A Study on Counterfactual Explanations

Investigating the impact of inter-class distance and data imbalance

Ivor Zagorac

Delft University of Technology

TU Delft

A Study on Counterfactual Explanations

Investigating the impact of inter-class distance and data imbalance

by

Ivor Zagorac

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Thursday June 13, 2024 at 9:00 AM.

Student number:	4691202	
Project duration:	September 13, 2023 – June 13, 2024	
Thesis committee:	C.C.S Liem,	TU Delft, supervisor
	P. Altmeyer,	TU Delft, daily co-supervisor
	D.M.J. Tax	TU Delft

TU Delft

Summary

Counterfactual explanations (CEs) are emerging as a crucial tool in Explainable AI (XAI) for understanding model decisions. This research investigates the impact of various factors on the quality of CEs generated for classification tasks. We explore how inter-class distance, data imbalance, balancing techniques, the presence of biased classifiers, and decision thresholds influence CE quality. To answer these research questions, we conduct experiments on various datasets, classification models and counterfactual generators. The datasets include the MNIST and GMSC dataset. The models include well-established models like MLP and Random Forest, along with the novel NeuroTree model. The generators include the method proposed by Wachter et al. and the REVISE method. We evaluate how different factors affect CE quality by performing an extensive experimental analysis. Our findings demonstrate that increasing inter-class distance degrades CE quality, particularly explanation plausibility. Data imbalance showed minimal impact, while balancing techniques yielded a slight improvement in CE plausibility, especially for the minority class. Classifiers biased towards specific subgroups resulted in lower CE quality for those subgroups. We observed limited evidence for a consistent amplification effect of decision thresholds. This research utilizes various datasets and classification models, including the novel NeuroTree model. Our findings contribute to XAI by providing insights into factors affecting CE quality and highlighting areas for future development, particularly regarding fairness and handling imbalanced data.

Contents

1
Introduction

Machine learning applications are increasingly popular in domains like the criminal justice system [13], healthcare [39], and finance [76], directly influencing decisions that impact individuals. As AI systems gain traction in these real-world applications, there is a growing demand for transparency and explainability. This demand is further emphasized by legal regulations like the General Data Protection Regulation (GDPR) [71] and, more recently, the Artificial Intelligence Act (AI Act) [56] adopted by the European Parliament. The AI act requires high-risk AI systems to satisfy safety requirements where ”AI systems are always considered high-risk if it profiles individuals, i.e. automated processing of personal data to assess various aspects of a person’s life, such as ... economic situation” [42]. Additionally, GDPR contains a specific policy on the right of citizens to receive an explanation for algorithmic decisions, placing a legal obligation on explainable decision-making [27]. This type of decision-making has been the focus of the Explainable AI (XAI) field that has emerged in recent years [88].

There are two primary approaches to achieving XAI: inherently interpretable models and post-hoc explanation techniques. This is the distinction which we adhere by in this work and has also been found in literature such as the work of Rudin [73] and Guidotti et al. [31]. Inherently interpretable models are those whose internal workings are readily understandable by humans. Examples include linear regression, decision trees, and rule sets. These models often provide clear insights into how features contribute to the final prediction. However, their expressive power can be limited for complex problems. In contrast, other models, such as random forests and deep neural networks, are often highly effective but lack inherent interpretability. Their decision-making processes can be opaque, making it challenging to understand how they arrive at specific outputs.

Several studies have highlighted the prevalence of both interpretable and non-interpretable models in the fields of criminal justice, healthcare, and finance. The research by Hassani et al. [41], found that the criminal justice system utilises a variety of classification models, including decision trees, random forests, and neural networks. Similarly, the choice of model in healthcare is highly problem-dependent, ranging from decision trees and Support Vector Machines (SVMs) to neural networks [39, 80]. The financial sector also employs a broad spectrum of machine learning techniques, including random forests and neural networks, for tasks like credit risk assessment [10]. These findings show that there is a need for explanation approaches which help to understand a wide range of non-interpretable models.

Counterfactual Explanations (CEs) are a model-agnostic approach, meaning it is not tailored to specific classification models, which focuses on understanding the relationship between inputs and outputs of classification models, rather than investigating the internal mechanics of these models directly. CEs aim to provide individuals impacted by an automated decision-making system with a data point that is similar to the individual at stake but would be classified as a different class. This can be used as a diagnostic tool to understand the workings of the model but can also be used to generate (algorithmic) recourse (AR) first defined by Ustun et al. [81]. AR is about giving non-expert individuals actionable recommendations on how to change their classification. This is particularly valuable when a clear distinction exists between desirable and undesirable classifications. Consider the credit risk domain

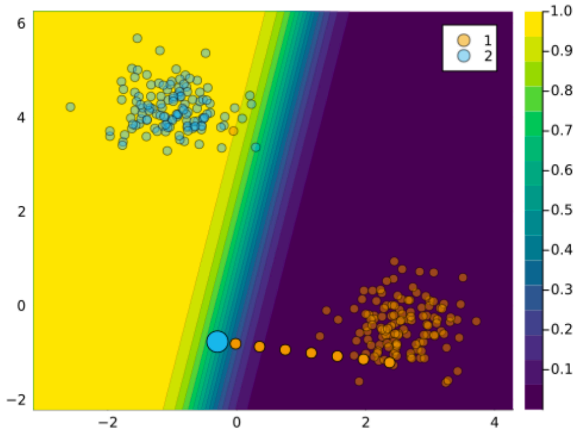


Figure 1.1: Counterfactual path using generic counterfactual generator for conventional binary classifier [1].

as an illustrative example. For an applicant whose loan request is rejected (undesirable outcome), algorithmic recourse generated from a counterfactual explanation can offer actionable feedback that helps them improve their creditworthiness and potentially secure loan approval (desirable outcome). Figure 1.1 gives an example of a CE. In the figure we see two classes (orange and blue dots) and the decision boundary of a classifier depicted as a heatmap. The highlighted orange dots show the path of a data point to its generated counterfactual (large blue dot).

