.2 shows the mean AUC of the HIGGS experiment for all algorithms. We can see how the metric increases when we apply rejection by sacrificing coverage.

Regarding the model interpretability, as the HIGGS dataset domain is related to sub-atomic particles, it is hard to interpret. We discuss the interpretability in the next section.

[5]TabNet is pretrained as an unsupervised model (i.e., with a self-supervised learning task).

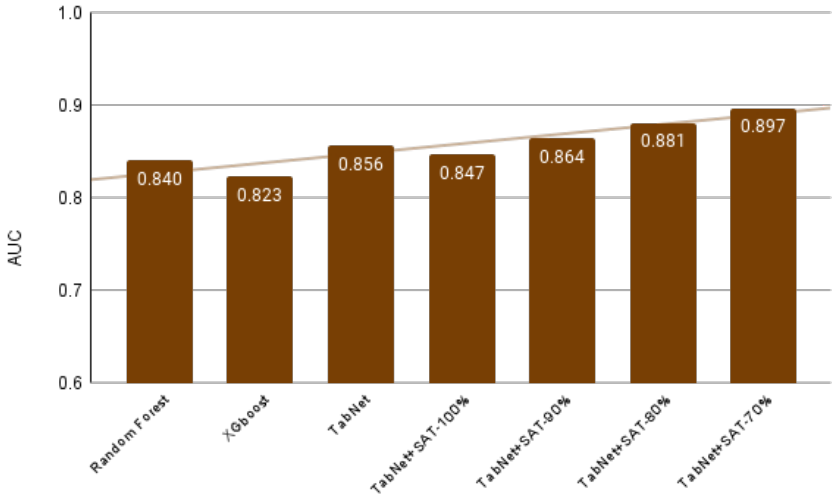


Figure 6.2: HIGGS AUC results for all algorithms.

6.2.4 Adult Census

TabNet authors chose the Adult Census dataset to demonstrate the interpretation capabilities of TabNet. Its features are comprehensible and do not require any domain knowledge. Following TabNet's paper [8], we also use this dataset to evaluate and demonstrate TabNet+SAT interpretability.

TabNet [8] reported accuracy of 85.7% in the test set (not available) using the following hyper-parameters: $N_d = N_a = 16$, $\lambda_{sparse} = 0.0001$, $B = 4096$, $B_v = 128$, $m_B = 0.98$, $N_{steps} = 5$, and $\gamma = 1.5$. The original paper does not mention the settings for learning rate and training epochs. Dreamquark-ai Pytorch example uses the following hyper-parameters: $N_d = N_a = 8$, $\lambda_{sparse} = 0.0001$, $B = 1024$, $B_v = 128$, $m_B = 0.02$, $N_{steps} = 5$, and $\gamma = 1.3$. The learning rate initial value is 0.02 and decreases every 50 epochs with a 0.9 ratio. The only pre-processing applied was encoding the non-numerical features using the scikit-learn label encoder and imputing missing values using the feature's mean value. The provided source code is reproducible and achieved an accuracy of 82.7%. Although originally, the accuracy metric was used, we also computed AUC. Dreamquark-ai Adult Census example achieved an AUC of 0.918.

The same hyper-parameters the TabNet's paper provided and the Dreamquark-ai example pre-processed were used. We also used a 5-fold cross-validation twice, resulting in 10 experiments. We achieved an average accuracy of 82.5% and an average AUC of 0.924. Table 6.3 shows the results for TabNet, Random Forest, and XGBoost algorithms, and our TabNet+SAT results under coverage rates ranging from 70% to 100%.

TabNet performance was practically equal to XGBoost and better than Random Forest, confirming the TabNet findings. We also confirm that applying SAT on TabNet significantly increases its performance. With 70% of coverage, the AUC increases from 0.924 to 0.969. This means that the TabNet+SAT model learns which samples had the highest uncertainty and thus rejects them. Figure 6.3 shows the mean AUC of the Adult Census experiment for all algorithms. We can see how the metric increases when we apply rejection by sacrificing coverage.

Table 6.3: Adult Census AUC results averaged of all 10 runs for Random Forest, XGBoost, TabNet, and TabNet+SAT for various coverage rates. TabNet+SAT-c% stands for TabNet+SAT under c coverage rate.

	AUC
Random Forest	0.907 ± 0.004
XGBoost	0.926 ± 0.003
TabNet	0.926 ± 0.004
TabNet+SAT-100%	0.926 ± 0.004
TabNet+SAT-90%	0.942 ± 0.004
TabNet+SAT-80%	0.957 ± 0.004
TabNet+SAT-70%	0.969 ± 0.003

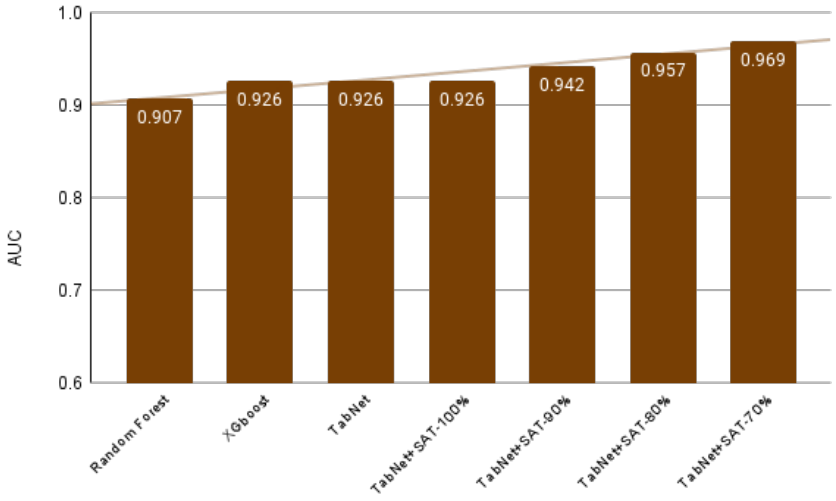


Figure 6.3: Adult Census results for all algorithms.

Let us now consider the model interpretability. Figure 6.4 represents feature importances for all 10 training splits. The vertical axis contains the names of the features, while the horizontal axis contains the relative importance in percentage, summing up to 100%. We bring the attention that the importance varies according to the training split, and this variance is not negligible. Despite the variance, we can still interpret which feature group is the most important. For instance, *capital-gain* (see Table 4.8) is the most important feature on 9 out of 10 runs. The feature *age* is also important, but our results differ from TabNet paper because they report *age* as the most important one. It is reasonable to assume that this difference is actually due to different models learning different representations based on different feature importances.

To answer the question of rejection interpretation, we propose a selective classification interpretation map inspired by the heatmap plot of the SHAP python package [45]. Figure 6.5 displays this plot, with samples sorted by uncertainty, with the least uncertain at the top and the most uncertain at the bottom. The left side heatmap is similar to TabNet aggregated attention map, with values normalized considering the entire dataset. Points

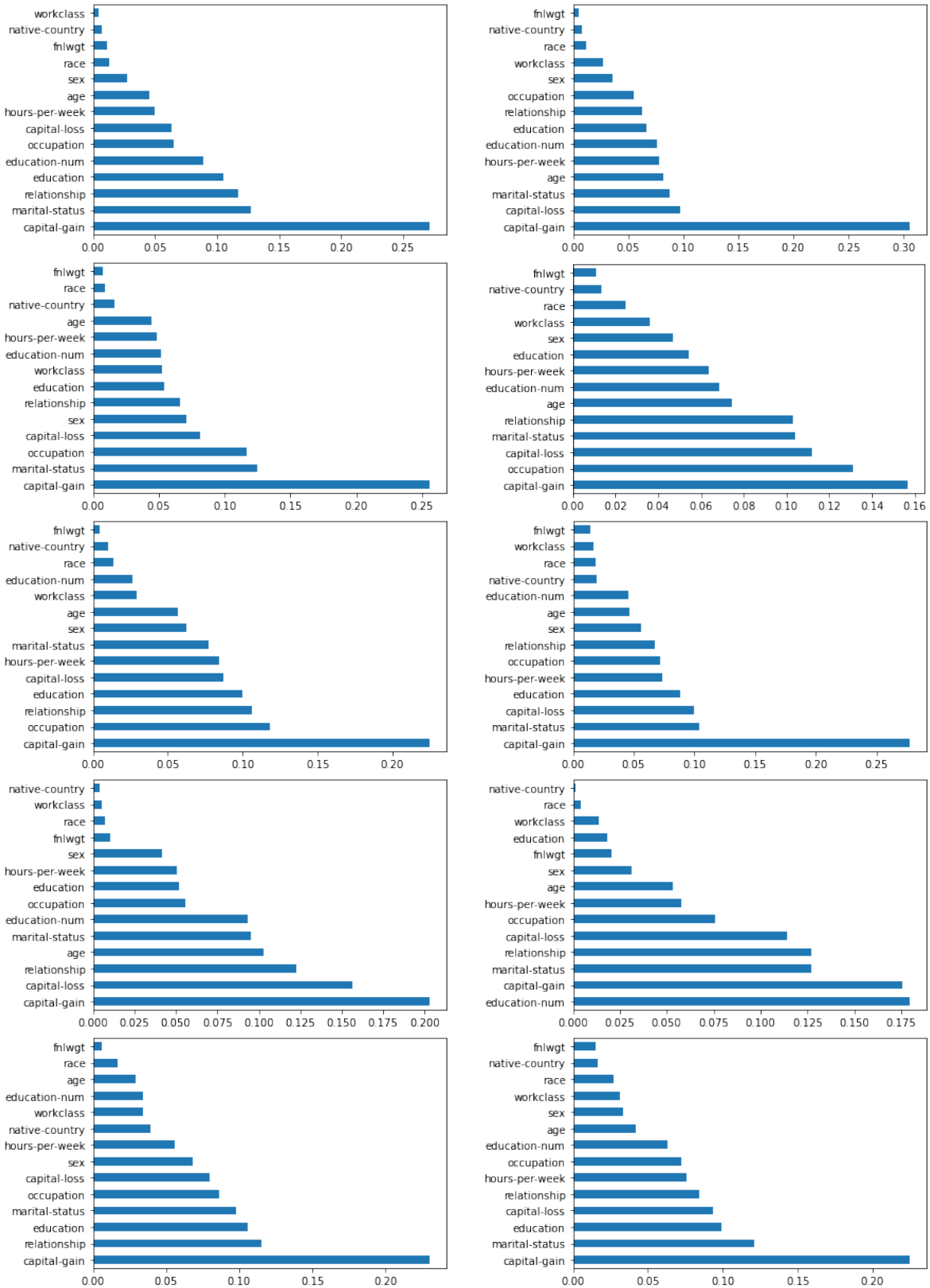


Figure 6.4: The TabNet+SAT feature importance plots for each run. There is considerable variability between the runs.

with colors near yellow received most of the neural network attention. The purple points received the slightest attention. The right side heatmap provides insight into the labels. *Pre_ Negative*, *Pred_ Positive*, and *uncertainty* sum to one, as explained in Section 2.2.2. Using the blue to red scale, white stands right in the middle and can be interpreted as doubt. *Predicted* is equal to one if *Pred_ Positive* is greater than *Pred_ Negative*, and zero otherwise. *True_ Positive* is the ground truth label. Placing *Pred_ Positive* side by side with *Predicted* helps to identify the regions that the model is classifying correctly or not. *Norm_ uncertainty* is the normalized value of uncertainty, from 0 to 1, considering the entire dataset. Finally, *rejected* column indicates if the sample was rejected, which happens when uncertainty is above the calculated threshold for the given coverage. For Figure 6.5, the rejected blue area covers 70% of the image, as this was the requested coverage. TabNet' paper does not sort this plot, but we found that sorting could help interpretability and naturally provide meaning for certain regions.

We can distinctly see three regions. Region 1, samples ranging from 0 to 5,000 in Figure 6.6, has a very consistent group of samples with no rejection and very accurate predictions. The *capital-gain* is the most active feature for this group. A high *capital-gain* means a higher *income*. The model ignores *marital-status*, *sex*, and *relationship* in this case. We will further explore whether those features are the cause of uncertainty.

In region 2, samples ranging from 5,000 to 20,000, we see most samples classified as low *income*. Despite the small amount, the positive samples here are still correctly classified, and, for those, we observe that the TabNet+SAT pays attention to *capital-gain*. For negative samples, the attention in this region is shifted towards *marital-status* and *relationship*. In the third region, samples ranging from 20,000 to 32,000, we see a gradual increase in uncertainty and a growing misclassification. In this region, we see divided attention on capital gain, occupation, *marital-status*, and *relationship*. Although these are, in fact, features that relate to *income*, they are not decisive. In other words, *occupation*, for example, can also contribute significantly to *income*, but this cannot be a general rule. The exceptions for those cases impact the accuracy and thus contribute to a higher uncertainty value. As can be seen, TabNet+SAT correctly identifies those samples and reject classifying them.

The categorical features *occupation*, *marital-status*, and *relationship* are also sources of uncertainty. Figure 6.7 is a Marimekko[6] chart that demonstrates which categorical values are driving uncertainty. We observe that "craft-repair" and "transport-moving" both have around 50% of rejection rate. This means that for the dataset samples with the categorical feature *occupation* with such values, TabNet+SAT has a higher error rate, and thus while minimizing the SAT loss, the network understands the pattern and rejections from classifying half of the samples which have "craft-repair" and "transport-moving". We cannot further drill down and understand how those features were collected, but interestingly, TabNet+SAT points to a problem with a specific feature and on specific values of those features. In real-world datasets, data scientists could further engineer those features, for example, separating the "transport-moving" feature into airplane pilot, truck driver, or taxi driver.

We observe similar behavior in *marital-status* and *relationship*, which are proxies of

[6] `https://datavizcatalogue.com/methods/marimekko_chart.html`

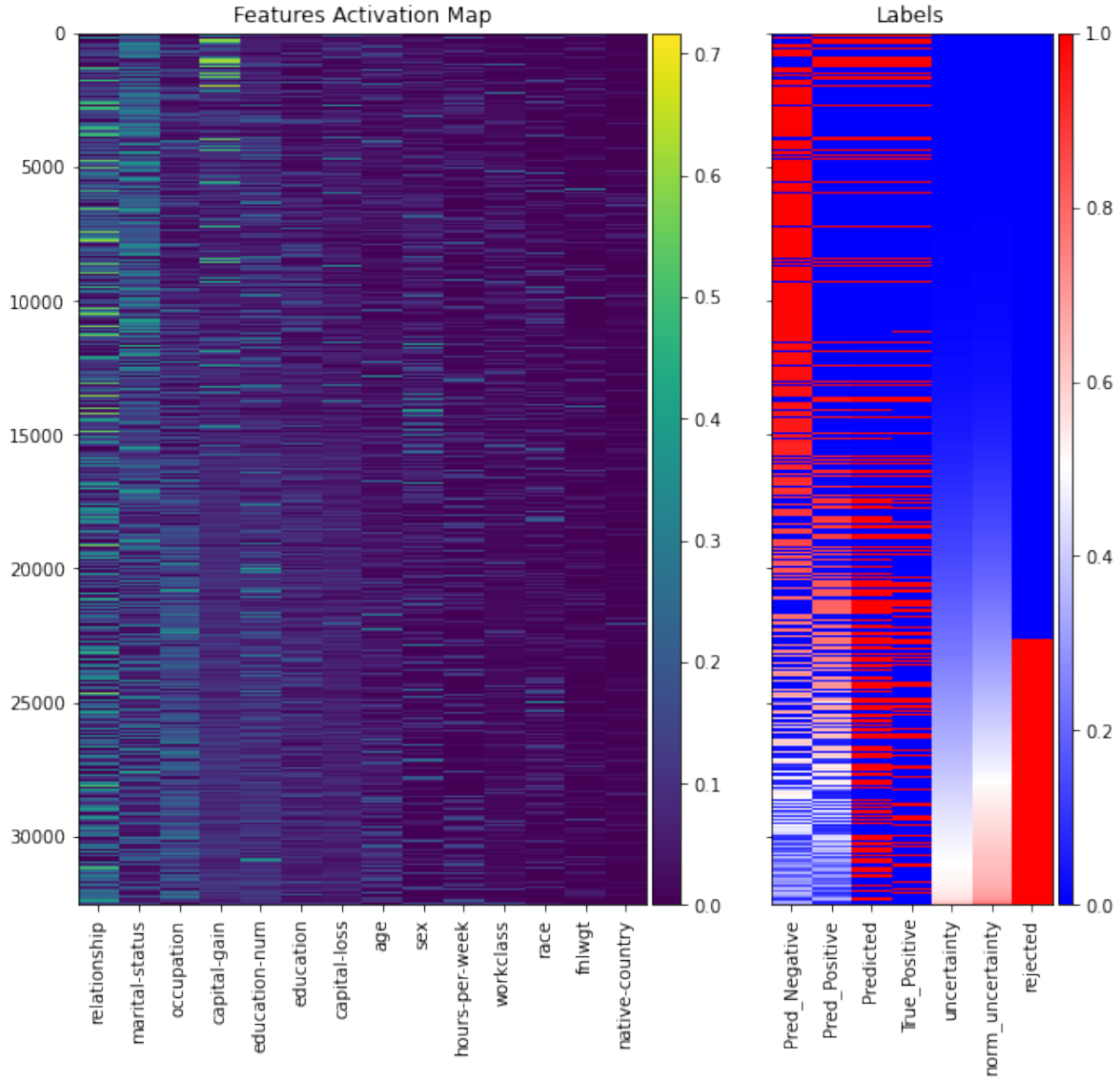


Figure 6.5: The whole Adult Census dataset rejection option interpretation map with feature activation map and prediction outputs. Features are sorted by importance, and samples are sorted by uncertainty.