One key question regarding counterfactual explanation methods is how the distance between classes impacts CE quality. Since CE generation involves transitioning from a factual instance to the desired class, it is expected that as this distance increases, the quality of counterfactuals degrades. To the best of our knowledge, there is currently no existing research that investigates this relationship between inter-class distance and CE quality. This thesis aims to fill this gap in the literature by examining how the inter-class distance affects the quality of generated counterfactuals.

Besides explainability and interpretability, the issue of imbalanced data in classification tasks has been an important direction for research in machine learning and AI. The survey by Haixiang et al. [35] provides a detailed summary of this research. Among other things, the authors present an extensive list of studies that tackle imbalanced learning in various application domains. This list includes domains such as ’chemical, biomedical engineering’ and ’financial management’. Notably, these domains intersect with healthcare and finance, fields where explainability has been previously highlighted as critical. This reinforces the notion that the need for interpretable classification models is prevalent in scenarios with imbalanced datasets. However, existing research on counterfactual explanations lacks investigations in this specific context. This thesis aims to address this gap by evaluating the quality of CEs generated for imbalanced datasets.

Furthermore, the interplay between balancing techniques and CEs remains unexplored. Balancing techniques aim to mitigate the negative effects of data imbalance. This research investigates whether these techniques can also positively influence CE quality.

The quality of CE has been extensively researched and surveyed. We begin by examining how CE quality is defined and assessed in Explainable AI before connecting it to the field of CEs. These findings are crucial for understanding how to evaluate generated counterfactuals in different scenarios.

Beyond the aforementioned aspects, we also explore how classifiers biased towards minority sub-groups in the data affect CE quality. We introduce a classifier exhibiting negative bias towards a specific sub-group, extending the concept beyond the minority class. While related to existing research on fairness in CEs, our approach differs by explicitly considering the potential bias of the classification model. This novel research direction offers valuable insights into current CE generation methods.

The following section presents the research questions that have been defined to help this research. To answer these research questions we perform experiments which are described in detail in Section 6.2.

We perform these experiments on different datasets such as MNIST [50] and Give Me Some Credit (GMSC) [17]. The experiments will test the behaviour of various counterfactual generators in combination with different classification models such as Multi-Layer Perceptron (MLP), Random Forest and the NeuroTree model implemented in [20]. This NeuroTree model combines the differentiability of a neural network with the performance on tabular data of a Random Forest. This is relevant in the context of CE because many state-of-the-art generators rely on differentiable classifiers. To the best of our knowledge, this is the first work that utilises the NeuroTree model in the field of CE.

1.1. Research Questions

Research Question 1
How does inter-class distance impact the quality of generated counterfactuals?

By answering this research question, we try to find out whether there is a relationship between the distance between the factual and target class, and the quality of the generated counterfactual.

Hypothesis: A larger inter-class distance will lead to a lower quality of counterfactuals.

Research Question 2
What is the effect of data imbalance on the quality of counterfactuals?

This question intends to identify whether imbalanced datasets lead to poorer quality of counterfactual explanations compared to balanced datasets.

Hypothesis: The quality of counterfactuals generated for the minority class of an imbalanced dataset will be lower than for the same class in a balanced environment.

Research Question 3
How do balancing techniques employed on imbalanced datasets affect the quality of counterfactuals?

We want to see whether using balancing techniques to balance datasets has any effect on the quality of counterfactuals.

Hypothesis: Balancing imbalanced datasets with techniques such as SMOTE and RUS will have a positive effect on the quality of the counterfactuals.

Research Question 4
How does the quality of counterfactuals for subgroups in the data which the classifier is biased towards compare to other data points in the same class?

With this research question we intend to observe whether the quality of counterfactuals generated for biased subgroups is worse than other points in the dataset. A biased subgroup in this context means a group of data points bounded by feature constraints for which the classifier will predict a negative label with a high probability. We want to know whether there is a difference in quality between the counterfactuals generated for points in the biased subgroup compared to other data points in the negatively labeled class.

Hypothesis: We hypothesise that the counterfactuals generated for points in the biased subgroup are of worse quality compared to other points outside the subgroup with the same label.

Each of the research questions described above will have its own set of experiments, presented in Section 6.2, that help to answer it. Additionally, we became interested in exploring the potential influence of the decision threshold on various factors. The decision threshold is a user-defined value between 0 and 1 that determines the termination point for the counterfactual search. A threshold of 0.5 implies that a counterfactual crossing the classification boundary is deemed sufficient in the context of binary classification. Conversely, a value closer to 1 signifies a stricter requirement for the counterfactual to be confidently classified within the target class. To guide this exploration, we formulated the following research question and hypothesis:

Research Question 5
How does the decision threshold affect the impact of inter-class distance, data imbalance, balancing techniques and negatively biased subgroups on the quality of generated counterfactuals?

Hypothesis: The impact of inter-class distance, data imbalance, balancing techniques and negatively biased subgroups on the quality of generated counterfactuals is larger when the decision threshold is closer to 1

This research question will not require a separate set of experiments. Instead, we will investigate its answer throughout the experiments conducted for the other research questions.

1.2. Contributions

Overall, the key contributions of this research can be summarized as follows:

- We investigate the relationship between inter-class distance and the quality of counterfactual explanations.
- We explore the effect of data imbalance on the quality of counterfactual explanations.
- We analyze the impact of common data balancing techniques on the quality of counterfactual explanations.
- We evaluate whether biased classifiers exhibit differences in the quality of counterfactual explanations generated for different subgroups of classes.