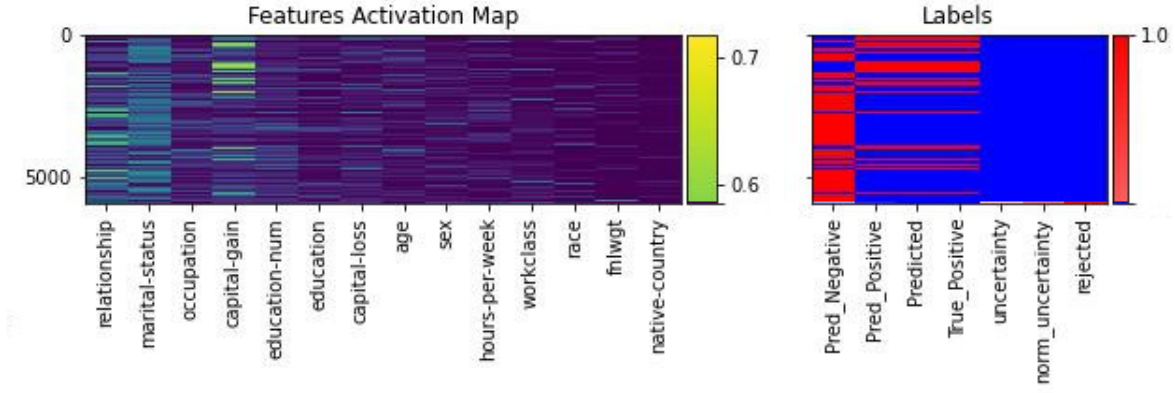
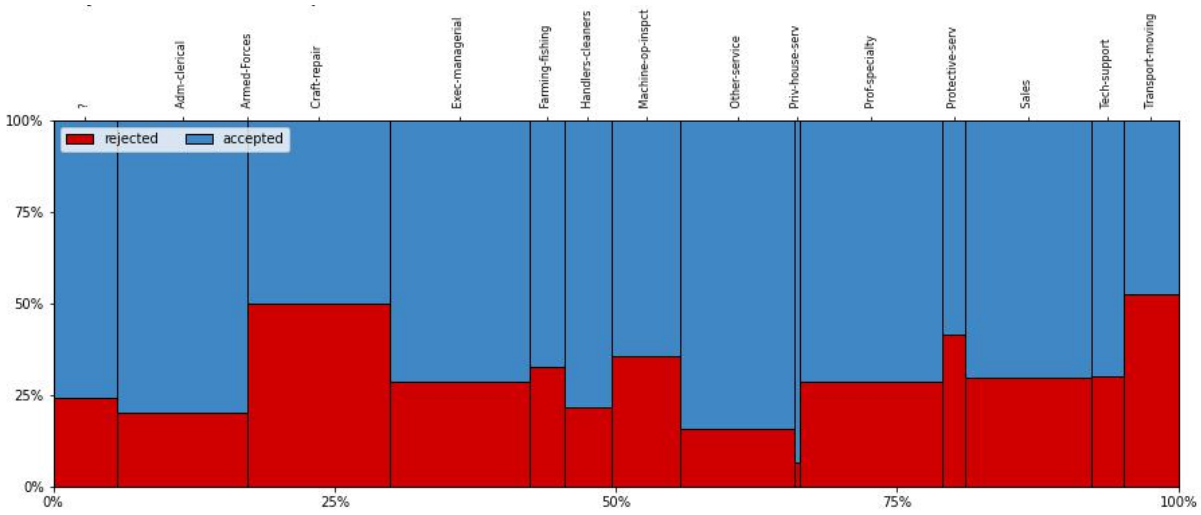
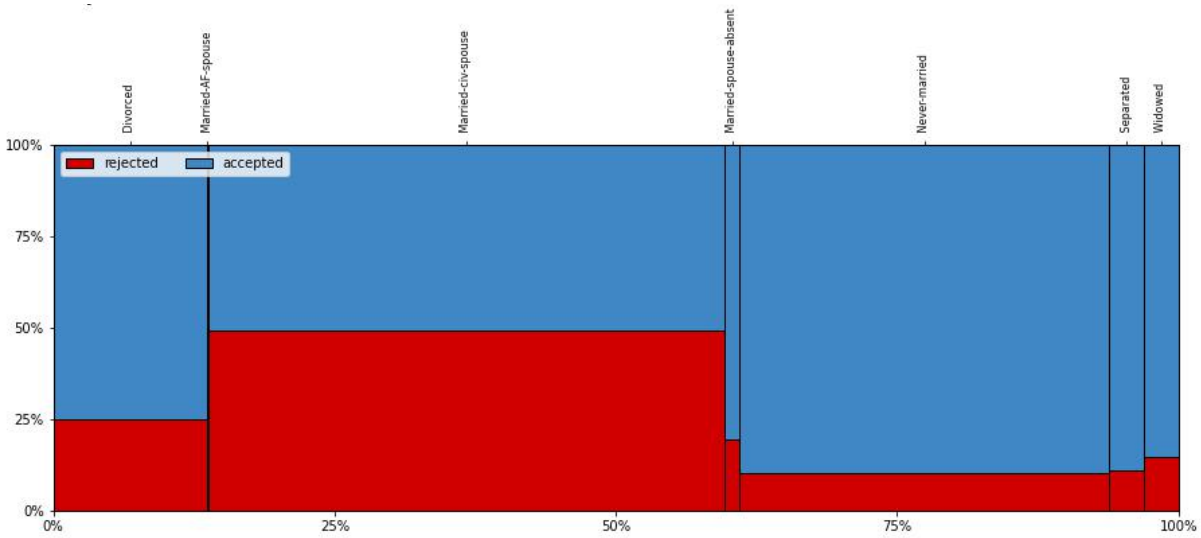


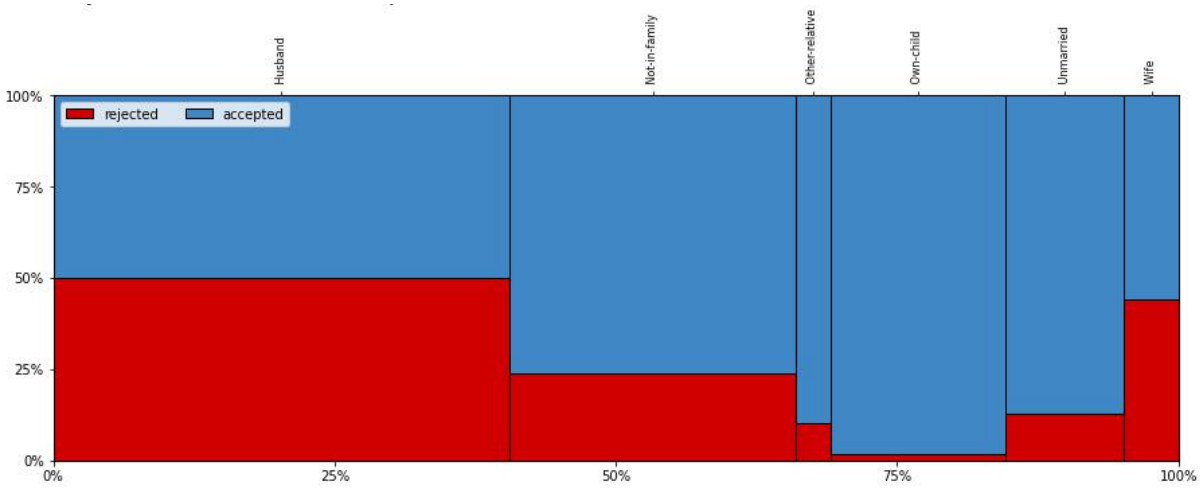
Figure 6.6: Regions 1. Samples from 0 to 5000 demonstrate accurate predictions with a high emphasis on the capital gain.



(a) Occupations "craft-repair" and "transport-moving" are the ones with higher rejection rates.



(b) Relationship "married-civ-spouse" has near 50% of rejection rate.



(c) Marital-status "husband" and "wife" have near 50% of rejection rate.

Figure 6.7: Adult Census Marimekko plots of categorical features demonstrating which categories are responsible for higher rejection rates, thus, sources of uncertainty.

each other. Clearly, TabNet+SAT has a problem classifying samples of husbands. As we also observe in Table 6.4, TabNet+SAT predicts that out of 13,193 husbands, 9,479 have an *income* higher than 50K. This is incorrect because the ground truth tells that only 5,918 husbands have such *income*. TabNet+SAT compensates for this error with a higher rejection rate, rejecting 6,614 samples of the dataset that have the *relationship* feature as a husband. Here, we have a clear case of a neural network being affected by a dataset bias and exacerbating it. In other words, if a sample is from a husband, guessing a higher *income* will favor the classification results, so the neural network exacerbates the bias already present in the dataset. But with SAT, there is also an option to abstain because the algorithm is not forced to guess. Thus, being a source of error, TabNet+SAT learns to attribute uncertainty to those samples and thus reject it.

Table 6.4: Baseline characteristics of the Adult Census dataset grouped by predicted class, ground truth class, and rejection.

			Grouped by Predicted		Grouped by True_Positive		Grouped by rejected	
		Overall	0	1	0	1	0	1
n		32561	20687	11874	24720	7841	22632	9929
age, mean (SD)		21.6 (13.6)	18.5 (14.2)	26.9 (10.6)	19.8 (14.0)	27.2 (10.5)	20.1 (14.1)	25.0 (11.9)
workclass, n (%)	?	1836 (5.6)	1525 (7.4)	311 (2.6)	1645 (6.7)	191 (2.4)	1390 (6.1)	446 (4.5)
	Federal-gov	960 (2.9)	475 (2.3)	485 (4.1)	589 (2.4)	371 (4.7)	636 (2.8)	324 (3.3)
	Local-gov	2093 (6.4)	1173 (5.7)	920 (7.7)	1476 (6.0)	617 (7.9)	1422 (6.3)	671 (6.8)
	Never-worked	7 (0.0)	7 (0.0)		7 (0.0)		6 (0.0)	1 (0.0)
	Private	22696 (69.7)	15141 (73.2)	7555 (63.6)	17733 (71.7)	4963 (63.3)	16221 (71.7)	6475 (65.2)
	Self-emp-inc	1116 (3.4)	306 (1.5)	810 (6.8)	494 (2.0)	622 (7.9)	694 (3.1)	422 (4.3)
	Self-emp-not-inc	2541 (7.8)	1285 (6.2)	1256 (10.6)	1817 (7.4)	724 (9.2)	1356 (6.0)	1185 (11.9)
	State-gov	1298 (4.0)	765 (3.7)	533 (4.5)	945 (3.8)	353 (4.5)	897 (4.0)	401 (4.0)
	Without-pay	14 (0.0)	10 (0.0)	4 (0.0)	14 (0.1)		10 (0.0)	4 (0.0)
fnlwgt, mean (SD)		189778.4 (105550.0)	191916.7 (106956.3)	186052.9 (102952.4)	190340.9 (106482.3)	188005.0 (102541.8)	193704.5 (107892.7)	180829.1 (99432.5)
education, n (%)	10th	933 (2.9)	899 (4.3)	34 (0.3)	871 (3.5)	62 (0.8)	716 (3.2)	217 (2.2)
	11th	1175 (3.6)	1136 (5.5)	39 (0.3)	1115 (4.5)	60 (0.8)	986 (4.4)	189 (1.9)
	12th	433 (1.3)	380 (1.8)	53 (0.4)	400 (1.6)	33 (0.4)	346 (1.5)	87 (0.9)
	1st-4th	168 (0.5)	163 (0.8)	5 (0.0)	162 (0.7)	6 (0.1)	136 (0.6)	32 (0.3)
	5th-6th	333 (1.0)	321 (1.6)	12 (0.1)	317 (1.3)	16 (0.2)	307 (1.4)	26 (0.3)
	7th-8th	646 (2.0)	601 (2.9)	45 (0.4)	606 (2.5)	40 (0.5)	498 (2.2)	148 (1.5)
	9th	514 (1.6)	506 (2.4)	8 (0.1)	487 (2.0)	27 (0.3)	403 (1.8)	111 (1.1)
	Assoc-acdm	1067 (3.3)	632 (3.1)	435 (3.7)	802 (3.2)	265 (3.4)	688 (3.0)	379 (3.8)
	Assoc-voc	1382 (4.2)	805 (3.9)	577 (4.9)	1021 (4.1)	361 (4.6)	803 (3.5)	579 (5.8)
	Bachelors	5355 (16.4)	2331 (11.3)	3024 (25.5)	3134 (12.7)	2221 (28.3)	3756 (16.6)	1599 (16.1)
	Doctorate	413 (1.3)	42 (0.2)	371 (3.1)	107 (0.4)	306 (3.9)	325 (1.4)	88 (0.9)
	HS-grad	10501 (32.3)	7451 (36.0)	3050 (25.7)	8826 (35.7)	1675 (21.4)	6592 (29.1)	3909 (39.4)
	Masters	1723 (5.3)	471 (2.3)	1252 (10.5)	764 (3.1)	959 (12.2)	1285 (5.7)	438 (4.4)
	Preschool	51 (0.2)	51 (0.2)		51 (0.2)		51 (0.2)	
	Prof-school	576 (1.8)	49 (0.2)	527 (4.4)	153 (0.6)	423 (5.4)	478 (2.1)	98 (1.0)
	Some-college	7291 (22.4)	4849 (23.4)	2442 (20.6)	5904 (23.9)	1387 (17.7)	5262 (23.3)	2029 (20.4)
education-num, mean (SD)		9.1 (2.6)	8.3 (2.4)	10.4 (2.2)	8.6 (2.4)	10.6 (2.4)	9.0 (2.7)	9.2 (2.2)
marital-status, n (%)	Divorced	4443 (13.6)	3934 (19.0)	509 (4.3)	3980 (16.1)	463 (5.9)	3339 (14.8)	1104 (11.1)
	Married-AF-spouse	23 (0.1)	12 (0.1)	11 (0.1)	13 (0.1)	10 (0.1)	9 (0.0)	14 (0.1)
	Married-civ-spouse	14976 (46.0)	4267 (20.6)	10709 (90.2)	8284 (33.5)	6692 (85.3)	7616 (33.7)	7360 (74.1)
	Married-spouse-absent	418 (1.3)	379 (1.8)	39 (0.3)	384 (1.6)	34 (0.4)	337 (1.5)	81 (0.8)
	Never-married	10683 (32.8)	10200 (49.3)	483 (4.1)	10192 (41.2)	491 (6.3)	9572 (42.3)	1111 (11.2)
	Separated	1025 (3.1)	971 (4.7)	54 (0.5)	959 (3.9)	66 (0.8)	911 (4.0)	114 (1.1)
	Widowed	993 (3.0)	924 (4.5)	69 (0.6)	908 (3.7)	85 (1.1)	848 (3.7)	145 (1.5)
occupation, n (%)	?	1843 (5.7)	1532 (7.4)	311 (2.6)	1652 (6.7)	191 (2.4)	1396 (6.2)	447 (4.5)
	Adm-clerical	3770 (11.6)	2929 (14.2)	841 (7.1)	3263 (13.2)	507 (6.5)	3011 (13.3)	759 (7.6)
	Armed-Forces	9 (0.0)	8 (0.0)	1 (0.0)	8 (0.0)	1 (0.0)	7 (0.0)	2 (0.0)
	Craft-repair	4099 (12.6)	2362 (11.4)	1737 (14.6)	3170 (12.8)	929 (11.8)	2054 (9.1)	2045 (20.6)
	Exec-managerial	4066 (12.5)	1414 (6.8)	2652 (22.3)	2098 (8.5)	1968 (25.1)	2902 (12.8)	1164 (11.7)
	Farming-fishing	994 (3.1)	867 (4.2)	127 (1.1)	879 (3.6)	115 (1.5)	668 (3.0)	326 (3.3)
	Handlers-cleaners	1370 (4.2)	1256 (6.1)	114 (1.0)	1284 (5.2)	86 (1.1)	1075 (4.7)	295 (3.0)
	Machine-op-inspct	2002 (6.1)	1524 (7.4)	478 (4.0)	1752 (7.1)	250 (3.2)	1288 (5.7)	714 (7.2)
	Other-service	3295 (10.1)	3063 (14.8)	232 (2.0)	3158 (12.8)	137 (1.7)	2777 (12.3)	518 (5.2)
	Priv-house-serv	149 (0.5)	146 (0.7)	3 (0.0)	148 (0.6)	1 (0.0)	139 (0.6)	10 (0.1)
	Prof-specialty	4140 (12.7)	1681 (8.1)	2459 (20.7)	2281 (9.2)	1859 (23.7)	2960 (13.1)	1180 (11.9)
	Protective-serv	649 (2.0)	305 (1.5)	344 (2.9)	438 (1.8)	211 (2.7)	380 (1.7)	269 (2.7)
	Sales	3650 (11.2)	2101 (10.2)	1549 (13.0)	2667 (10.8)	983 (12.5)	2565 (11.3)	1085 (10.9)
	Tech-support	928 (2.9)	522 (2.5)	406 (3.4)	645 (2.6)	283 (3.6)	650 (2.9)	278 (2.8)
	Transport-moving	1597 (4.9)	977 (4.7)	620 (5.2)	1277 (5.2)	320 (4.1)	760 (3.4)	837 (8.4)
relationship, n (%)	Husband	13193 (40.5)	3714 (18.0)	9479 (79.8)	7275 (29.4)	5918 (75.5)	6579 (29.1)	6614 (66.6)
	Not-in-family	8305 (25.5)	7420 (35.9)	885 (7.5)	7449 (30.1)	856 (10.9)	6319 (27.9)	1986 (20.0)
	Other-relative	981 (3.0)	927 (4.5)	54 (0.5)	944 (3.8)	37 (0.5)	879 (3.9)	102 (1.0)
	Own-child	5068 (15.6)	5011 (24.2)	57 (0.5)	5001 (20.2)	67 (0.9)	4975 (22.0)	93 (0.9)
	Unmarried	3446 (10.6)	3233 (15.6)	213 (1.8)	3228 (13.1)	218 (2.8)	3005 (13.3)	441 (4.4)
	Wife	1568 (4.8)	382 (1.8)	1186 (10.0)	823 (3.3)	745 (9.5)	875 (3.9)	693 (7.0)
race, n (%)	Amer-Indian-Eskimo	311 (1.0)	246 (1.2)	65 (0.5)	275 (1.1)	36 (0.5)	223 (1.0)	88 (0.9)
	Asian-Pac-Islander	1039 (3.2)	611 (3.0)	428 (3.6)	763 (3.1)	276 (3.5)	727 (3.2)	312 (3.1)
	Black	3124 (9.6)	2557 (12.4)	567 (4.8)	2737 (11.1)	387 (4.9)	2509 (11.1)	615 (6.2)
	Other	271 (0.8)	225 (1.1)	46 (0.4)	246 (1.0)	25 (0.3)	204 (0.9)	67 (0.7)
	White	27816 (85.4)	17048 (82.4)	10768 (90.7)	20699 (83.7)	7117 (90.8)	18969 (83.8)	8847 (89.1)
sex, n (%)	Female	10771 (33.1)	9192 (44.4)	1579 (13.3)	9592 (38.8)	1179 (15.0)	9151 (40.4)	1620 (16.3)
	Male	21790 (66.9)	11495 (55.6)	10295 (86.7)	15128 (61.2)	6662 (85.0)	13481 (59.6)	8309 (83.7)
capital-gain, mean (SD)		6.5 (23.3)	2.4 (11.8)	13.7 (34.1)	2.0 (11.0)	20.6 (40.2)	9.3 (27.4)	0.2 (3.7)
capital-loss, mean (SD)		2.1 (10.1)	1.3 (7.5)	3.4 (13.3)	1.1 (7.2)	5.0 (15.7)	2.9 (11.7)	0.2 (3.7)
hours-per-week, mean (SD)		39.4 (12.1)	36.9 (11.9)	43.8 (11.3)	37.8 (12.1)	44.4 (10.7)	38.2 (12.2)	42.1 (11.6)

6.3 Medical Datasets

6.3.1 Framingham Heart Study

Framingham Heart Study is our first medical dataset. TabNet's original paper did not report it, but it is well-known by cardiologists and widely used. As the dataset is small and unbalanced, it was necessary to use a portion of the data for the validation split. Initially, we applied a stratified k-fold cross-validation with five folds, resulting in very few positive samples on the validation set. Thus, we used $k = 3$ for three rounds, with random splits, to evaluate experiment variance.

To compare TabNet results with traditional tabular algorithms, we fit the same data on Random Forest [12] and XGBoost [13] as a baseline. We fit the data on an unmodified version of TabNet without selective classification as a sanity check. The default hyper-parameters were used for Random Forest and XGBoost, as our objective was not to fine-tune any model but compare the results. TabNet hyper-parameters were $N_a = 8$, $N_d = 8$, $N_{steps} = 3$, learning rate = $2e^{-2}$, learning rate decay = 0.9, decay steps = 10, batch size = 256, virtual batch size = 128, and patience = 50 epochs.

The average AUC for TabNet was 0.764, for Random Forest 0.760, and 0.736 for XGBoost. TabNet+SAT achieved an average AUC of 0.768 at 100% coverage. Again, SAT proved to be a useful tool to improve metrics. A significant increase in AUC was achieved, reaching 0.815 (+0.051) at 70% coverage. The values for all nine runs can be found in Table 6.5.

Table 6.5: Framingham Heart Study AUC results averaged of all 9 runs for Random Forest, XGBoost, TabNet, and TabNet+SAT for various coverage rates. TabNet+SAT-c% stands for TabNet+SAT under c coverage rate.

	AUC
Random Forest	0.760 ± 0.005
XGBoost	0.736 ± 0.009
TabNet	0.764 ± 0.003
TabNet+SAT-100%	0.768 ± 0.005
TabNet+SAT-90%	0.783 ± 0.005
TabNet+SAT-80%	0.799 ± 0.006
TabNet+SAT-70%	0.815 ± 0.007

To interpret the results, we used the aggregated TabNet activation map, side by side with the results of the predictions (raw positive and negative probabilities, predicted value, ground truth, uncertainty, normalized uncertainty, and rejection result). As we have raw probabilities for positive and negative classes, uncertainty is the raw probability of the unknown class. Features are sorted by importance, while samples are sorted in Figure 6.8 by uncertainty, from the lowest to the highest uncertainty. Details of this chart were given in Section 6.2.4. As the samples are sorted by uncertainty, we can interpret this figure by observing the distinct sections of samples. In the topmost part, from samples 0 to about 1,000, we can see a section where the uncertainty is practically zero, and all samples

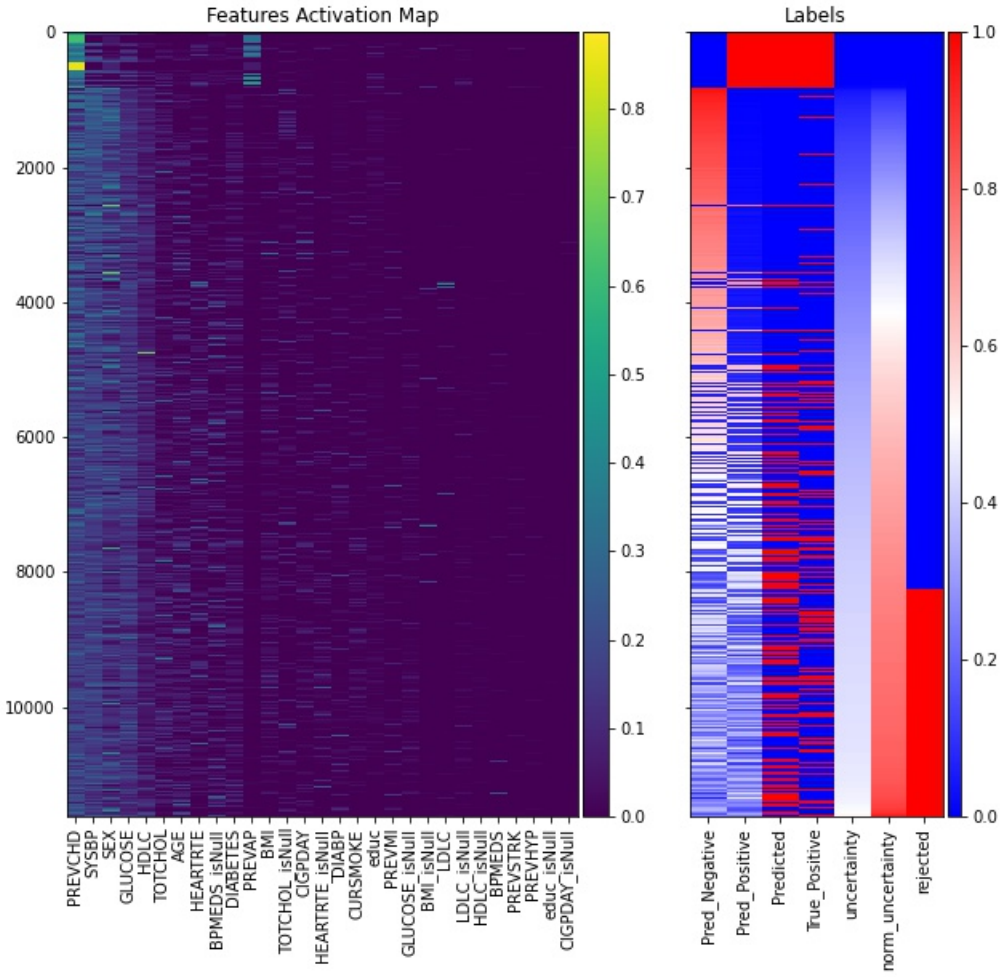


Figure 6.8: Framingham Heart Study rejection option interpretation map with feature activation map and prediction outputs. Features are sorted by importance, and samples are sorted by uncertainty.

are correctly classified as positive. We can identify the dominance of the PREVCHD (prevalent coronary heart disease) feature. In the second section, from samples 1,000 to 5,000, there is a low uncertainty and low error section with samples predominantly classified as negative samples.

From sample 5,000 below, there is a gradual increase in uncertainty and prediction error. This means that the network could correctly identify the features responsible for the prediction error increase. This should not be taken for granted. This is happening because TabNet+SAT has learned to reject while training and not after it due to its special SAT loss. Finally, as we opted for a 70% dataset coverage, samples below 8,000 and beyond are marked as rejected based on a calculated threshold.

An alternative way of understanding is the interpretation results table, with baseline characteristics of each feature. This table counts every categorical feature value and compares the samples by predicted, true_positive (ground truth), and rejection. For numerical features, it gives the mean and standard deviation values. For healthcare pro-

fessionals used to this view, it is possible to identify which feature drives both prediction and rejection. We omitted some irrelevant features for brevity and kept the most relevant ones in Table 6.6.

			Grouped by Predicted		Grouped by True_Positive		Grouped by rejected	
		Overall	0	1	0	1	0	1
n		11627	7496	4131	8469	3158	8253	3374
AGE, mean (SD)		54.8 (9.6)	52.7 (9.3)	58.6 (8.7)	53.8 (9.4)	57.4 (9.5)	54.5 (9.7)	55.4 (9.1)
	0	6598 (56.7)	4364 (58.2)	2234 (54.1)	4804 (56.7)	1794 (56.8)	4784 (58.0)	1814 (53.8)
CURSMOKE, n (%)	**1**	5029 (43.3)	3132 (41.8)	1897 (45.9)	3665 (43.3)	1364 (43.2)	3469 (42.0)	1560 (46.2)
	0	11097 (95.4)	7327 (97.7)	3770 (91.3)	8219 (97.0)	2878 (91.1)	7837 (95.0)	3260 (96.6)
DIABETES, n (%)	**1**	530 (4.6)	169 (2.3)	361 (8.7)	250 (3.0)	280 (8.9)	416 (5.0)	114 (3.4)
DIABP, mean (SD)		83.0 (11.7)	80.4 (10.1)	87.9 (12.7)	82.1 (11.2)	85.6 (12.5)	82.3 (11.8)	84.8 (11.2)
GLUCOSE, mean (SD)		73.7 (36.3)	71.4 (32.6)	77.9 (41.7)	72.5 (33.7)	76.8 (42.3)	73.7 (37.2)	73.6 (33.9)
	0	11000 (94.6)	7496 (100.0)	3504 (84.8)	8469 (100.0)	2531 (80.1)	7626 (92.4)	3374 (100.0)
PREVAP, n (%)	**1**	627 (5.4)		627 (15.2)		627 (19.9)	627 (7.6)	
	0	10785 (92.8)	7496 (100.0)	3289 (79.6)	8469 (100.0)	2316 (73.3)	**7411 (89.8)**	**3374 (100.0)**
PREVCHD, n (%)	**1**	842 (7.2)		842 (20.4)		842 (26.7)	842 (10.2)	
	0	6283 (54.0)	4980 (66.4)	1303 (31.5)	5017 (59.2)	1266 (40.1)	4674 (56.6)	1609 (47.7)
PREVHYP, n (%)	**1**	5344 (46.0)	2516 (33.6)	2828 (68.5)	3452 (40.8)	1892 (59.9)	3579 (43.4)	1765 (52.3)
	0	11253 (96.8)	7496 (100.0)	3757 (90.9)	8469 (100.0)	2784 (88.2)	7879 (95.5)	3374 (100.0)
PREVMI, n (%)	**1**	374 (3.2)		374 (9.1)		374 (11.8)	374 (4.5)	
	0	11475 (98.7)	7436 (99.2)	4039 (97.8)	8388 (99.0)	3087 (97.8)	8142 (98.7)	3333 (98.8)
PREVSTRK, n (%)	**1**	152 (1.3)	60 (0.8)	92 (2.2)	81 (1.0)	71 (2.2)	111 (1.3)	41 (1.2)
	0	5022 (43.2)	2183 (29.1)	2839 (68.7)	3255 (38.4)	1767 (56.0)	**3043 (36.9)**	**1979 (58.7)**
SEX, n (%)	**1**	6605 (56.8)	5313 (70.9)	1292 (31.3)	5214 (61.6)	1391 (44.0)	**5210 (63.1)**	**1395 (41.3)**
SYSBP, mean (SD)		136.3 (22.8)	129.4 (18.5)	148.9 (24.4)	133.6 (21.6)	143.5 (24.4)	135.2 (23.5)	139.1 (20.8)
TOTCHOL, mean (SD)		232.7 (62.9)	225.3 (60.5)	246.1 (64.9)	229.7 (61.7)	240.7 (65.3)	227.7 (66.2)	244.8 (52.3)

Table 6.6: Framingham Heart Study interpretation results.

This table highlights two interesting things. First, the different distribution of PREVCHD=0 samples rejected (3,374 rejected, 7,411 accepted) compared with ground truth distribution (8,169 negatives, 2,316 positives). Secondly, observing the SEX feature, the male/female proportion of the rejected samples is quite different from the accepted samples. These two highlight which predictors are the most confusing for the model. This can be confirmed and better understood with the density plots of Figure 6.9 and 6.10.

In Figure 6.9, we show a density plot of PREVCHD attention $\times$ norm_uncertainty. This figure demonstrates that all samples with a PREVCHD value of 1, in red, have no uncertainty, while samples with a PREVCHD value of zero have weaker but still relevant attention and a wide range of uncertainty, from 0 to 0.8. As 98% of the dataset has a PREVCHD value of 0, we need to further search for the focus of doubt. We do this by analyzing Figure 6.10, a density plot of SEX attention $\times$ norm_uncertainty. Confirming our findings in Table 6.6, there is a significant difference of uncertainty depending on the SEX value, and we see a higher concentration of male patients around 0.8 norm_uncertainty value, while female patients distribution is more widespread. As can be seen in Table 6.6, from a total of 5,022 men, 1,979 were rejected ($\sim$40%), while from a total of 6,605 women, 1,395 ($\sim$20%) were rejected. This indicates that our trained model has difficulties predicting cardiovascular risk for men without PREVCHD. This fact is curious and confirms that risk factors affect men and women differently [46,74]. Although the discussion of risk factors for men and women is beyond the objective of this Master's thesis, our objective here is to demonstrate that with our approach, TabNet+SAT, the trained model now "knows" when it does not know and which features are driving the uncertainty.

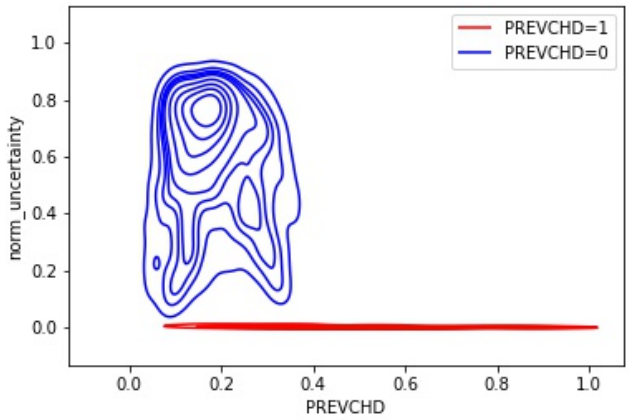


Figure 6.9: Framingham Heart Study PREVCHD attention × Uncertainty indicating high uncertainty for PREVCHD=0 (no prevalent coronary heart disease).

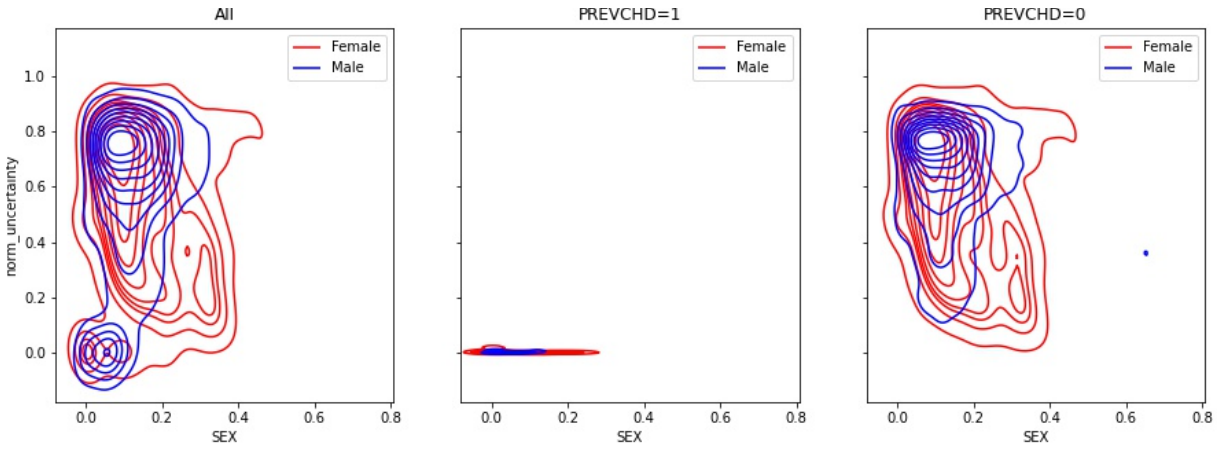


Figure 6.10: Framingham Heart Study PREVCHD × uncertainty segmented by male/female, for patients with (PREVCHD=1) and without (PREVCHD=0) prevalent coronary heart disease.