1.3. Structure

The rest of this report is structured as follows. Chapter 2 provides detailed descriptions of the classifications models that are utilised in this research. In Chapter 3 we delve into the challenges that arise in classification tasks due to data imbalance. Additionally, we discuss balancing techniques that could help to overcome these challenges. Literature on the notions of quality in XAI and how they are linked to desiderata for CEs is provided in Chapter 4. In Chapter 5 we present the related work of this research focusing on data imbalance in CE and researching how fairness is evaluated in recourse and CE. Chapter 6 presents the methodology which consists of the experimental setup and design. In Chapter 7, the results of the experiments are displayed and analysed. Chapter 8 concludes the research. Finally, Chapter 9 discusses the research and proposes future work.

2 Classification Models

As discussed in Chapter 1, machine learning applications are increasingly popular domains such as the criminal justice system, healthcare, and finance. Specifically classification models are commonly implemented to help the decision-making process. This chapter provides a concise overview of the classification models employed in this research. Each model's architecture and functionality are described, accompanied by an illustrative figure to enhance comprehension. The models covered include the Multi-Layer Perceptron (MLP), Decision Tree, Random Forest, and the recently introduced NeuroTree model. In-depth implementation details for the NeuroTree can be found in the referenced work by Desgagné-Bouchard et al. [20].

2.1. Multi-Layer Perceptron

Artificial neural networks (ANNs) are a powerful tool for machine learning tasks, originally inspired by the structure and function of the biological brain. A fundamental type of ANN is the Multi-Layer Perceptron (MLP), also known as a feedforward neural network. MLPs consist of an input layer, an output layer, and one or more hidden layers, all containing interconnected artificial neurons.

Each neuron applies a weighted sum of its inputs and passes the result through a non-linear activation function. This function introduces non-linearity into the network, allowing it to model complex relationships between the input and output data. Unlike simpler Perceptrons, which are limited to linearly separable problems, MLPs can effectively learn non-linear patterns through the stacked layers. The connections between neurons hold weights, which are adjusted during a training process to minimize an error function. This training process, typically performed using backpropagation, allows the MLP to learn intricate mappings from input to output.

Figure 2.1 depicts a simple MLP architecture with one hidden layer. Each circle represents a neuron, and the arrows denote connections with associated weights. The network progressively transforms the input data through the hidden layer(s) to generate the desired output.

2.2. Decision Tree

Even though this research does not use the decision tree model for classification, the knowledge of how the model works is a necessary prerequisite to understand how the Random Forest and NeuroTree model works.

Decision trees (DTs) are a fundamental supervised learning approach for classification tasks. They leverage a tree-like structure where internal nodes represent features (attributes) of the data, and branches represent decision rules based on those features. Leaf nodes, also known as terminal nodes, contain the final predicted class labels.

During classification, a data point traverses the tree starting from the root node. At each internal node, the value of the corresponding feature is compared to a threshold or specific value. Based on this comparison, the data point is directed down the appropriate branch towards the next decision node. The

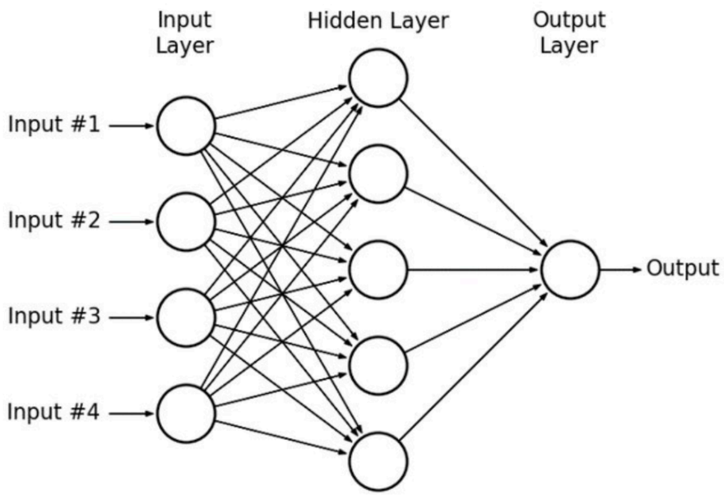


Figure 2.1: Example of a Multi-Layer Perceptron architecture [40].

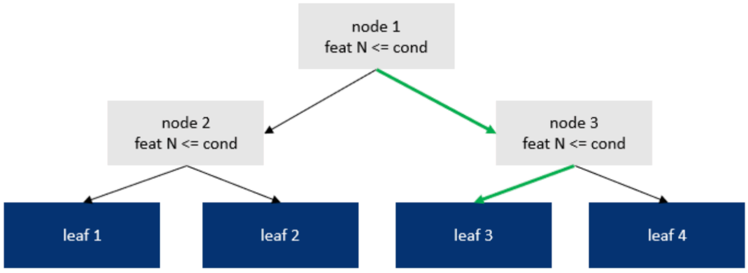


Figure 2.2: Example of a Decision Tree [62].

traversal continues until a leaf node is reached, and the associated class label becomes the predicted class for the data point.

Figure 2.2 depicts an example decision tree where the gray nodes represent internal nodes and the blue nodes represent the leaf nodes. The set of green arrows in the figure is an example of a traversal of a data point.

2.3. Random Forest

Random forests are ensemble learning algorithms that leverage the collective power of multiple decision trees. Each decision tree partitions the data based on features (attributes) to arrive at a final prediction. By combining the predictions of numerous trees, random forests aim to improve overall accuracy and address the issue of overfitting present in single decision trees.

During training, random forests construct a multitude of decision trees. Each tree is built using a subset of the training data drawn with replacement (bootstrapping). Additionally, at each node in the tree, a random subset of features is considered for splitting the data. This is different from a single decision tree where for each node, all features are considered. This randomness in tree generation and feature selection helps prevent the forest from overfitting to the training data.

When making a prediction, each tree in the forest votes on the class or value for a new data point. In classification tasks, the most popular class among the trees' votes becomes the final prediction. For regression tasks, the average predicted value across all trees is considered the final output. An exmaple of a Random Forest archictecture is given in Figure 2.3.