6.3.2 MI-SIEVE ACC

We created the MI-SIEVE ACC dataset with real-world data collected from hospitals and clinics in Brazil. It has also never been used for ML before. To better know the available data, we started with exploratory experiments. Initially, there were several outcomes in the "evolution" table. For the exploratory experiments, we chose "sucesso_clinico" as it was the outcome that combined all other outcomes. However, we soon realized that we would need to compare our results with baseline risk calculators.

Thus, we decided to perform two experiments on this dataset. The first was the comparison with the baseline risk calculator TIMI [6], and the second used all features available. We could not evaluate TIMI-50 [11] due to an incompatible endpoint, as TIMI-50 has a long-term endpoint, and all outcomes of the MI-SIEVE ACC dataset are short-term.

We also tried to use the GRACE 2.0 calculator, but it was not feasible. As GRACE 2.0 can only be used in acute cases, we had to filter out the chronic cases from the dataset to predict Hard MACE events, resulting in 4,822 samples. Also, this calculator needs 8 input variables: age, heart rate, systolic blood pressure, CHF (congestive heart failure,

given in Killip class[7]), creatinine, ST-segment deviation[8], elevated troponin, and cardiac arrest at admission.

MI-SIEVE ACC does not have 4 out of 8 inputs: CHF, ST-segment deviation, elevated troponin, and cardiac arrest at admission. Hoping that information could be inferred from other variables, a domain specialist analyzed the data, indicating that all dataset samples should be considered as having: elevated troponin = True, cardiac arrest at admission = False, and CHF (Killip class) = Killip I. Thus, 3 of the inputs would be constants. The other inputs, heart rate, systolic blood pressure, and creatinine, had 2,781, 2,626, and 3,479 missing values. To apply data imputation on those values would make no sense at all. How to guess a patient's heart rate? One cannot assume it to be the average of other patients or apply any other statistical technique. Thus, we abandoned the attempt to use GRACE 2.0 due to these limitations.

For both experiments, "TIMI Features" and "All Features", we used the same TabNet hyper-parameters. The only change is the data fed. TabNet hyper-parameters were: $N_d = N_a = 8$, $\lambda_{sparse} = 0.001$, $B = 256$, $B_v = 128$, $m_B = 0.9$, $N_{steps} = 3$, and $\gamma = 1.5$. Random Forest and XGBoost algorithms used the default libraries' hyper-parameters. We also used the same outcome, Hard MACE, a computation of outcomes from the table "evolution", for both experiments. We considered all samples with death or infarction after the intervention as Hard MACE.

MI-SIEVE ACC Experiment: TIMI Features

Following our methodology (Chapter 5), we performed the pre-processing of the TIMI experiment data using the MI-SIEVE ACC dataset. TIMI is a simple computation considering a risk score based on the total count of risk factors (inputs). The inputs are:

- Age greater or equal to 65;
- Three or more of these risk factors: family history of coronary artery disease, hypertension, hypercholesterolemia, Diabetes, or currently smoking;
- Known coronary artery disease;
- Use of aspirin in the past 7 days;
- Severe angina (greater or equal to 2 episodes in the last 24 hours);
- ST-segment deviation;
- Elevated serum cardiac marker (troponin).

[7]Killip is a classification of congestive heart failure (CHF) that independently predicts mortality in patients with myocardial infarction: I. No clinical sign of CHF; II. Presence of rales (crackles) in the lungs, raised jugular venous pressure, or third heart sound (S3 gallop); III. Acute pulmonary edema; IV. Cardiogenic shock.

[8]ST-segment is a segment of the cardiac wave recorded by electrocardiography. A deviation of this part of the wave is a strong indication of myocardial infarction. Troponin is a protein in heart cells released into the bloodstream following cardiac damage. Elevated troponin in the bloodstream on admission strongly predicts mortality and infarct size.

Patients achieved 1 point for yes and 0 points for no. Patients with 0–2 scores are in low-risk groups, 3–4 scores are medium-risk, and 5–7 scores are in high-risk groups. In order to compare TIMI results against a binary classifier algorithm, we had to establish a threshold above which a sample should be considered positive. We used 8% (3 points) as the threshold, as suggested by the domain specialist.

Not all inputs were readily available in the MI-SIEVE ACC dataset. The dataset missed the information about aspirin use, ST-segment deviation, and troponin. With the aid of a domain specialist, we inferred those features from other features. Usually, people with known coronary artery disease use aspirin to prevent infarction. Thus, aspirin used in the past 7 days was assumed true for those with known coronary artery disease. The ST-segment deviation was assumed to be true for patients that received thrombolytic types rt-PA/SK, or with a primary/rescue intervention type. Finally, a high troponin marker was assumed when the intervention type was primary, rescue, or facilitated.

While the pre-processing was satisfactory for this study due to the assumed feature values, the findings of this Master's thesis cannot be used directly in the medical field. Our goal here is to demonstrate the use of machine learning on practical medical problems. A proper risk score could be obtained if the data had all the necessary information.

In addition to the Random Forest and XGBoost baselines used in all other experiments, we also developed a scikit-learn compatible TIMI classifier for this experiment. We followed the details provided by Antman et al. [6]. A risk score from 4.7% to 40.9% is computed based on the total count of risk factors. For binary classification, we established a threshold in which the outcome for a Hard MACE is considered positive above a particular risk value. We used 8% as the threshold, as suggested by the domain specialist. We also used Logistic Regression as the baseline. TIMI counts risk factors without weight those factors. Eventually, a logistic regressor found better weights and obtained better results. We applied logistic regression only in this experiment.

Experiment results achieved a very low AUC for all algorithms. The average AUC and standard deviation are shown in Table 6.7. TabNet and Logistic Regression had a better performance than the TIMI calculator. This low AUC demonstrates that only considering the 7 risk factors is inadequate for a better risk prediction.

Using TabNet+SAT, without pre-training, there was no increase in performance by sacrificing coverage. With pre-training, AUC increased from 0.634 at 100% coverage to 0.663 at 70% coverage. This could indicate that TabNet is suffering from a small dataset, with only 7 features and 4,821 samples, of which only 176 are positive (3.6%). After the 5-fold cross-validation (twice), the dataset is even smaller, with 3,214 training samples and 1,607 validation samples, for each fold.

This experiment demonstrates a TabNet+SAT limitation. SAT is also based on the learning capacity of the deep neural network. If the deep neural network does not learn how to classify samples, it is possible to learn how to reject them. SAT can only improve performance where there is some improvement. Despite that, it is interesting to observe that most predictions have, in fact, high uncertainty values. TabNet+SAT could not figure out what exactly it was, but it knew something was wrong.

Table 6.7: MI-SIEVE ACC AUC results averaged of all 10 runs for Random Forest, XGBoost, TabNet, and TabNet+SAT for various coverage rates. TabNet+SAT-c% stands for TabNet+SAT under c coverage rate.

	AUC	
	TIMI Features	All Features
TIMI	0.599 ± 0.023	–
Logistic Regression	0.629 ± 0.012	–
Random Forest	0.596 ± 0.031	0.807 ± 0.031
XGBoost	0.599 ± 0.033	0.806 ± 0.028
TabNet	0.634 ± 0.032	0.822 ± 0.023
TabNet+SAT-100%	0.634 ± 0.025	0.803 ± 0.043
TabNet+SAT-90%	0.640 ± 0.028	0.809 ± 0.050
TabNet+SAT-80%	0.644 ± 0.030	0.820 ± 0.054
TabNet+SAT-70%	0.663 ± 0.021	0.827 ± 0.063

MI-SIEVE ACC Experiment: All Features

This experiment is the most crucial experiment of this work, as it applies our approach in the real world challenging data. In order to obtain a better risk calculator, we did not limit ourselves to the 7 TIMI features in the second experiment with the MI-SIEVE ACC dataset. Instead, we used all 207 features. Those features were obtained by cleaning, encoding, and filling in the missing information. Each data preparation operation was discussed with the domain specialist during several meetings. In this experiment, we considered all samples, regardless of acute/chronicle cases. We aim to let the model indicate what is going on and further check with the domain specialist, the opposite as most commonly done.

Table 6.7 shows that the performance significantly increased compared to the TIMI experiment. Regarding TabNet+SAT, we confirmed that better performance was achieved by sacrificing coverage. TabNet+SAT average AUC was 0.803 (±0.043) at 100% coverage and increased to 0.827 (±0.063) at 70% coverage.

To interpret the results, we again used the aggregated TabNet activation map, side by side with the results of the predictions. The global interpretation is made using the interpretation map in Figure 6.11. The horizontal axis contains the features ordered by global importance, and, from left to right, the most important features were "complicacao_Grau - Complicação_Grave_sum", "stent_Timi pós_3.0_mean" and "procedimento_Duração Interv. (min)". Those features are engineered features that aggregate values for the auxiliary tables "complicação" and "stent", as there can be more than one complication on the same intervention. There can also be more than one stent used in one intervention. So, to keep a tabular, single-row format, we used the sum, mean and minimal values to aggregate those features. This means that the model pays attention to the total of severe complications and the mean TIMI flow value for all stents used. TIMI 3.0 means a normal flow of blood after the coronary obstruction. We can infer that the interventions that had severe complications and/or did not have a normal flow of blood are the ones that

will most likely lead to a Hard MACE outcome. The total intervention time in minutes is also a relevant feature, and we can interpret that interventions that take a long time are the ones that are treating the most severe cases and thus have a greater chance of leading to Hard MACE after the interventions. This is also confirmed by the density plot in Figure 6.12. The vertical axis contains the samples, ranging from 0 to 9,635 samples, sorted by the "uncertainty" column. The left area ("Features Activation Map") can be understood as a heatmap of features importance, and each pair feature x sample contains the attention given by the model to that feature for that instance. We use a scale from dark blue, meaning low attention, to yellow, meaning high attention. The right area (Labels) shows the prediction results using the blue-red scale.

We can notice that in the top region of the heatmap, many samples were Predicted and True_Prositive match, meaning that the model predicts correctly. We can infer that this is due to higher attention given for Complicação_Grave (severe complication) feature. The rest of the heatmap reveals an increasing number of false positives, especially at the bottom region, with samples ranging from 6,000 to 9,635. The attention on the most important features decreases while the attention on Diabetes features increases. Although Diabetes is a risk factor, the interpretation demonstrates that the model overestimates it.

Our approach also permits local interpretation. Four patients were selected, covering four different cases:

1. Patient 560: correctly classified as positive Hard MACE, with virtually no uncertainty;
2. Patient 5412: wrongly classified as positive Hard MACE, but not rejected due to a small uncertainty;
3. Patient 8048: wrongly classified as positive Hard MACE, but rejected due to high uncertainty;
4. Patient 6336: correctly classified as positive Hard MACE, but rejected due to high uncertainty.

Let us first interpret patient 560. Prediction results with probabilities, true label, uncertainty, and rejection results are found in Table 6.8. Features values with the respective attention are found in Table 6.9. As we can see, TabNet+SAT is paying attention and deciding based on three main features. Patient 560 had a severe complication, and the normal TIMI flow (3.0) was zero, meaning that the intervention could not reestablish the blood flow. Also, the intervention took 180 minutes, much higher than the average. TabNet+SAT does not doubt that those combinations impose a very high risk of a Hard MACE event.

The second patient is patient 5412, a false positive patient wrongly classified as positive Hard MACE, but not rejected due to a small uncertainty. Prediction results with probabilities, true label, uncertainty, and rejection results are also found in Table 6.8. Features values with the respective attention are found in Table 6.10. The model predicted with very low uncertainty that this patient had a high risk of Hard Mace. This is also due to a severe complication and no normal TIMI flow (3.0). The third most impor-

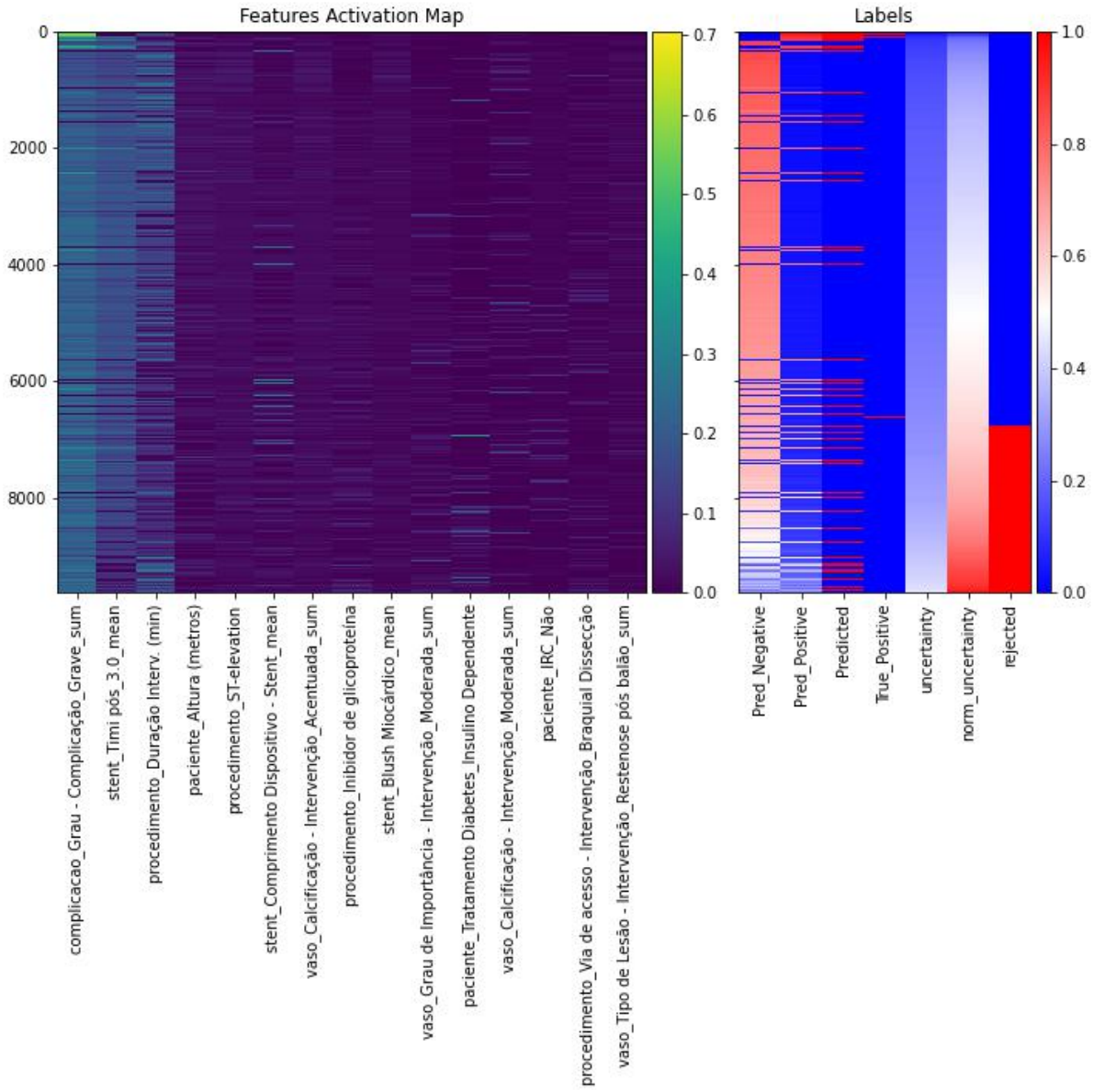


Figure 6.11: Interpretation map of whole MI-SIEVE-ACC dataset rejection option with feature activation map and prediction outputs. Features are sorted by importance, and samples are sorted by uncertainty.

Table 6.8: Prediction results of some patients.

	Patient ID			
	560 (true positive)	5412 (false positive)	8048 (rejected false positive)	6336 (rejected true negative)
Pred_Negative	0.000001	0.000133	0.229612	0.304241
Pred_Positive	0.997144	0.986709	0.299017	0.220693
Predicted	1.000000	1.000000	1.000000	0.000000
True_Positive	1.000000	0.000000	0.000000	0.000000
uncertainty	0.002855	0.013157	0.471372	0.475066
norm_uncertainty	0.000000	0.021816	0.992176	1.000000
rejected	0.000000	0.000000	1.000000	1.000000

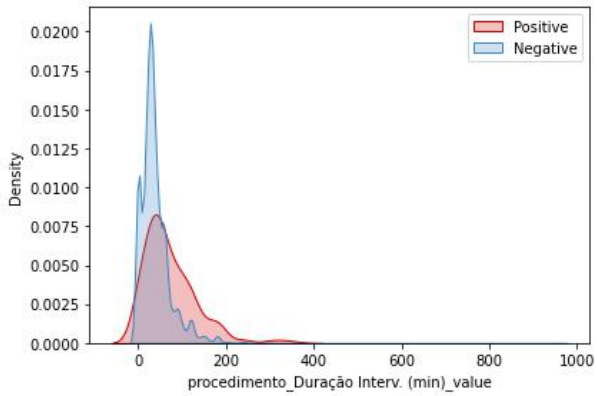


Figure 6.12: MI-SIEVE ACC density plot of the intervention time feature.

tant feature is MDRD[9], which is calculated based on patient creatinine and is a measure of the kidney filtration rate. Although this value is abnormally high, according to the domain specialist, there is no correlation with a MACE event. If the value was very low, this could indicate kidney failure and, thus a higher risk. TabNet+SAT overestimated the risk based on the complication and TIMI flow, and maybe got confused with MDRD. This example demonstrated that a possible enhancement of TabNet+SAT could be the direction of the relationship of the predictor with the outcome, defining if the outcome probability increases when the feature value increases or it goes in the opposite direction.

The third patient is patient 8048, who was wrongly classified as positive Hard MACE but rejected due to high uncertainty. Prediction results with probabilities, true label, uncertainty, and rejection results are found in Table 6.8. Features values with the respective attention are found in Table 6.11. The model prediction was ambiguous, with a positive class probability almost equal to the negative class probability. Although we have TIMI flow (3.0), indicating complete perfusion of blood after the intervention, which means a successful intervention, and also no severe complications, we see the highest attention paid to the feature that indicates that the patient has Diabetes and under treatment. This feature is probably a source of confusion for the model. Nevertheless, the model learned how to identify it and correctly attributed a high uncertainty.

The last patient is patient 6336, correctly classified as positive Hard MACE but rejected due to high uncertainty. Prediction results with probabilities, true label, uncertainty, and rejection results are found in Table 6.8. Features values with the respective attention are found in Table 6.12. The rejection option can cause correctly classified samples to be rejected when the model sees a pattern of high uncertainty. The features, in fact, indicate a successful intervention, with normal TIMI flow (3.0), no severe complications, and close to average intervention time. Diabetes seems to also bring confusion to the model along with the coronary vessel, namely, DPD.

Those results were interpreted in conjunction with the domain specialist MD. Luiz Sérgio Fernandes de Carvalho and TabNet+SAT results satisfactorily shed light on model interpretation. Feature values and importance do present reasonable explanations according to medical practice.

[9] `http://www.mdrd.com`

Table 6.9: Features values and attention of patient 560 (true positive instance).

Feature	Value	Attention (%)
stent_Timi pós_0.0_mean	1.0	31.98
procedimento_Duração Interv. (min)	180.0	26.78
complicacao_Grau - Complicação_Grave_sum	1.0	17.71
stent_Timi pós_3.0_mean	0.0	13.04
paciente_IRC_Não	1.0	1.72
stent_Timi pós_1.0_max	0.0	1.61
vaso_Calcificação - Intervenção_Discreta_sum	0.0	1.32
vaso_Calcificação - Intervenção_Acentuada_sum	0.0	1.21
paciente_Tipo Antecedentes Familiares_Precoce	1.0	0.84
paciente_Clearance Cr (ml/min)	0.0	0.82
Top 10 Total		97.04

Table 6.10: Features values and attention of patient 5412 (false positive instance).

Feature	Value	Attention (%)
complicacao_Grau - Complicação_Grave_sum	1.0	28.19
stent_Timi pós_3.0_mean	0.0	19.77
paciente_MDRD (ml/min/173m^2)	182.0	18.15
paciente_Temperatura (^{o}C)	36.0	7.10
vaso_Angulação - Intervenção_gt90_sum	0.0	4.06
balao_Timi pós_mean	3.0	3.70
vaso_Tipo de Lesão - Focal - Intervenção_Multif...	0.0	2.54
vaso_Local da Lesão - Intervenção_Óstio do enxe...	0.0	2.25
paciente_Intervenção Percutânea Prévia	1.0	1.93
complicacao_Tipo Complicação_Neurológica_sum	0.0	1.76
Top 10 Total		89.44

Table 6.11: Features values and attention of patient 8048 (rejected false positive).

Feature	Value	Attention (%)
paciente_Tratamento Diabetes_Insulino Dependente	1.0	24.58
complicacao_Grau - Complicação_Grave_sum	0.0	17.53
stent_Timi pós_3.0_mean	1.0	8.03
procedimento_Duração Interv. (min)	90.0	5.73
paciente_Altura (metros)	1.5	5.58
stent_Grau de estenose pós_mean	0.0	5.34
complicacao_Tipo Complicação_Hematoma_sum	0.0	3.65
procedimento_Via de acesso - Intervenção_Braqui...	0.0	3.40
stent_Pressão Final de Liberação (ATM)_min	9.0	3.33
complicacao_Tipo Complicação_Arritmia_sum	0.0	3.14
Top 10 Total		80.31

Table 6.12: Features values and attention of patient 6336 (rejected true positive instance).

Feature	Value	Attention (%)
vaso_Vaso coronário - Intervenção_DPD_sum	1.0	40.24
paciente_Tratamento Diabetes_Insulino Dependente	1.0	18.96
complicacao_Grau - Complicação_Grave_sum	0.0	9.67
procedimento_Duração Interv. (min)	120.0	8.52
stent_Pressão Final de Liberação (ATM)_min	14.0	4.00
stent_Timi pós_3.0_min	1.0	3.52
stent_Grau de estenose pós_mean	0.0	2.35
complicacao_Tipo Complicação_Hematoma_sum	0.0	1.87
balao_Pressão Final (ATM)_min	0.0	1.83
paciente_AVC prévio	0.0	1.43
Top 10 Total		92.40

Chapter 7

Conclusions

7.1 Motivation and Results

This Master's thesis is motivated by applying advanced machine learning techniques to predict the risk of Major Adverse Cardiovascular Events (MACE). Despite already many works proposing machine learning for health risk prediction, adopting this technology is impractical because those approaches are "black boxes". We identified the opportunity to employ an interpretable deep neural network, TabNet [8], and enhance this network with a rejection option approach (Self Adaptive Training (SAT) [33]), informing data scientists and medical doctors when the prediction is not reliable, which, as far as we know, is the first time this attempt was made on tabular data.

We evaluated our TabNet+SAT on two medical datasets, Framingham Heart Study (FHS) [1] and Myocardial ISchemIa prognostic EValuation AngioCardio-Clarity (MI-SIEVE ACC) (collected from real hospitals and clinic data from Brazil). In both datasets, we achieved better results than the baseline. We obtained 0.760 AUC on Random Forest in the FHS dataset and 0.764 on TabNet. Using TabNet+SAT, the final AUC at 70% coverage rate was 0.815. In the MI-SIEVE ACC dataset, we obtained an AUC of 0.807 for Random Forest and 0.822 for TabNet. Using TabNet+SAT, the final AUC at 70% coverage was 0.827. To ensure that the results were consistent, we also confirmed our results on four non-medical datasets, and the results are detailed in Chapter 6.

We further demonstrated that given TabNet interpretation capabilities, not only can the prediction be interpreted but also the rejection of samples by the SAT technique. This approach can be used to give data scientists or domain specialists insights into *when* and *why* the model cannot learn to classify a subset of samples. We also showed that this interpretation could be used both globally and locally. Global interpretation proved a great tool for data scientists to assess their models' quality and find biases, although the latter was not originally intended. Local interpretation can be a powerful tool for users of such TabNet+SAT models, such as medical doctors, enabling them to evaluate the model uncertainty and make an informed decision if they should or not trust the model. As TabNet+SAT can give relative importance to each feature for a given sample at inference time, medical doctors can also be informed and understand based on which features the model is computing its prediction.

7.2 Answers to the Research Questions

The research questions were fundamental to giving direction to this work. The questions and their respective answers are listed in the following.

Q1. What is the impact of using the deep neural network architecture TabNet on the risk prediction of cardiovascular diseases compared with the conventional risk scores?

Due to restrictions imposed by the available data, we evaluated only the performance of TabNet against the TIMI risk calculator. We found the performance of all tested models is awful, using only the 7 features required by the TIMI calculator. TIMI AUC was 0.599, while TabNet achieved 0.634. We obtained a much better result using all features of the MI-SIEVE ACC dataset, not only those required by the TIMI calculator. This demonstrates that, despite TabNet performing better than TIMI using 7 features, its full capability is shown when using the maximum amount of data available. The AUC obtained was 0.822.

Q2. How to integrate a selective classifier/rejection option on TabNet? What is the impact on the results?

We demonstrated that it is not only possible to integrate the rejection option on TabNet but also that it consistently improved the performance in all experiments. We achieved AUC gains from 2 to 6 percentage points at a 70% coverage rate.

7.3 Future Work

In future work, we intend to compare TabNet's activation map interpretation with state-of-the-art techniques such as SHAP (SHapley Additive exPlanations) [45] and LIME (Local Interpretable Model-Agnostic) [61]. We also plan to investigate if those tools could work with rejection interpretation, as our approach does.

Very recent works are presenting alternatives to TabNet, such as GATE (Gated Additive Tree Ensemble) [36], NBM (Neural Basis Model) [58], Hopular (Modern Hopfield Networks for Tabular Data) [65], FT-Transformer (Feature Tokenizer Transformer) [25], and SAINT (Self-Attention and Intersample Attention Transformer) [68]. We intend to evaluate the Self Adaptive Training on those deep neural networks.

DNNs also can enable transfer learning, while traditional tree-based algorithms are always trained from scratch. This advantage was not explored in this Master's thesis as the use of DNN for tabular data is very recent and not yet established. Nevertheless, a recent work [42] has explored this advantage in other DNN architectures.

Although we evaluated our TabNet+SAT on 4 non-medical and 2 medical datasets, recent research [27] has appointed that tree-based models still outperform DNNs on tabular data on 45 datasets. The paper also gives insights into why tree-based models thrive, being uninformative features the primary reason. However, the mentioned paper did not conduct experiments using TabNet. With our experiment pipeline, we can evaluate it

on the 45 datasets. That will be an excellent chance to show that the attention mechanism of TabNet can filter out uninformative features, while SAT can reject ambiguous or inconclusive samples, proving to be more robust than traditional tree-based models.

Bibliography

[1] Framingham Heart Study. https://framinghamheartstudy.org/fhs-about. Online; accessed 27 Feb 2022. 29, 31, 65

[2] Composição racial no Brasil - IBGE. https://educa.ibge.gov.br/jovens/conheca-o-brasil/populacao/18319-cor-ou-raca.html, 2020. 31

[3] Kaggle Competition: COVID19 Global Forecasting. https://www.kaggle.com/c/covid19-global-forecasting-week-5, 2020. 14

[4] Kaggle Competition: RSNA Intracranial Hemorrhage Detection. https://www.kaggle.com/c/rsna-intracranial-hemorrhage-detection, 2020. 14

[5] Kaggle Competition: RSNA STR Pulmonary Embolism Detection. https://www.kaggle.com/c/rsna-str-pulmonary-embolism-detection, 2020. 14

[6] Elliott M Antman, Marc Cohen, Peter JLM Bernink, Carolyn H McCabe, Thomas Horacek, Gary Papuchis, Branco Mautner, Ramon Corbalan, David Radley, and Eugene Braunwald. The timi risk score for unstable angina/non–st elevation mi: a method for prognostication and therapeutic decision making. *Jama*, 284(7):835–842, 2000. 13, 27, 56, 58

[7] MARK A APPLEBY, BRAD G ANGEJA, KENT DAUTERMAN, and C MICHAEL GIBSON. Angiographic assessment of myocardial perfusion: Timi myocardial perfusion (tmp) grading system. *Heart*, 86(5):485–486, 2001. 19

[8] Sercan Ö Arik and Tomas Pfister. TabNet: Attentive interpretable tabular learning. In *AAAI Conference on Artificial Intelligence*, pages 6679–6687, 2021. 14, 19, 21, 22, 24, 27, 39, 41, 43, 45, 46, 65

[9] Firdaus Aziz, Sorayya Malek, Khairul Shafiq Ibrahim, Raja Ezman Raja Shariff, Wan Azman Wan Ahmad, Rosli Mohd Ali, Kien Ting Liu, Gunavathy Selvaraj, and Sazzli Kasim. Short-and long-term mortality prediction after an acute st-elevation myocardial infarction (stemi) in asians: A machine learning approach. *PloS one*, 16(8):e0254894, 2021. 27

[10] Pierre Baldi, Peter Sadowski, and Daniel Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature communications*, 5(1):1–9, 2014. 35

[11] Erin Bohula, Marc Bonaca, Eugene Braunwald, Philip Aylward, Ramon Corbalan, Gaetano De Ferrari, Ping He, Basil Lewis, Piera Merlini, Sabina Murphy, Marc Sabatine, Benjamin Scirica, and David Morrow. Atherothrombotic risk stratification and the efficacy and safety of vorapaxar in patients with stable ischemic heart disease and prior myocardial infarction. *Circulation*, 134:CIRCULATIONAHA.115.019861, 07 2016. 13, 27, 56

[12] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. 14, 15, 53

[13] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 785–794, 2016. 14, 15, 53

[14] Chi-Keung Chow. An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*, (4):247–254, 1957. 24

[15] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972. 13

[16] Jan Salomon Cramer. The origins of logistic regression. 2002. 13

[17] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International conference on machine learning*, pages 933–941. PMLR, 2017. 20

[18] LS de Carvalho, Rebeca Gouget Sergio Miranda, Silvio Gioppato, M Fernandez, Bernardo Carvalho Trindade, JC Quinaglia e Silva, Sandra Avila, and Andrei Sposito. Pcv20 machine learning better predicts risk after acute coronary syndromes, identifies high-cost individuals and those with elevated burden of uncontrolled risk factors. *Value in Health*, 22:S545, 2019. 27

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019. 14

[20] Dheeru Dua, Casey Graff, et al. Uci machine learning repository. 2017. 20, 33, 35, 36

[21] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. The kdd process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11):27–34, 1996. 16, 39

[22] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé, and Kate Crawford. Datasheets for datasets, 2018. 31

[23] Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *International Conference on Machine Learning (ICML)*, pages 2151–2159, 2019. 24

[24] Chuanxing Geng, Sheng-jun Huang, and Songcan Chen. Recent advances in open set recognition: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3614–3631, 2020. 23, 24

[25] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems*, 34:18932–18943, 2021. 28, 66

[26] Christopher B. Granger, Robert J. Goldberg, Omar Dabbous, Karen S. Pieper, Kim A. Eagle, Christopher P. Cannon, Frans Van de Werf, Álvaro Avezum, Shaun G. Goodman, Marcus D. Flather, Keith A. A. Fox, and for the Global Registry of Acute Coronary Events Investigators. Predictors of Hospital Mortality in the Global Registry of Acute Coronary Events. *Archives of Internal Medicine*, 163(19):2345–2353, 10 2003. 13, 27

[27] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on tabular data? *arXiv preprint arXiv:2207.08815*, 2022. 66

[28] Malay Haldar, Mustafa Abdool, Prashant Ramanathan, Tao Xu, Shulin Yang, Huizhong Duan, Qing Zhang, Nick Barrow-Williams, Bradley C Turnbull, Brendan M Collins, et al. Applying deep learning to airbnb search. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1927–1935, 2019. 27

[29] Alon Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12, 2009. 15

[30] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014. 14

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 20

[32] Lang Huang, Chao Zhang, and Hongyang Zhang. Self-adaptive training: beyond empirical risk minimization. *arXiv preprint arXiv:2002.10319*, 2020. 14

[33] Lang Huang, Chao Zhang, and Hongyang Zhang. Self-adaptive training: beyond empirical risk minimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 19365–19376, 2020. 23, 24, 40, 65

[34] Mikael Huss. Tabular Data and Deep Learning: Where Do We Stand? 14, 27

[35] Ankush Jamthikar, Deep Gupta, Luca Saba, Narendra N Khanna, Tadashi Araki, Klaudija Viskovic, Sophie Mavrogeni, John R Laird, Gyan Pareek, Martin Miner, et al. Cardiovascular/stroke risk predictive calculators: a comparison between statistical and machine learning models. *Cardiovascular Diagnosis and Therapy*, 10(4):919, 2020. 27

[36] Manu Joseph and Harsh Raj. Gate: Gated additive tree ensemble for tabular classification and regression. *arXiv preprint arXiv:2207.08548*, 2022. 28, 66

[37] Ioannis A Kakadiaris, Michalis Vrigkas, Albert A Yen, Tatiana Kuznetsova, Matthew Budoff, and Morteza Naghavi. Machine learning outperforms acc/aha cvd risk calculator in mesa. *Journal of the American Heart Association*, 7(22):e009476, 2018. 27

[38] S Kasim, S Malek, KS Ibrahim, JH Hiew, and MF Aziz. Acs mortality prediction in asian in-hospital patients with deep learning using machine learning feature selection. *European Heart Journal*, 42(Supplement_1):ehab724–3069, 2021. 27

[39] SS Kasim, S Malek, KKS Ibrahim, and MF Aziz. Risk stratification of asian patients after st-elevation myocardial infarction using machine learning methods. *European Heart Journal*, 41(Supplement_2):ehaa946–3494, 2020. 27

[40] Ronny Kohavi and Barry Becker. Data mining and visualization. silicon graphics. *Extraction done by Barry Becker from the database of 1994 Census*, 1994. 36

[41] Christine Lee. *Developing Predictive Models for Risk of Postoperative Complications and Hemodynamic Instability in Patients Undergoing Surgery*. PhD thesis, UC Irvine, 2019. 27

[42] Roman Levin, Valeriia Cherepanova, Avi Schwarzschild, Arpit Bansal, C Bayan Bruss, Tom Goldstein, Andrew Gordon Wilson, and Micah Goldblum. Transfer learning with deep tabular models. *arXiv preprint arXiv:2206.15306*, 2022. 66