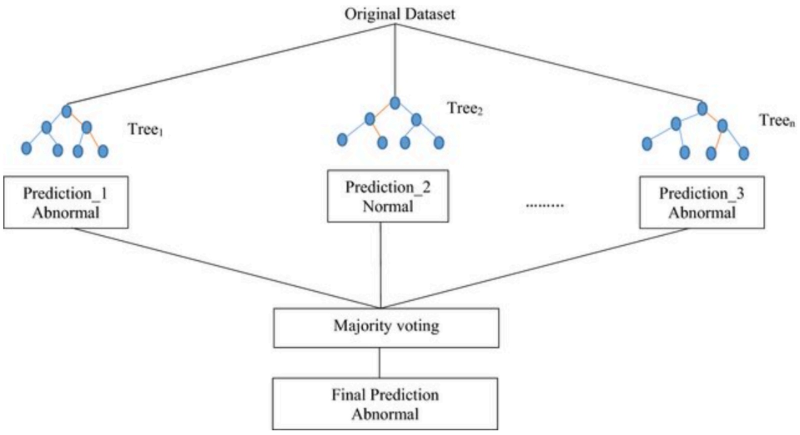


Figure 2.3: Example of a Random Forest with multiple decision trees. Each tree makes a prediction based on the data point, and the final output is determined by aggregating the individual tree predictions (majority vote for classification, average for regression) [21].

2.4. NeuroTree

In this section we describe how the NeuroTree model works and what its architecture looks like according to the description in [62].

The NeuroTree model, like a Random Forest, is a collection of decision trees. Specifically, the decision trees are complete binary trees which means that they do not have any pruned nodes. Figure 2.2 exemplifies a complete binary tree with a depth of two. The key distinction compared to a Random Forest lies in NeuroTree's differentiability, enabling training through first-order gradient-based methods. This is achieved by employing "soft decisions" along each tree path instead of the standard "hard decisions" in traditional trees.

Consider the highlighted path in Figure 2.2 (`node1` → `node3` → `leaf3`). Traditionally, internal nodes make binary choices (`true` or `false`), directing the data point to a specific prediction index (index 3 in this case). An alternative perspective views leaf nodes as offering weights associated with each prediction. The tree's prediction then becomes the weighted sum of leaf values and their corresponding weights. However, with traditional hard decisions at internal nodes, the leaf weights essentially act as a mask (e.g., `[0, 0, 1, 0]` for reaching `leaf3`).

This approach of a weighted sum of the leaves' values and weights is the basis of the predictions in the NeuroTree model. By relaxing the hard conditions into soft ones, the mask now becomes a form of a probability vector where each element of the vector is equal to the weight of the leaf node. This means that the following holds true: $\sum$`(leaf_weights) = 1` and `leaf_weight = [0, 1]`.

The NeuroTree model builds upon the weighted sum concept for predictions. The key lies in relaxing the "hard conditions" at internal nodes into "soft decisions." Consequently, the mask is transformed into a probability vector. Each element represents a leaf node's weight, constrained by the conditions:

- $\sum$`(leaf_weights) = 1` (sum of leaf weights equals 1)
- `leaf_weight` $\in$ `[0, 1]` (each weight lies between 0 and 1)

This probabilistic approach using soft decisions is the foundation of NeuroTree's predictions.

Figure 2.4 illustrates the architecture of a single decision tree within the NeuroTree model. We first focus on the leftmost section, which depicts the internal nodes of the decision tree (from Figure 2.2) as `NW1 - NW3` (node weights). Each `NW` represents the proportion of data points considered `true` based on the corresponding node's decision condition. For instance, a 50/50 split at a node translates to `NW = 0.5`.

We derive the cumulative node weights `CNW1-CNW7` from the `NW` values. As described earlier, each `NW` dictates which fraction of data points is considered `true` and `false` (illustrated by blue and red arrows in

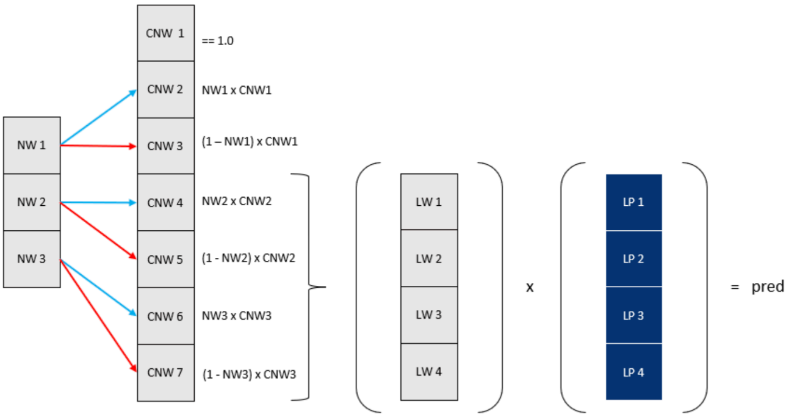


Figure 2.4: Example of how a basic decision tree can be represented as a single differentiable tree with the NeuroTree model [62].

Figure 2.4). Initially, `CNW1 = 1` since all data points pass through the root node. Subsequent cumulative node weights are calculated by multiplying either `NW` or `NW - 1` with the preceding cumulative weight (calculations shown in Figure 2.4).

The resulting vector `CNW1-CNW7` represents the fraction of data points reaching each corresponding node. In the example (the decision tree in Figure 2.2), leaf nodes correspond to `CNW4-CNW7`.

Finally, the tree's prediction is obtained by calculating the dot product of leaf weights `LW1-LW4` and leaf predictions `LP1-LP4`. During NeuroTree training, both leaf weights (adjusting internal node decisions) and leaf predictions (altering class predictions for leaf nodes) are optimized.

Lastly, the final prediction of the NeuroTree model is a combination of all the individual trees in the model, similar to how a Random Forest works. The average of the predictions of the individual trees is calculated to get the final prediction. As summing and averaging are differentiable operations, this differentiability allows the NeuroTree model to be trained using gradient-based methods.