[43] Gary Lincoff. *The National Audubon Society Field Guide to Mushrooms*. Knopf, 1981. 33

[44] Ziyin Liu, Zhikang Wang, Paul Pu Liang, Russ R Salakhutdinov, Louis-Philippe Morency, and Masahito Ueda. Deep gamblers: Learning to abstain with portfolio theory. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 24

[45] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc. 20, 27, 47, 66

[46] Rosalinda Madonna, Carmela Rita Balistreri, Salvatore De Rosa, Saverio Muscoli, Stefano Selvaggio, Giancarlo Selvaggio, Péter Ferdinandy, and Raffaele De Caterina. Impact of sex differences and diabetes on coronary atherosclerosis and ischemic heart disease. *Journal of Clinical Medicine*, 8(1), 2019. 55

[47] Andre Martins and Ramon Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International Conference on Machine Learning*, pages 1614–1623, 2016. 23, 28

[48] José Mena, Oriol Pujol, and Jordi Vitrià. A survey on uncertainty estimation in deep learning classification systems from a bayesian perspective. *ACM Computing Surveys (CSUR)*, 54(9):1–35, 2021. 23

[49] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences, 2017. 14

[50] David Moher, Alessandro Liberati, Jennifer Tetzlaff, and Douglas G Altman. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ*, 339, 2009. 26

[51] Gláucia Maria Moraes de Oliveira, Luisa Campos Caldeira Brant, Carisi Anne Polanczyk, Andreia Biolo, Bruno Ramos Nascimento, Deborah Carvalho Malta, Maria de Fatima Marinho de Souza, Gabriel Porto Soares, Gesner Francisco Xavier Junior, M. Julia Machline-Carrion, Marcio Sommer Bittencourt, Octavio M. Pontes Neto, Odilson Marcos Silvestre, Renato Azeredo Teixeira, Roney Orismar Sampaio, Thomaz A. Gaziano, Gregory A. Roth, and Antonio Luiz Pinho Ribeiro. Estatística Cardiovascular Brasil 2020. *Arquivos Brasileiros de Cardiologia*, 115:308 – 439, 09 2020. 13, 31

[52] Saarang Panchavati, Carson Lam, Nicole S Zelin, Emily Pellegrini, Gina Barnes, Jana Hoffman, Anurag Garikipati, Jacob Calvert, Qingqing Mao, and Ritankar Das. Retrospective validation of a machine learning clinical decision support tool for myocardial infarction risk stratification. *Healthcare technology letters*, 8(6):139, 2021. 27

[53] Ben Peters, Vlad Niculae, and André FT Martins. Sparse sequence-to-sequence models. *arXiv preprint arXiv:1905.05702*, 2019. 28

[54] Sergei Popov, Stanislav Morozov, and Artem Babenko. Neural oblivious decision ensembles for deep learning on tabular data. In *International Conference on Learning Representations*, 2020. 27, 28

[55] Ishan Poudel, Chavi Tejpal, and Nusrat Jahan Hamza Rashid. Major adverse cardiovascular events: an inevitable outcome of st-elevation myocardial infarction? a literature review. *Cureus*, 11(7), 2019. 13

[56] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, pages 6638–6648, 2018. 43

[57] Giorgio Quer, Ramy Arnaout, Michael Henne, and Rima Arnaout. Machine learning and the future of cardiovascular care: Jacc state-of-the-art review. *Journal of the American College of Cardiology*, 77(3):300–313, 2021. 27

[58] Filip Radenovic, Abhimanyu Dubey, and Dhruv Mahajan. Neural basis models for interpretability. *arXiv preprint arXiv:2205.14120*, 2022. 28, 66

[59] Anthony Ralston and Herbert S Wilf. Mathematical methods for digital computers. Technical report, 1960. 13

[60] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier, 2016. 20

[61] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you? explaining the predictions of any classifier. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016. 66

[62] Tim Sabsch. *Visual Model Interpretation for Epidemiological Cohort Studies*. PhD thesis, University of Magdeburg, 2018. 27

[63] Fatima Sanchez-Cabo, Xavier Rossello, Valentin Fuster, Fernando Benito, Jose Pedro Manzano, Juan Carlos Silla, Juan Miguel Fernández-Alvira, Belen Oliva, Leticia Fernandez-Friera, Beatriz Lopez-Melgar, et al. Machine learning improves cardiovascular risk definition for young, asymptomatic individuals. *Journal of the American College of Cardiology*, 76(14):1674–1685, 2020. 27

[64] Ashish Sarraju, Andrew Ward, Sukyung Chung, Jiang Li, David Scheinker, and Fàtima Rodríguez. Machine learning approaches improve risk stratification for secondary cardiovascular disease prevention in multiethnic patients. *Open Heart*, 8(2), 2021. 27

[65] Bernhard Schäfl, Lukas Gruber, Angela Bitto-Nemling, and Sepp Hochreiter. Hopular: Modern hopfield networks for tabular data. *arXiv preprint arXiv:2206.00664*, 2022. 28, 66

[66] Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772, 2012. 14

[67] Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, 2022. 14

[68] Gowthami Somepalli, Micah Goldblum, Avi Schwarzschild, C Bayan Bruss, and Tom Goldstein. Saint: Improved neural networks for tabular data via row attention and contrastive pre-training. *arXiv preprint arXiv:2106.01342*, 2021. 28, 66

[69] Kristian Thygesen, Joseph S Alpert, Allan S Jaffe, Bernard R Chaitman, Jeroen J Bax, David A Morrow, Harvey D White, and Executive Group on behalf of the Joint European Society of Cardiology (ESC)/American College of Cardiology (ACC)/American Heart Association (AHA)/World Heart Federation (WHF) Task Force for the Universal Definition of Myocardial Infarction. Fourth universal definition of myocardial infarction (2018). *Journal of the American College of Cardiology*, 72(18):2231–2264, 2018. 18

[70] Adam Timmis, Nick Townsend, Chris P Gale, Aleksandra Torbica, Maddalena Lettino, Steffen E Petersen, Elias A Mossialos, Aldo P Maggioni, Dzianis Kazakiewicz, Heidi T May, Delphine De Smedt, Marcus Flather, Liesl Zuhlke, John F Beltrame, Radu Huculeci, Luigi Tavazzi, Gerhard Hindricks, Jeroen Bax, Barbara Casadei, Stephan Achenbach, Lucy Wright, Panos Vardas, and European Society of Cardiology. European Society of Cardiology: Cardiovascular Disease Statistics 2019. *European Heart Journal*, 41(1):12–85, 12 2019. 13

[71] Eduardo Valle, Michel Fornaciali, Afonso Menegola, Julia Tavares, Flávia Vasques Bittencourt, Lin Tzy Li, and Sandra Avila. Data, depth, and design: Learning reliable models for skin lesion analysis. *Neurocomputing*, 383:303–313, 2020. 14

[72] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. 22

[73] Salim Virani, Alvaro Alonso, Emelia Benjamin, Márcio Bittencourt, Clifton Callaway, April Carson, Alanna Chamberlain, Alex Chang, Susan Cheng, Francesca Delling, Luc Djoussé, Mitchell Elkind, Jane Ferguson, Myriam Fornage, Sadiya Khan, Brett Kissela, Kristen Knutson, Tak Kwan, Daniel Lackland, and Connie Tsao. Heart disease and stroke statistics—2020 update: A report from the american heart association. *Circulation*, 141, 01 2020. 13

[74] Laura Young and Leslie Cho. Unique cardiovascular risk factors in women. *Heart*, 105(21):1656–1660, 2019. 55

Appendix A

MI-SIEVE ACC Datasheet

A.1 Motivation

For what purpose was the dataset created? (Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.) This dataset serves as a registry of cardiac catheterization procedures across Brazil, including patient-level data regarding clinical information, exam finding, procedures, and outcomes. The dataset was funded by SBHCI (Sociedade Brasileira de Hemodinâmica e Cardiologia Intervencionista), and MI-SIEVE ACC represents a high-quality subset of the full original dataset, including only exams from Angiocardio labs. This particular subset aims to enable research in cardiovascular risk prediction. Current cardiovascular risk calculators are based on cohort studies with no information regarding percutaneous coronary intervention. The intention was to produce a better risk calculator using many attributes currently collected from real patients by hospitals and clinics in Brazil.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)? This dataset was created by MD Silvio Giopatto, Prof MD Luiz Sérgio Fernandes de Carvalho, Prof. Sandra Eliza Fontes de Avila, and MSc. student Tito Barbosa Rezende. At the time of creation, Silvio Giopatto was a senior researcher at SBHCI (Sociedade Brasileira de Hemodinâmica e Cardiologia Intervencionista), Luiz Sérgio was the founder of the company Clarity Health and a researcher at the University of Campinas. Sandra Avila was a researcher and professor at the University of Campinas and supervisor of the M.Sc. student Tito Barbosa Rezende.

Who funded the creation of the dataset? (If there is an associated grant, please provide the name of the grantor and the grant name and number.) The dataset was created by SBHCI (Sociedade Brasileira de Hemodinâmica e Cardiologia Intervencionista) with its funding.

Any other comments? None.

A.2 Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? (Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.) Each instance is a patient who was submitted to percutaneous coronary intervention. Attributes are composed of both patients' characteristics and intervention information. The dataset is labeled after the outcome of the intervention, describing if the patient had or did not have a MACE, a major adverse cardiovascular event.

How many instances are there in total (of each type, if appropriate)? The dataset consists of 9,635 patients.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? (If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances because instances were withheld or unavailable).) It is a sample of all possible patients. We cannot assume that every possible instance is contained in the dataset, and we expect the contrary, that there are the least typical patients and conditions not contained in this dataset. A larger representative set of the dataset is not present because instances were unavailable.

What data does each instance consist of? ("Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.) Each instance consists of patients/intervention features already cleaned and pre-processed. We provide here the list of all features in Table A.1, in Portuguese and English:

Table A.1: MI-SIEVE ACC list of features in Portuguese and English.

Feature (Portuguese)	**Feature (English)**
evolucao_Hard_MACE	evolution_hard_mace
paciente_Sexo	sex
paciente_Idade no procedimento	age at intervention
paciente_Procedência	origin
paciente_HAS	hypertension
paciente_Diabetes	patient_diabetes
paciente_Dislipidemia	patient_dislipidemia
paciente_Tabagista	patient_tabagist
paciente_Cigarros (dia) - Tabagista	patient_cigarrets per day- smoking
paciente_Tempo	years of smoking
paciente_Ex-Tabagista	patient_ex-tabagist
paciente_Cigarros (dia) - Ex-Tabagista	patient_cigarrets per day- former tabagist
paciente_Tempo de Interrupção	years since quit smoking
paciente_IAM Prévio	previous myocardial infarction
paciente_Síndrome Metabólica	metabolic syndrome
paciente_Antecedentes Familiares	family history
paciente_AVC prévio	previous stroke
paciente_História de Trombose Venosa Profunda	patient_ deep venous thrombosis history

paciente_DPOC	chronic obstructive pulmonary disease
paciente_Alergia ao iodo	iodine allergy
paciente_DAC conhecida	known cad
paciente_Cate Prévio	previous pci
paciente_Cirurgia de Revascularização Prévia	previous revascularization surgery
paciente_Intervenção Percutânea Prévia	previous percutaneous intervention
paciente_Frequência Cardíaca	heart frequency
paciente_Frequência Respiratória	breath frequency
paciente_Temperatura (^{o}C)	temperature
paciente_Peso (Kg)	weight
paciente_Altura (metros)	height
paciente_Índice de Massa Corporal	body max index
paciente_Creatinina Sérica	serumic creatinine
paciente_MDRD (ml/min/173m^{2})	modification of diet in renal disease (mdrd) (ml/min/173m^{2})
paciente_Clearance Cr (ml/min)	patient_clearance cr (ml/min)
paciente_IRC_Dialítica	dialitical chronic kidney disease
paciente_IRC_Não	no chronic kidney disease
paciente_IRC_Não Dialítica	chronic kidney disease other than dialitical
paciente_IRC_Não_especificado	chronic kidney disease - not specified
paciente_Doença Vascular Periférica_Não	no peripheral vascular disease
paciente_Doença Vascular Periférica_Não Operada	not treated peripheral vascular disease
paciente_Doença Vascular Periférica_Operada / Intervenção prévia	treated peripheral vascular disease
paciente_Pressão Arterial_min	blood_pressure_min
paciente_Pressão Arterial_max	blood_pressure_max
paciente_Pressão Arterial Diastólica_min	diastolic_blood_pressure_min
paciente_Pressão Arterial Diastólica_max	diastolic_blood_pressure_max
paciente_Grau de Instrução_1^{a} a 4^{a} série	education_1^{a} to 4th grade
paciente_Grau de Instrução_2^{o} grau	education_2^{o} degree
paciente_Grau de Instrução_5^{a} a 8^{a} série	education_5th to 8th grade
paciente_Grau de Instrução_Mestrado ou Doutorado	education_master or doctorate
paciente_Grau de Instrução_Não alfabetizado	education_no literacy
paciente_Grau de Instrução_Não sabe/Sem declaração	education_not_informed
paciente_Grau de Instrução_Superior	education_graduate
paciente_Tratamento Diabetes_Dieta	patient_treatment diabetes_diet
paciente_Tratamento Diabetes_Hipoglicemiante Oral	patient_treatment diabetes_hipoglycemic
paciente_Tratamento Diabetes_Insulino Dependente	patient_treatment diabetes_insulin_dependent
paciente_Tratamento Diabetes_Sem Tratamento	patient_treatment diabetes_no treatment
paciente_Tratamento Dislipidemia_Dieta	patient_treatment dyslipidemia_diet
paciente_Tratamento Dislipidemia_Estatina	patient_treatment dyslipidemia_estatin
paciente_Tratamento Dislipidemia_Fibrato	patient_treatment dyslipidemia_fibrate
paciente_Tratamento Dislipidemia_Fibrato e Estatina	patient_treatment dyslipidemia_fibrate and statin
paciente_Tratamento Dislipidemia_Sem Tratamento	patient_tramenting dyslipidemia_no treatment
paciente_Tipo Antecedentes Familiares_Não Precoce	patient_tipo family background_not early
paciente_Tipo Antecedentes Familiares_Precoce	patient_type family background_early
procedimento_Convênio	intervention_health_insurance
procedimento_Qtde (ml)	intervention_amount of líquid (ml)
procedimento_Porta/Interv.(hrs) - Urgência	intervention_from door to intervention (hours) - urgency
procedimento_Dor/ATC (horas) - Resgate	intervention_pain (hours) - rescue
procedimento_Dor/ATC (dias) - Eletiva - Ad-hoc	intervention_pain (days) - elective - ad -hoc
procedimento_Duração Interv. (min)	intervention_intervention duration (min)
procedimento_Ano da Intervenção	intervention_year
procedimento_AAS	intervention_aspirin
procedimento_Ticlopidina	intervention_tyclopidine
procedimento_Clopidogrel	intervention_clopidogrel
procedimento_HNF	intervention_hnf
procedimento_HBPM	intervention_hbpm
procedimento_Abciximab	intervention_abciximab

procedimento_Tirofiban	intervention_tirofiban
procedimento_Eptifibatide	intervention_eptifibrate
procedimento_Adenosina	intervention_adenosin
procedimento_Ultrasom Intracoronário	intervention_intracoronary ulttrasound
procedimento_Angiografia Quantitativa	intervention_quantitative_angiography
procedimento_Inibidor de glicoproteína	intervention_ glycoprotein_inhibitor
procedimento_Clopidogrel.1	intervention_clopidogrel.1
procedimento_Dose (mg)	intervention_clopidogrel_dose (mg)
procedimento_Anestesia - Intervenção_Geral Inalatória	intervention_anesthesia - inhaled_general intervention
procedimento_Anestesia - Intervenção_Geral Sedação	intervention_anesthesia - intervention_geral sedation
procedimento_Anestesia - Intervenção_Local	intervention_anesthesia - intervention_local
procedimento_Droga - Intervenção_Fentanil	intervention_droga - intervention_fentanil
procedimento_Droga - Intervenção_Midazolan	intervention_droga - intervention_midazolan
procedimento_Droga - Intervenção_Propafenona	intervention_droga - intervention_propafenona
procedimento_Droga - Intervenção_Thionembutal	intervention_droga - intervention_thionembutal
procedimento_Droga - Intervenção_Xilocaína 2%	intervention_drog - intervention_xilocaine 2%
procedimento_Técnica - Intervenção_Convencional	intervention - technique_conventional
procedimento_Técnica - Intervenção_kissing-balloon	intervention -technique_kissing -balloon
procedimento_Introdutor - Intervenção_10F	intervention_introductor - 10f
procedimento_Introdutor - Intervenção_11F	intervention_introductor - 11f
procedimento_Introdutor - Intervenção_5F	intervention_introductor - 5f
procedimento_Introdutor - Intervenção_6F	intervention_introductor - 6f
procedimento_Introdutor - Intervenção_7F	intervention_introductor - 7f
procedimento_Introdutor - Intervenção_8F	intervention_introductor - 8f
procedimento_Introdutor - Intervenção_9F	intervention_introductor - 9f
procedimento_Via de acesso - Intervenção_Braquial Dissecção	access_ intervention _brakial dissection
procedimento_Via de acesso - Intervenção_Braquial Punção	access_ intervention - braquial punction
procedimento_Via de acesso - Intervenção_Femoral	access_ intervention - femoral
procedimento_Via de acesso - Intervenção_Radial	access_ intervention - radial
procedimento_Lado Via de Acesso - Intervenção_Direito	intervention_lass access way - right
procedimento_Lado Via de Acesso - Intervenção_Direito e Esquerdo	intervention_lalad access - right and left
procedimento_Lado Via de Acesso - Intervenção_Esquerdo	intervention_lass access way - left
procedimento_Tipo Contraste - Intervenção_Iônico de alta osmolaridade	intervention_type contrast - ionic high osmolarity
procedimento_Tipo Contraste - Intervenção_Iônico de baixa osmolaridade	intervention_type contrast - ionic low osmolarity
procedimento_Tipo Contraste - Intervenção_Não iônico de baixa osmolaridade	intervention_type contrast - low osmolarity not ionic
procedimento_Tipo Contraste - Intervenção_Não iônico isoosmolar	intervention_type contrast - isomolatity not ionic
procedimento_Tipo de Intervenção_Ad-hoc	intervention_type ad-hoc
procedimento_Tipo de Intervenção_Eletiva	intervention_type elective
procedimento_Tipo de Intervenção_Facilitada	intervention_type facilitated
procedimento_Tipo de Intervenção_Primária	intervention_type primary
procedimento_Tipo de Intervenção_Resgate	intervention_type rescue
procedimento_Tipo de Intervenção_Urgência	intervention_type urgency
procedimento_Trombolítico Prévio_Nenhum	intervention thrombolitical - none
procedimento_Trombolítico Prévio_Não_especificado	intervention thrombolitical - not specified
procedimento_Trombolítico Prévio_SK	intervention thrombolitical - sk
procedimento_Trombolítico Prévio_TNK	intervention thrombolitical - tnk
procedimento_Trombolítico Prévio_TPA	intervention thrombolitical - tpa
procedimento_Trombolítico Prévio_rPA	intervention thrombolitical - rpa
procedimento_Mês Intervenção	month of intervention
procedimento_ST-elevation	intervention_st-elevation
complicacao_Quando_Até 1 semana_sum	number of complications_1 week
complicacao_Quando_Até 24 horas_sum	number of complications_24 hours