3
Data Imbalance in Classification Tasks

This chapter will elaborate on the existing literature on data imbalance in classification tasks. Firstly, we will discuss the challenges that have been defined in current work. Second, we will discuss the different forms of sampling techniques which could potentially help to overcome these challenges.

The survey by Haixiang et al. [35] summarises the literature on learning from class-imbalanced data. In this paper, the authors state that the challenges of learning from class-imbalanced data based on existing work are fivefold:

1. According to the empirical analysis by López et al. [53], standard classifiers such as logistic regression, Support Vector Machines (SVMs) and decision trees which are suitable for balanced training sets often provide suboptimal classification results when facing imbalanced datasets. The classifier will perform well on the majority examples and perform poorly on the minority examples.
2. In the study by Loyola-González et al. [54], the authors state that guiding the learning process of a model by global performance metrics such as accuracy induces a bias towards the majority class. This results in the rare episodes remaining unknown even if the model has a high overall precision.
3. Beyan and Fisher [9] discuss how rare minority examples may possibly be treated as noise by the classification model and vice versa because both the minority class and noise are rare patterns in the data space.
4. The study by Denil and Trappenberg [19] discusses the overlapping problem under imbalanced datasets. This problem is defined by the fact that minority examples usually overlap wiht other regions where the prior probabilities of both classes are almost identical.
5. According to the study by Branco et al. [11], a lack of density and small sample size with high feature dimensionality is a challenge to imbalanced learning.

To overcome the challenges in class-imbalanced classification tasks, balancing techniques have become an important part of the machine learning systems. In the survey by Haixiang et al. [35], the authors show that resampling techniques are commonly used to rebalance the dataset to alleviate the effect of skewed class distributions in the learning process. The research categorises the techniques into three groups:

- Over-sampling methods: balancing by means of creating new minority class samples. Two widely-used methods are randomly duplicating the minority samples and SMOTE [15]. This technique creates synthetic samples by drawing a line between two minority examples and generating a new data point along that line.
- Under-sampling methods: balancing by means of discarding samples in the majority class. Random Under-Sampling (RUS) is the simplest yet most effective method [78].
- Hybrid methods: these are a combination of an over-sampling and an under-sampling method.

Even though various studies have shown how balancing techniques positively impact the performance of the classification model [54, 90, 61], there has also been some skepticism on the degree to which these techniques improve performance. More specifically, the paper by Prati et al. [69] proposes an experimental design that tests how much of the class imbalance performance loss was recovered by rebalancing the data. In this study, the authors perform performance tests on 22 different datasets with a variety of classification models such as a decision tree, neural network and SVM. First, the datasets are artificially imbalanced to different class ratios ranging from 60/40 to 99/1. Next, the original performance of the classifier is compared to the performance on the imbalanced datasets. Lastly, the authors compare how much of the loss, which was present for all classifiers, could be recovered by the sampling techniques. For the sampling techniques, they opted to use SMOTE and two variations of SMOTE named Borderline-SMOTE [37] and ADASYN [34]. The study concludes that ”the sampling techniques were able to occasionally recover a significant proportion (between 50 and 60%) of the performance lost” [69]. The authors continue by stating that in most cases the recovery was below 30% which according to them can be considered as a quite modest recovery rate.

4

Quality of Explanations

This chapter delves into the existing literature on quality within Explainable Artificial Intelligence (XAI) and Counterfactual Explanations (CEs). First, we will look at how quality of explanations is defined and assessed in the field of XAI. Next, we will try to link the notions of quality in XAI to desiderata for CEs. Specifically, we will look at desiderata such as validity, distance, plausibility and robustness. Additionally, we examine recent surveys in the field of CE. These surveys offer valuable insights into how researchers are currently evaluating and benchmarking quality and more specifically the desiderata mentioned before.

4.1. Quality in XAI

Various literature can be found on what defines quality and explainability in Explainable AI and how it should be evaluated. The works by Vilone et al. [85], Barredo Arrieta et al. [7] and Lipton et al. [52] have tried to summarise this literature and categorise the different notions and goals around explainability. In these works we find that there are certain notions and goals more profoundly found in the literature. These include fidelity, actionability, trustworthiness, causality, transferability and robustness. The remainder of this section presents how different studies define and evaluate these terms in more detail.

Fidelity

The work by Guidotti et al. [31] defines fidelity as the extent to which an interpretable model is able to imitate a black-box predictor. This can be measured by comparing the ouputs of the black-box predictor to the outputs of the interpretable model in terms of accuracy, F1 score, etc. This definition very closely resembles the definition in [59]. In this work the author further states that ”high fidelity is one of the most important properties of an explanation, because an explanation with low fidelity is useless to explain the machine learning model” [59].

Actionability

The term actionability in the field of XAI refers to making sure that users who get explanations can act upon these explanations. The authors of [48] state that ”both theory and prior empirical findings suggest end users will ignore explanations when the benefit of attending to them is unclear”. Therefore, they opt to create actionable explanations in hope of lowering the perceived cost of attending to the explanation.

Trustworthiness

Trustworthiness is an aspect of XAI that has come up in numerous works and in two different forms namely trustworthy models and trustworthy explanations. According to [7], several authors agree upon the fact that the primary goal of XAI should be the search for trustworthy models. These authors include Fox et al. [25], Ribeiro et al. [72] and Došilović et al. [24]. Besides this notion of trustworthiness, we also want to make sure that explanations in of itself are trustworthy. This is commonly linked to the need for causality, transferability and robustness which the following paragraphs delve into.

Causality

According to Vilone et al. [85], several of the papers that were surveyed consider causality as a fundamental attribute of explainability. This is further emphasised by the work of Lipton [52], who says that interpreting models could possibly lead to initially unobserved causal relationships being discovered. The works by [70] and [87] further back up the idea that explainable models might ease the task of finding relationships that could be tested further for a stronger causal link between the involved variables. Additionally, in [33], the authors show that assigning causal attribution to events is part of human nature and systems that provide causal explanations are perceived as more human-like.

Transferability

Transferability in XAI is commonly defined as the capacity of a method to transfer prior knowledge to unfamiliar situations [85]. Both [7] and [52], state the importance of transferability by giving an example of a case where transferability was not achieved, namely the study by Caruana et al. [14]. [7] says that this is an example of a case in which ”the lack of a proper understanding of the model might drive the user toward incorrect assumptions and fatal consequences”. [52] explains how this is an example of ”models being deployed in settings where their use might alter the environment, invalidating future predictions”.

Robustness

In [3] and [4], the authors investigate the robustness of different interpretability methods. They state that intuitively, if the input being explained is modified slightly - subtly enough so as to not change the prediction of the model too much - then we would hope the explanation provided by the interpretability method for that new input does not change much either [4]. Both studies show that currently this is not always the case for state-of-the-art interpretability methods. The notion used by Alvarez-Melis et al. is in line with how Hancox-Li talks about robustness in [38]. In their research, the focus lies on investigating what makes a good explanation. Specifically, the author tries to find out how important it is that explanations reflect real patterns in the data or the world. They argue that robustness is an important aspect of this and that it is desirable to the extent that we’re concerned about finding real patterns in the world [38].

4.2. Quality in CEs

The following sections describe how the notions and goals presented in Section 4.1 are linked to desiderata and evaluation metrics in the field of CE. Moreover, we look at surveys of CEs to find out how these desiderata are being evaluated.

Validity

The validity evaluation metric of CEs which states that a counterfactual is valid if the classification model classifies it in the target class, is closely linked to the fidelity notion found in XAI. According to Altmeyer et al. [2], valid counterfactuals always have full local fidelity because counterfactual explanations work directly with the black-box model. However, the authors also state that full local fidelity is not enough to guarantee faithfull counterfactual explanations. Therefore, in the following sections we will discuss further what makes a good counterfactual explanation.

Distance and Sparsity

In the field of CE, actionability is often linked to feasibility [47] or plausibility. We will explore this link in the following sections and when looking into CE surveys. Besides the link to feasibility, we argue that actionability can also be linked to the distance between factual and counterfactual. This distance, which is sometimes also referred to as cost, is an attribute that is commonly minimised in counterfactual generation methods. The method by Wachter et al. [86] is one of these methods where the distance from the factual gets penalised during the generation process. In this work, the authors state that ”an ideal counterfactual explanation would alter values as little as possible and represent a closest world under which score p' is returned instead of p” where p' is the counterfactual of the factual p. The authors continue by saying that in many situations it will be more informative to provide a diverse set of counterfactual explanations on which case-specific considerations will be relevant rather than a theoretically ideal counterfactual according to a specific distance metric.

Another method which tries to minimise the distance to the counterfactual is the work by Mothilal et al. [60]. Their method, named Diverse Counterfactual Explanations (DiCE), generates multiple counterfactual explanations for each factual because the authors believe that diversity is an important desideratum. Here, the authors talk about proximity instead of distance which refers to how close the counterfactual is to the original input. Another interesting desideratum is sparsity which the authors of [60] argue is closely connected to distance/proximity. Sparsity is here defined as: ”how many features does a user need to change to transition to the counterfactual class”. Mothilal et al. say that intuitively a counterfactual will be more feasible if it changes a fewer number of features. When looking at CE surveys in Section 4.2.1, we will see that sparsity is often part of the desiderata of various counterfactual generation methods.

Even though distance/cost/proximity is an important aspect of counterfactual explanations, various studies that look at robustness have shown that close counterfactuals generally are less robust [23, 77, 64]. This shows that the desiderata might not be independent and in order to achieve good quality counterfactuals, we need to look at a diverse spectrum of desiderata. We go into more detail in the following sections.

Plausibility

The paper by Del Ser et al. [18], talks about generating trustworthy counterfactual explanations. The authors present a framework that balances three objectives: plausibility, the intensity of changes, and adversarial power. In their research, they conclude that this framework improves the overall trustworthiness of the audience in the classification model’s output. This is a clear link between the notion of trustworthiness in XAI and the plausibility desideratum in CE research.

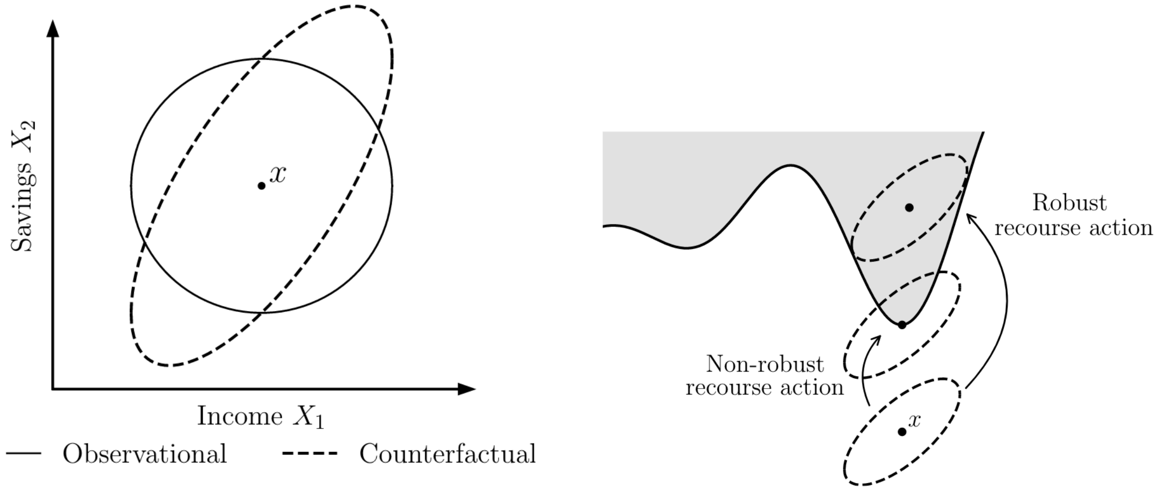
The work by Mahajan et al. [57], links feasibility and plausibility to causality. The authors formulate the challenge of feasibility and plausibility as preserving causal relationships among input features. Furthermore, they state that plausible counterfactuals should respect causal relationships. This link between feasibility, plausibility and causality is also found in the work of Karimi et al. [46]. In this work, the authors state that causal relationships need to be investigated before going from counterfactual explanation to recourse.