complicacao_Quando_Intra-exame até a alta_sum	number of complications_intra-exam to discharge
complicacao_Quando_Recuperação_sum	number of complications_recovery
complicacao_Quando_Sala de exame_sum	number of complications_exam_room
complicacao_Destino Alta_Domiciliar_sum	number of complications_home discharge
complicacao_Destino Alta_Enfermaria_sum	number of complications_nursery dischargge
complicacao_Destino Alta_Hospital Origem_sum	number of complications_hospital of origin
complicacao_Óbito_Não_sum	complication_no_death
complicacao_Óbito_Sim_sum	complication_death
complicacao_Tipo Complicação_Alérgica_sum	number of complications_type alergic
complicacao_Tipo Complicação_Arritmia_sum	number of complications_type arrhythmia
complicacao_Tipo Complicação_Congestiva_sum	number of complications_type congestive
complicacao_Tipo Complicação_Embólica_sum	number of complications_type embolic
complicacao_Tipo Complicação_Hematoma_sum	number of complications_type bruise
complicacao_Tipo Complicação_Isquêmica_sum	number of complications_type ischemic
complicacao_Tipo Complicação_Neurológica_sum	number of complications_type neurological
complicacao_Tipo Complicação_R. Contraste_sum	number of complications_type contrast
complicacao_Tipo Complicação_Vagal_sum	number of complications_type vagal
complicacao_Tipo Complicação_Vascular_sum	number of complications_type vascular
complicacao_Grau - Complicação_Grave_sum	number of complications_major degree
complicacao_Grau - Complicação_Leve_sum	number of complications_light degree
complicacao_Grau - Complicação_Moderado_sum	number of complications_moderate degree
balao_Diâmetro Cateter - Balão_count	catheter diameter - balloon_count
balao_Diâmetro Cateter - Balão_mean	catheter diameter - balloon_mean
balao_Diâmetro Cateter - Balão_max	catheter diameter - balloon_max
balao_Diâmetro Cateter - Balão_min	catheter diameter - balloon_min
balao_Comprimento Cateter - Balão_mean	catheter lenght - balloon_mean
balao_Comprimento Cateter - Balão_max	catheter lenght - balloon_max
balao_Comprimento Cateter - Balão_min	catheter lenght - balloon_min
balao_Pressão Final (ATM)_mean	final balloon pressure (atm) - mean
balao_Pressão Final (ATM)_max	final balloon pressure (atm) - max
balao_Pressão Final (ATM)_min	final balloon pressure (atm) - min
balao_Grau de estenose pós_mean	degree of estenosis after intervention - mean
balao_Grau de estenose pós_max	degree of estenosis after intervention - max
balao_Grau de estenose pós_min	degree of estenosis after intervention - min
balao_Timi pós_mean	balao_timi flow post_mean
balao_Timi pós_max	balao_timi flow post_max
balao_Timi pós_min	balao_timi flow post_min
balao_Blush Miocárdico_mean	balao_myocardic blush _mean
balao_Blush Miocárdico_max	balao_myocardic blush _max
balao_Blush Miocárdico_min	balao_myocardial blush _min
balao_N. Injeções - Adenosina - Balão_sum	balao_n. injections - adenosina - balloon_sum
balao_Total Injetado (mg) - Adenosina - Balão_sum	balao_total injected (mg) - adenosine - balloon_sum
balao_N. Injeções - Papaverina - Balão_sum	balao_n. injections - papaverine - balloon_sum
balao_Total Injetado (mg) - Papaverina - Balão_sum	balao_total injected (mg) - papaverine - balloon_sum
balao_N. Injeções - Nitroglicerina - Balão_sum	balao_n. injections - nitroglycerin - balloon_sum
balao_Total Injetado (mg) - Nitroglicerina - Balão_sum	balao_total injected (mg) - nitroglicerine - balloon_sum
balao_N. Injeções - Monocordil - Balão_sum	balao_n. injections - monocordil - balloon_sum
balao_Total Injetado (mg) - Monocordil - Balão_sum	balao_total injected (mg) - monocordil - balloon_sum
balao_N. Injeções - Nitroprussiato - Balão_sum	balao_n. injections - nitroprussito - balloon_sum
balao_Total Injetado (mg) - Nitroprussiato - Balão_sum	balao_total injected (mg) - nitroprussito - balloon_sum
balao_N. Injeções - Adrenalina - Balão_sum	balao_n. injections - adrenaline - balloon_sum
balao_Total Injetado (mg) - Adrenalina - Balão_sum	balao_total injected (mg) - adrenaline - balloon_sum
balao_Device Adjunto - Balão_Braquiterapia_sum	balao_device assistant - balloon_braquiterapic_sum
balao_Device Adjunto - Balão_Exciser_sum	balao_device assistant - balloon_exciser_sum
balao_Resultado Angiográfico - Balão_I1 - Não ultrapassou a lesão_sum	balao_angiographic result - balloon_i1 - did not exceed the lesion_sum
balao_Resultado Angiográfico - Balão_I2 - Ultrapassou e não dilatou_sum	balao_angiographic result - balloon_i2 - overdue and did not dilate_sum
balao_Resultado Angiográfico - Balão_I3 - Oclusão Aguda_sum	balao_angiographic result - balloon_i3 - acute occlusion_sum

balao_Resultado Angiográfico - Balão_I4 - Dissecção_sum	balao_angiographic result - balloon_i4 - dissection_sum
balao_Resultado Angiográfico - Balão_S - Sucesso_sum	balao_angiographic result - balão_s - success_sum
balao_Tipo - Dissecção - Balão_A_sum	dissection - balloon_a_sum
balao_Tipo - Dissecção - Balão_B_sum	dissection - balloon_b_sum
balao_Tipo - Dissecção - Balão_C_sum	dissection - balloon_c_sum
balao_Tipo - Dissecção - Balão_D_sum	dissection - balloon_d_sum
balao_Tipo - Dissecção - Balão_E_sum	dissection - balloon_e_sum
balao_Tipo - Dissecção - Balão_F_sum	dissection - balloon_f_sum
balao_Classificação da Lesão (ACC/AHA) - Balão_A_sum	classification of the lesion (acc/aha) - balloon_a_sum
balao_Classificação da Lesão (ACC/AHA) - Balão_B1_sum	classification of the lesion (acc/aha) - balloon_b1_sum
balao_Classificação da Lesão (ACC/AHA) - Balão_B2_sum	classification of the lesion (acc/aha) - balloon_b2_sum
balao_Classificação da Lesão (ACC/AHA) - Balão_C_sum	classification of the lesion (acc/aha) - balloon_c_sum
balao_Distúrbio de Fluxo - Balão_Não_sum	flow disturbance_no
balao_Distúrbio de Fluxo - Balão_Sim_sum	flow disturbance_yes
balao_Tipo do Distúrbio de Fluxo - Balão_No-reflow_sum	type of flow disturbance_no-reflow
balao_Tipo do Distúrbio de Fluxo - Balão_Slow flow_sum	type of flow disturbance_slow flow
balao_Tipo No-Reflow_Inalterado após as medicações descritas_sum	type of flow disturbance_no-reflow_no change after medication
balao_Tipo No-Reflow_Melhora completa_sum	type of flow disturbance_no-reflow_complete recover
balao_Tipo No-Reflow_Melhora parcial_sum	type of flow disturbance_no-reflow_partial recover
balao_Vaso coronário - Balão_CD_sum	coronary vessel - balloon_cd_sum
balao_Vaso coronário - Balão_CX_sum	coronary vessel - balloon_cx_sum
balao_Vaso coronário - Balão_DA_sum	coronary vessel - balloon_da_sum
balao_Vaso coronário - Balão_DPD_sum	coronary vessel - balloon_dpd_sum
balao_Vaso coronário - Balão_DPE_sum	coronary vessel - balloon_dpe_sum
balao_Vaso coronário - Balão_Dg1_sum	coronary vessel - balloon_dg1_sum
balao_Vaso coronário - Balão_Dg2_sum	coronary vessel - balloon_dg2_sum
balao_Vaso coronário - Balão_Dg3_sum	coronary vessel - balloon_dg3_sum
balao_Vaso coronário - Balão_Intermédio-RI_sum	coronary vessel - balloon_intermediary_ri_sum
balao_Vaso coronário - Balão_MamáriaDA_sum	coronary vessel - balloon_mamaryda_sum
balao_Vaso coronário - Balão_Mg1_sum	coronary vessel - balloon_mg1_sum
balao_Vaso coronário - Balão_Mg2_sum	coronary vessel - balloon_mg2_sum
balao_Vaso coronário - Balão_Mg3_sum	coronary vessel - balloon_mg3_sum
balao_Vaso coronário - Balão_MgD_sum	coronary vessel - balloon_mgd_sum
balao_Vaso coronário - Balão_PVSCD_sum	coronary vessel - balloon_pvscd_sum
balao_Vaso coronário - Balão_PVSDA_sum	coronary vessel - balloon_pvsda_sum
balao_Vaso coronário - Balão_PVSDPD_sum	coronary vessel - balloon_pvsdpd_sum
balao_Vaso coronário - Balão_PVSDg1_sum	coronary vessel - balloon_pvsdg1_sum
balao_Vaso coronário - Balão_PVSMg1_sum	coronary vessel - balloon_pvsmg1_sum
balao_Vaso coronário - Balão_PVSMg2_sum	coronary vessel - balloon_pvsmg2_sum
balao_Vaso coronário - Balão_RadialDg_sum	coronary vessel - balloon_radialdg_sum
balao_Vaso coronário - Balão_TCE_sum	coronary vessel - balloon_tce_sum
balao_Vaso coronário - Balão_VPD_sum	coronary vessel - balloon_vpd_sum
balao_Vaso coronário - Balão_VPE_sum	coronary vessel - balloon_vpe_sum
balao_Kissing wire - Balão_Não_sum	kissing wire - balloon_no_sum
balao_Kissing wire - Balão_Sim_sum	kissing wire - balloon_yes_sum
balao_Kissing balloon - Balão_Não_sum	kissing balloon - balloon_no_sum
balao_Kissing balloon - Balão_Sim_sum	kissing balloon - balloon_yes_sum
stent_Diâmetro Dispositivo - Stent_mean	device diameter - stent_mean
stent_Diâmetro Dispositivo - Stent_max	device diameter - stent_max
stent_Diâmetro Dispositivo - Stent_min	device diameter - stent_min
stent_Diâmetro Dispositivo - Stent_count	device diameter - stent_count
stent_Comprimento Dispositivo - Stent_mean	device lenght - stent_mean
stent_Comprimento Dispositivo - Stent_max	device lenght - stent_max
stent_Comprimento Dispositivo - Stent_min	device lenght - stent_min

stent_Diâmetro Cateter Balão Adjunto Pré_mean	stent_catheter dimeter of attached balloon_pre_mean
stent_Diâmetro Cateter Balão Adjunto Pré_max	stent_catheter dimeter of attached balloon_max
stent_Diâmetro Cateter Balão Adjunto Pré_min	stent_catheter dimeter of attached balloon_pre_min
stent_Diâmetro Cateter Balão Adjunto Pós_mean	stent_catheter dimeter of attached balloon_post_mean
stent_Diâmetro Cateter Balão Adjunto Pós_max	stent_catheter dimeter of attached balloon_post_max
stent_Diâmetro Cateter Balão Adjunto Pós_min	stent_catheter dimeter of attached balloon_post_min
stent_Pressão Final de Liberação (ATM)_mean	stent_final release pressure (atm) _mean
stent_Pressão Final de Liberação (ATM)_max	stent_final release pressure (atm) _max
stent_Pressão Final de Liberação (ATM)_min	stent_final release pressure (atm) _min
stent_Grau de estenose pós_mean	stent_degree of stenosis post_mean
stent_Grau de estenose pós_max	stent_degree of stenosis post_max
stent_Grau de estenose pós_min	stent_degree of stenosis post_min
stent_Blush Miocárdico_mean	stent_myocardial blush_mean
stent_Blush Miocárdico_max	stent_myocardial blush_max
stent_Blush Miocárdico_min	stent_myocardial blush_min
stent_N. Injeções - Adenosina - Stent_sum	stent_n. injections - adenosine - stent_sum
stent_Total Injetado (mg) - Adenosina - Stent_sum	injected stent_total (mg) - adenosine - stent_sum
stent_N. Injeções - Papaverina - Stent_sum	stent_n. injections - papaverine - stent_sum
stent_Total Injetado (mg) - Papaverina - Stent_sum	injected stent_total (mg) - papaverine - stent_sum
stent_N. Injeções - Nitroglicerina - Stent_sum	stent_n. injections - nitroglycerin - stent_sum
stent_Total Injetado (mg) - Nitroglicerina - Stent_sum	injected stent_total (mg) - nitroglycerin - stent_sum
stent_N. Injeções - Monocordil - Stent_sum	stent_n. injections - monocordil - stent_sum
stent_Total Injetado (mg) - Monocordil - Stent_sum	injected stent_total (mg) - monocordil - stent_sum
stent_N. Injeções - Nitroprussiato - Stent_sum	stent_n. injections - nitroprussian - stent_sum
stent_Total Injetado (mg) - Nitroprussiato - Stent_sum	stent_total injected (mg) - nitroprusside - stent_sum
stent_N. Injeções - Adrenalina - Stent_sum	stent_n. injections - adrenaline - stent_sum
stent_Total Injetado (mg) - Adrenalina - Stent_sum	stent_total injected (mg) - adrenaline - stent_sum
stent_Timi pós_0.0_mean	stent_timi post_0.0_mean
stent_Timi pós_0.0_max	stent_timi post_0.0_max
stent_Timi pós_0.0_min	stent_timi post_0.0_min
stent_Timi pós_1.0_mean	stent_timi post_1.0_mean
stent_Timi pós_1.0_max	stent_timi post_1.0_max
stent_Timi pós_1.0_min	stent_timi post_1.0_min
stent_Timi pós_2.0_mean	stent_timi post_2.0_mean
stent_Timi pós_2.0_max	stent_timi post_2.0_max
stent_Timi pós_2.0_min	stent_timi post_2.0_min
stent_Timi pós_3.0_mean	stent_timi post_3.0_mean
stent_Timi pós_3.0_max	stent_timi post_3.0_max
stent_Timi pós_3.0_min	stent_timi post_3.0_min
vaso_Grau de estenose pré - Intervenção_max	vessel_degree of stenosis_max
vaso_Grau de estenose pré - Intervenção_min	vessel_degree of stenosis_min
vaso_Grau de estenose pré - Intervenção_mean	vessel_degree of stenosis_mean
vaso_Grau de estenose pré - Intervenção_sum	vessel_degree of stenosis_sum
vaso_Anatomia - Intervenção_Ateromatose moderada_sum	vessel_anatomy - intervention_ateromatosis moderate_sum
vaso_Anatomia - Intervenção_Ateromatose severa_sum	vessel_anatomy - intervention_ateromatosis severe_sum
vaso_Anatomia - Intervenção_Com irregularidades parietais_sum	vessel_anatomy - intervention_with parietal irregularities_sum
vaso_Anatomia - Intervenção_Com lesão_sum	vessel_anatomy - intervention_injury_sum
vaso_Vaso coronário - Intervenção_CD_sum	coronary vessel - intervention_cd_sum
vaso_Vaso coronário - Intervenção_CX_sum	coronary vessel - intervention_cx_sum
vaso_Vaso coronário - Intervenção_DA_sum	coronary vessel - intervention_da_sum
vaso_Vaso coronário - Intervenção_DPD_sum	coronary vessel - intervention_dpd_sum
vaso_Vaso coronário - Intervenção_DPE_sum	coronary vessel - intervention_dpe_sum
vaso_Vaso coronário - Intervenção_Dg1_sum	coronary vessel - intervention_dg1_sum
vaso_Vaso coronário - Intervenção_Dg2_sum	coronary vessel - intervention_dg2_sum