In [45] by Joshi et al. the authors propose a recourse algorithm that models the underlying data distribution or manifold. According to the authors, this algorithm results in realistic and actionable recourse and explanations. Thus here we see a link to the actionability notion of XAI. Additionally, the study by Altmeyer et al. [2] mentions the work by Joshi et al. as a study towards more plausible explanations.

In [74], the authors state that counterfactual explanations that focus only on closeness have a great similarity to adversarial examples. They argue that the distinguishing feature between CEs and adversarial examples is interpretability: ”while CEs should interpretable, adversarial examples need not be”. The study defines an interpretable CE as one that is realistic and unambiguous. According to the authors, methods such as the one presented in [45] focus largely on generating realistic CEs but do not consider ambiguity. We argue that this idea of realistic and unambiguous explanations can be seen as aiming towards more plausible explanations. The authors of [74] also make a link to the robustness desideratum. They state that even though their proposed method works with ”any classifier that both offers uncertainty estimates and for which we have access to the gradients”, classifiers that have been retrained using adversarial training can improve the realism of generated explanations. This resembles the notion of robustness in XAI where adversarial examples and small perturbations are an important factor of robustness. In the following section we see how this is applied in more detail in the field of CEs.

Robustness

The notion that we find of robustness in general XAI research, can also be found in research regarding robustness in CEs. Inspired by [4], Artelt et al. formalise robustness as the ability to withstand perturbations in the data [5]. They find that plausible explanations are more robust to small perturbations than the closest explanation. Again, we find that the desiderata of plausibility and robustness are intertwined. Additionally, the authors link the need for robustness and stability to the system’s trustworthiness. They say that ”missing stability and robustness of explanations can lead to unfair explanations and thus compromise the system’s trustworthiness” [5].



(a) Illustration of the individuals similar to x based solely on the feature space (observational) and based on the SCM (counterfactual) [23].

(b) Example of an adversarially robust recourse action [23].

Figure 4.1

The work by Slack et al. [77] also investigates whether counterfactual explanations are vulnerable to small perturbations. The authors show how counterfactual explanations generated by state-of-the-art generators such as Wachter [86] and DiCE [60] can be manipulated. This entails that the generators are not robust and therefore the trustworthiness of counterfactual explanations is brought into question.

In [58], the authors link the robustness desideratum to the notion of transferability. The paper extends the robustness literature by introducing robustness to model changes. The authors claim that when an existing model is updated, it is often desirable that the already provided explanations to individuals should remain valid under the new classification model. For instance, consider an applicant who was denied loan, and the counterfactual explanation provided to the applicant was to increase their income by 10K. Now suppose that they indeed increase their income by 10K and reapply for the loan. If the model is no longer the same, there is no guarantee that their loan will now be approved, leading to potential mistrust and liability concerns for the counterfactual explanations [58].

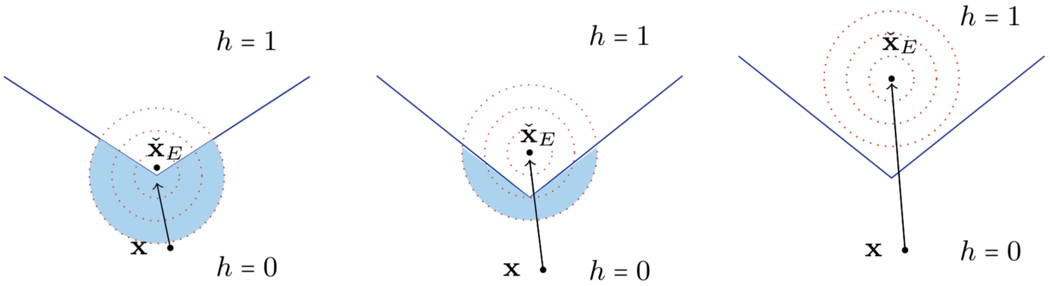
This link between transferability and robustness is also found in [64]. In this work, the authors compare ”sparse” approaches such as Wachter et al. [86] and Mothilal et al. [60] to ”data support” approaches such as Joshi et al. [45] and Mahajan et al. [57] on robustness. They show that ”data support” approaches are more robust and thus more transferable across different classifiers. This also shows the link between distance and robustness, and plausibility and robustness. The ”sparse” approaches, minimising distance, score worse on robustness. Whereas, the ”data support” approaches, leading to more plausible explanations, score better on robustness. The following studies also show how solely minimising distance might lead to worse robustness of CEs.

Dominguez-Olmedo et al. [23] argue that counterfactual explanations, in their work referred to as recourse recommendations, should be robust to modest feature uncertainty for the individual seeking recourse. This property, termed adversarial robustness, is formally defined as:

For some classifier h, individual $x \in X$, and uncertainty set $B(x)$, a recourse action a is adversarially robust if it is valid for all individuals in the uncertainty set $B(x)$.

The uncertainty set encompasses individuals similar to the observed instance x based on a known structural causal model (SCM) [66]. Figure 4.1a illustrates an uncertainty set (dotted line) compared to individuals similar only in the feature space (solid line). It is important to note that here the term ”counterfactual” does not refer to counterfactual explanations of the observed instance x but rather to counterfactuals generated by the SCM. Figure 4.1b presents an example of an adversarially robust recourse action.