vaso_Vaso coronário - Intervenção_Dg3_sum	coronary vessel - intervention_dg3_sum
vaso_Vaso coronário - Intervenção_Intermédio-RI_sum	coronary vessel - intervention_intermadio -ri_sum
vaso_Vaso coronário - Intervenção_MamáriaCD_sum	coronary vessel - intervention_mamáriacd_sum
vaso_Vaso coronário - Intervenção_MamáriaDA_sum	coronary vessel - intervention_mamádada_sum
vaso_Vaso coronário - Intervenção_MamáriaDg_sum	coronary vessel - intervention_mamáriadg_sum
vaso_Vaso coronário - Intervenção_MamáriaMg_sum	coronary vessel - intervention_mamámmg_sum
vaso_Vaso coronário - Intervenção_Mg1_sum	coronary vessel - intervention_mg1_sum
vaso_Vaso coronário - Intervenção_Mg2_sum	coronary vessel - intervention_mg2_sum
vaso_Vaso coronário - Intervenção_Mg3_sum	coronary vessel - intervention_mg3_sum
vaso_Vaso coronário - Intervenção_Mg4_sum	coronary vessel - intervention_mg4_sum
vaso_Vaso coronário - Intervenção_MgD_sum	coronary vessel - intervention_mgd_sum
vaso_Vaso coronário - Intervenção_PVSCD_sum	coronary vessel - intervention_pvscd_sum
vaso_Vaso coronário - Intervenção_PVSDA_sum	coronary vessel - intervention_pvsda_sum
vaso_Vaso coronário - Intervenção_PVSDPD_sum	coronary vessel - intervention_pvsdpd_sum
vaso_Vaso coronário - Intervenção_PVSDPE_sum	coronary vessel - intervention_pvsdpe_sum
vaso_Vaso coronário - Intervenção_PVSDg1_sum	coronary vessel - intervention_pvsdg1_sum
vaso_Vaso coronário - Intervenção_PVSDg2_sum	coronary vessel - intervention_pvsdg2_sum
vaso_Vaso coronário - Intervenção_PVSMg1_sum	coronary vessel - intervention_pvsmg1_sum
vaso_Vaso coronário - Intervenção_PVSMg2_sum	coronary vessel - intervention_pvsmg2_sum
vaso_Vaso coronário - Intervenção_PVSRI_sum	coronary vessel - intervention_pvsri_sum
vaso_Vaso coronário - Intervenção_PVSVPD_sum	coronary vessel - intervention_pvsvpd_sum
vaso_Vaso coronário - Intervenção_PVSVPE_sum	coronary vessel - intervention_pvsvpe_sum
vaso_Vaso coronário - Intervenção_RadialCD_sum	coronary vessel - intervention_radialcd_sum
vaso_Vaso coronário - Intervenção_RadialDg_sum	coronary vessel - intervention_radialdg_sum
vaso_Vaso coronário - Intervenção_RadialMg_sum	coronary vessel - intervention_radialmg_sum
vaso_Vaso coronário - Intervenção_TCE_sum	coronary vessel - intervention_tce_sum
vaso_Vaso coronário - Intervenção_VPD_sum	coronary vessel - intervention_vpd_sum
vaso_Vaso coronário - Intervenção_VPE_sum	coronary vessel - intervention_vpe_sum
vaso_Local da Lesão - Intervenção_Anastomose_sum	vessel_lesion - intervetion_anastomosis_sum
vaso_Local da Lesão - Intervenção_Difusa(s)_sum	vessel_lesion - diffuse intervention (s) _sum
vaso_Local da Lesão - Intervenção_Distal_sum	vessel_lesion - intervention_distal_sum
vaso_Local da Lesão - Intervenção_Distal do enxerto_sum	vessel_lesion - intervention_distal of graft_sum
vaso_Local da Lesão - Intervenção_Distal nativo_sum	vessel_lesion - intervention_distal native_sum
vaso_Local da Lesão - Intervenção_Médio_sum	vessel_lesion - intervention_medium_sum
vaso_Local da Lesão - Intervenção_Médio do enxerto_sum	vessel_lesion - intervention_graft medium_sum
vaso_Local da Lesão - Intervenção_Médio nativo_sum	vessel_lesion - intervention_medio nativo_sum
vaso_Local da Lesão - Intervenção_Proximal_sum	vessel_lesion - intervention_proximal_sum
vaso_Local da Lesão - Intervenção_Proximal do enxerto_sum	vessel_lesion - intervention_graft proximal_sum
vaso_Local da Lesão - Intervenção_Óstio_sum	vessel_lesion - intervention_ostium_sum
vaso_Local da Lesão - Intervenção_Óstio do enxerto_sum	vessel_lesion - intervention_graft ostium_sum
vaso_Tipo de Lesão - Intervenção_De Novo_sum	vessel_type of injury - intervention_new_sum
vaso_Tipo de Lesão - Intervenção_Dissecção espontânea_sum	vessel_type of injury - intervention_ spontaneous discussion_sum
vaso_Tipo de Lesão - Intervenção_Oclusão total crônica_sum	vessel_type of injury - intervention_ chronic occlusion_sum
vaso_Tipo de Lesão - Intervenção_Restenose intrastent_sum	vessel_type of injury - intervention_restenosis intrastent_sum
vaso_Tipo de Lesão - Intervenção_Restenose pós balão_sum	vessel_type of injury - intervention_ resteenosis post balloon_sum
vaso_Tipo de Lesão - Intervenção_Trombose intrastent_sum	vessel_type of injury - intervention_trombosis intrastent_sum
vaso_Tipo de Lesão - Intrastent - Intervenção_Difusa_sum	vessel_type of injury - intrastent - intervention_difuse_sum

vaso_Tipo de Lesão - Intrastent - Intervenção_Focal_sum	vessel_type of injury - intrastent - intervention_focal_sum
vaso_Tipo de Lesão - Intrastent - Intervenção_Oclusão_sum	vessel_type of injury - intrastent - intervention_oclusion_sum
vaso_Tipo de Lesão - Intrastent - Intervenção_Proliferativa_sum	vessel_type of injury - intrastent - intervention_proliferative_sum
vaso_Tipo de Lesão - Focal - Intervenção_Borda IB_sum	vessel_type of injury - focal - intervention_border ib_sum
vaso_Tipo de Lesão - Focal - Intervenção_Focal corpo stent IC_sum	vessel_type of injury - focal - intervention_focal body stent ic_sum
vaso_Tipo de Lesão - Focal - Intervenção_Multifocal ID_sum	vessel_type of injury - focal - intervention_multifocal id_sum
vaso_Tipo de Lesão - Focal - Intervenção_gap IA_sum	vessel_type of injury - focal - intervention_gap i_sum
vaso_Tipo de Lesão - Difusa - Intervenção_Intrastent II_sum	vessel_type of injury - diffuse - intervention_intra - sent ii_sum
vaso_Tipo de Lesão - Difusa - Intervenção_Oclusão total_sum	vessel_type of injury - diffuse - intervention_total oclusion_sum
vaso_Tipo de Lesão - Difusa - Intervenção_Proliferativca III_sum	vessel_type of injury - diffuse - intervention_proliferative iii_sum
vaso_Tipo de Lesão - Pós balão - Intervenção_Tardia_sum	vessel_type of injury - post balloon - intervention_late_sum
vaso_Calcificação - Intervenção_Acentuada_sum	vessel_calcification - intervention_accentuated_sum
vaso_Calcificação - Intervenção_Ausente_sum	vessel_calcification - intervention_absent_sum
vaso_Calcificação - Intervenção_Discreta_sum	vessel_calcification - intervention_discrete_sum
vaso_Calcificação - Intervenção_Moderada_sum	vessel_calcification - intervention_moderate_sum
vaso_Tortuosidade - Intervenção_Acentuada_sum	vessel_tortuous - intervention_accentuated_sum
vaso_Tortuosidade - Intervenção_Leve_sum	vessel_tortuous - intervention_light_sum
vaso_Tortuosidade - Intervenção_Moderada_sum	vessel_tortuous - intervention_moderate_sum
vaso_Angulação - Intervenção_gt45_sum	vessel_angulation - intervention_gt45_sum
vaso_Angulação - Intervenção_gt90_sum	vessel_angulation - intervention_gt90_sum
vaso_Angulação - Intervenção_Acentuada_sum	vessel_angulation - intervention_accentuated_sum
vaso_Angulação - Intervenção_Leve_sum	vessel_angulation - intervention_light_sum
vaso_Angulação - Intervenção_Moderada_sum	vessel_angulation - intervention_moderate_sum
vaso_Grau de Importância - Intervenção_Discreta_sum	vessel_grau of importance - intervention_discrete_sum
vaso_Grau de Importância - Intervenção_Grande_sum	vessel_grau of importance - intervention_big_sum
vaso_Grau de Importância - Intervenção_Moderada_sum	vessel_grau of importance - intervention_moderate_sum
vaso_Maior que 20mm - Intervenção_Não_sum	vessel_greater than 20mm - intervention_no_sum
vaso_Maior que 20mm - Intervenção_Sim_sum	vessel_greater than 20mm - intervention_yes_sum
vaso_Trombo - Intervenção_Não_sum	vessel_trombo - intervention_no_sum
vaso_Trombo - Intervenção_Sim_sum	vessel_trombo - intervention_yes_sum
vaso_Placa Rota - Intervenção_Não_sum	vessel_broken plate - intervention_no_sum
vaso_Placa Rota - Intervenção_Sim_sum	vessel_broken plate - intervention_yes_sum
vaso_Placa Ulcerada - Intervenção_Não_sum	vessel_ulcerous plate - intervention_no_sum
vaso_Placa Ulcerada - Intervenção_Sim_sum	vessel_ulcerous plate - intervention_yes_sum
vaso_Ectasia - Intervenção_Não_sum	vessel_ectasia - intervention_no_sum
vaso_Ectasia - Intervenção_Sim_sum	vessel_ectasia - intervention_sim_sum
vaso_Aneurisma - Intervenção_Não_sum	vessel_aneurisma - intervention_no_sum
vaso_Aneurisma - Intervenção_Sim_sum	vessel_aneurisma - intervention_yes_sum
vaso_Bifurcação - Intervenção_001_sum	vessel_bifurcation - intervention_001_sum
vaso_Bifurcação - Intervenção_010_sum	vessel_bifurcation - intervention_010_sum
vaso_Bifurcação - Intervenção_011_sum	vessel_bifurcation - intervention_011_sum
vaso_Bifurcação - Intervenção_100_sum	vessel_bifurcation - intervention_100_sum
vaso_Bifurcação - Intervenção_101_sum	vessel_bifurcation - intervention_101_sum
vaso_Bifurcação - Intervenção_110_sum	vessel_bifurcation - intervention_110_sum
vaso_Bifurcação - Intervenção_111_sum	vessel_bifurcation - intervention_111_sum

Is there a label or target associated with each instance? If so, please provide a description. The target is the dataset's last column and provides the intervention's

outcome, whether a MACE event had happened or not. A MACE event is considered when, during, or immediately after the intervention, the patient died, had a myocardial infarction, or had a stroke.

Is any information missing from individual instances? (If so, please explain why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.) All missing information was either excluded (feature-wise or instance-wise) or treated so that the resulting dataset had no missing information.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? (If so, please describe how these relationships are made explicit.) There are no relationships between individual instances.

Are there recommended data splits (e.g., training, development/validation, testing)? (If so, please provide a description of these splits, explaining the rationale behind them.) We expect this data to be used solely for testing purposes. We do not explicitly provide a training/validation/testing split; however, we recognize that people may wish to do this or to do some form of cross-validation. We suggest cross-validation, given that some phenomena only occur in a few instances and are likely to be lost in any random split.

Are there any errors, sources of noise, or redundancies in the dataset? (If so, please provide a description.) There are almost certainly some errors in data collection and annotation. We did our best to minimize these, but some indeed remain.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? (If it links to or relies on external resources, a) are there guarantees that they will exist and remain constant over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.)

The dataset is self-contained. Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? (If so, please provide a description.) No; Although the data represents patient confidential information, all data was anonymized and thus did not contain any personally identifiable information.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? (If so, please describe why.) No.

Does the dataset relate to people? (If not, you may skip the remaining questions in this section.) Yes.

Does the dataset identify any subpopulations (e.g., by age, gender)? (If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.) The dataset contains individuals ranging from 0 to 106 years, and the average age is 62 years old. Samples are from both genders, and Males comprise about 70% of the dataset. Regarding cardiovascular diseases, the dataset contains both chronic and acute cases.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? (If so, please describe how.) No. Data were anonymized.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? (If so, please provide a description.) No.

Any other comments? None.

A.3 Collection pocess

How was the data associated with each instance acquired? (Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.) The data was collected from databases of hospitals and clinics across 7 states in Brazil. Raw information was reported by subjects (doctors and nurses) and was not validated or verified.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? (How were these mechanisms or procedures validated?) Data was originally collected by each hospital and clinic using their own electronic health record software. It was then exported and aggregated in one single dataset.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? The dataset is not a sample from a larger dataset.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors), and how were they compensated (e.g., how much were crowdworkers paid)? The hospital and clinic's medical staff did all collection and annotation.

Over what timeframe was the data collected? (Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.) The data was collected from 2006 to 2018.

Were any ethical review processes conducted (e.g., by an institutional review board)? (If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.) The use of patient data in the MI-SIEVE ACC database was approved by the Research Ethics Committee (CEP/IGESDF, Comitê de Ética em Pesquisa do Instituto de Gestão Estratégica de Saúde do Distrito Federal) according to the approval number 3.854.051 on February 21st, 2020, and approval 4.263.940 on September 8th, 2020.

Does the dataset relate to people? (If not, you may skip the remaining questions in this section.) Yes;

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)? Directly from the individuals.

Were the individuals in question notified about the data collection? (If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.) Yes.

Did the individuals in question consent to the collection and use of their data? (If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.) Data collection is a usual practice in hospitals and clinics and patients consent to giving this data to aid in the treatment process. Data were collected from 2006 to 2018, prior to the creation of a specific regulation in Brazil, called LGPD, that became mandatory only in 2020. Nevertheless, the dataset followed common anonymization principles to preserve patients' privacy.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? (If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).) No.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? (If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.) No.

Any other comments? None.

A.4 Preprocessing/cleaning/labeling

Was any pre-processing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? (If so, please provide a description. If not, you may skip the remainder of the questions in this section.) Yes, there was extensive pre-processing and cleaning of the data. This work was supervised by the domain specialist MD. Luiz Sérgio Fernandes de Carvalho. As the details of each feature pre-processing are very extensive, we will not list them here, but a Jupyter Notebook with all data pre-processing and cleaning can be shared upon request.

Was the "raw" data saved in addition to the pre-processed/cleaned/labeled data (e.g., to support unanticipated future uses)? (If so, please provide a link or other access point to the "raw" data.) Yes, the original raw data was saved. Access to the raw data will be provided on a need basis upon request to Clarity Health, the company which owns the dataset.

Is the software used to pre-process/clean/label the instances available? (If so, please provide a link or other access point.) Yes. We used Python Jupyter Notebooks.

Any other comments? None.

A.5 Uses

Has the dataset been used for any tasks already? (If so, please provide a description.) The dataset has never been used before.

Is there a repository that links to any or all papers or systems that use the dataset? (If so, please provide a link or other access point.) N/A.

What (other) tasks could the dataset be used for? The dataset could be used to understand epidemiological factors related to acute coronary syndromes and assess risk factors associated with worse outcomes among individuals treated in cardiac catheterization labs in Brazil.

Is there anything about the composition of the dataset or the way it was collected and pre-processed/cleaned/labeled that might impact future uses? (For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?) Yes. The dataset was cleaned/pre-processed/labeled to be used in Machine Learning research. As data is not confirmed nor verified, it should NOT be used for medical research, and thus no medical conclusion or therapy should be derived based on this data.

Are there tasks for which the dataset should not be used? (If so, please provide a description.) As mentioned above, the dataset should NOT be used for medical research, or to obtain any medical conclusion, as it was not confirmed nor verified.

Any other comments? None.

A.6 Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? (If so, please provide a description.) The dataset is proprietary, and its distribution would be done upon request and approval by Clarity Health, the company which owns the dataset.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? (Does the dataset have a digital object identifier (DOI)?) By e-mail, on a need basis.

When will the dataset be distributed? Only in the cases approved by Clarity Health, the company which owns the dataset.

Will the dataset be distributed under a copyright or other intellectual property (IP) license and/or under applicable terms of use (ToU)? (If so, please describe this license and/or ToU and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.) Yes.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? (If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.) Not to our knowledge.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? (If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.) Not to our knowledge.

Any other comments? None.

A.7 Maintenance

Who is supporting/hosting/maintaining the dataset? The authors are maintaining the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., e-mail address)? E-mail address: luizsergiofc@gmail.com

Is there an erratum? (If so, please provide a link or other access point.) Currently, no. As errors are encountered, future versions of the dataset may be released (but will be versioned).

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances')? (If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?) Same as previous.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? (If so, please describe these limits and explain how they will be enforced.) No.

Will older versions of the dataset continue to be supported/hosted/maintained? (If so, please describe how. If not, please describe how its obsolescence will be communicated to users.) Yes, all data will be versioned.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? (If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.) Currently, we are not considering any contribution from the community, as the dataset is private.

Any other comments? None.