The study demonstrates that generators minimizing recourse cost (distance between factual and coun-



(a) Low recourse cost and low recourse robustness (e.g. Wachter et al. [86])

(b) Medium recourse cost and medium recourse robustness (PROBE)

(c) High recourse cost and high recourse robustness (e.g. Dominguez-Olmedo et al. [23])

Figure 4.2: Pictorial representation of the recourses (counterfactuals) output by various state-of-the-art recourse methods and our framework [65]

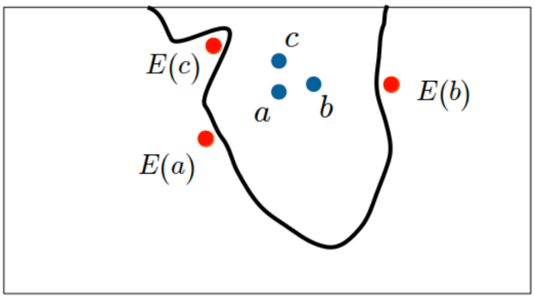


Figure 4.3: Illustration of the intuition behind the Stability notion for three instances *a*, *b* and *c*. [49]

terfactual) often fail to achieve adversarial robustness according to this definition. Furthermore, the authors propose methods for generating adversarially robust recourse for linear and differentiable classifiers.

The research by Pawelczyk et al. [65] extends upon the research by Dominguez-Olmedo et al. [23] by introducing the Probabilistically ROBust rEcourse (PROBE) framework. This method lets the user choose the probability with which a recourse could get invalidated if small changes are made to the recourse. The figures in Figure 4.2 show how the PROBE method compares to a state-of-the-art generator, e.g. the work of Wachter et al. [86], and the approach of Dominguez-Olmedo et al. [23].

The work by Laugel et al. [49] also dives into the need for robustness (in this case referred to as stability) for Counterfactual Explanations and how with the current definition (as defined by Alvarez-Melis et al. in [4]) it is difficult to properly assess whether a counterfactual generator is stable. They argue that it is not clear when an observed variation in explanations is the consequence of a lack robustness of the explainer or of normal variation in the data and in the decision boundary [49]. This can be illustrated with the example shown in Figure 4.3.

In Figure 4.3, we see three data points which are close to each other and thus similar in the data domain. However, the explanations (shown as *E(a)*, *E(b)* and *E(c)*) are very different where especially the explanation of the *b* instance is found in completely different part of the feature domain than the other explanations. This shows that specific information of the data point and the decision boundary of the classification model influence the stability metric. Therefore, while the notion of stability is intuitively important to engender trust to the user, it still lacks a proper definition [49].

4.2.1. Surveys

We look into multiple surveys and benchmarks of Counterfactual Explanations to find out how the aforementioned desiderata are being evaluated and surveyed. We do this by investigating three specific surveys, namely the surveys by Guidotti [31], Verma et al. [84] and Karimi et al. [47].

Guidotti [30] mentions all of the described desiderata explicitly. The author talks about validity as one of the desirable properties for counterfactual explanations. Additionally, almost all the surveyed counterfactual generation methods evaluate validity in their studies. The distance desideratum is here named similarity and the author adds that it is also often referred to as proximity. Furthermore, they describe the minimality or sparsity desideratum with the following logical expression: counterfactual x' of factual$x$ is minimal iif $\nexists x''$ s.t. $|\delta_{x',x''}| < |\delta_{x,x'}|$ where $|\delta_{a,b}|$ is the amount of different attribute value pairs between a and b. Besides mentioning the properties of similarity and minimality as desirable, they are also part of the benchmarking study as evaluation metrics.

The survey also mentions plausibility as one of the desirable properties of a counterfactual explanation. According to the author it is practically defined as a counterfactual not having higher/smaller values that those observable in existing examples, and that the counterfactual should not be labeled as an outlier with respect to existing examples. To test this property in the benchmark study, the author implements an implausibility metric which measures the distance from the counterfactual to the closest instance in the known dataset.

For the robustness desideratum, we find a similar definition as in [4] which is the following: given two similar instances x_1 and x_2 obtaining the same classification from the classifier b, i.e., $y = b(x_1) = b(x_2)$, then an explainer f should return two similar set $C1$, $C2$ of counterfactuals. Moreover, the research performs a benchmarking study which also utilises an instability metric. This metric measures to which extent the counterfactuals C obtained for x are close to the counterfactuals $\bar{C}$ obtained for $\bar{x} \in X$, where $\bar{x}$ is the closest instance to x and $\bar{x}$ receives the same black-box decision of x, i.e., $b(x) = b(\bar{x})$.

In the survey by Verma et al. [84], the authors cite the work by Wachter et al. [86] when talking about the validity desideratum. They state that the objective of a counterfactual explanation should be to minimise the distane between the counterfactual and the original datapoint subject to the constraint that the output of the classifier on the counterfactual is the desired label. This means that they incorporate the distance into the validity desideratum. Sparsity is also mentioned as a desideratum for counterfactual explanations by the authors. In this work, the plausibility desideratum is named Data Manifold closeness. The authors state that ”it would be hard to trust a counterfactual if it resulted in a combination of features that were utterly unlike any observations the classifier has seen before” [84]. Even though robustness is not a part of the initial survey, when discussing research challenges, the paper mentions that there are three kinds of robustness needs:

1. Robustness to model changes when models are retrained.
2. Robustness to the input datapoint (two individuals with a slight change in features should be given similar CFEs).
3. Robustness to small changes in the attained CFE (a CFE with minor changes to the originally suggested CFE should also be accepted).

Moreover, the research presents the current progress when it comes to this challenge. Among other things, they mention the earlier described studies by Slack et al. [77] and Artelt et al. [5].

Lastly, the study presents a list of evaluation metrics that can be used to measure the ease of acting on a recommended counterfactual. This list includes the mentioned desiderate in the form of the following metrics: validity, proximity, sparsity, closeness to the training data.

The survey by Karimi et al. [47] does not explicitly mention the idea of validity in counterfactual explanations. For the distance between counterfactual and factual, the study talks about dissimilarity. Additionally, it presents that distance and cost of actions can not be seen as a general 1-1 mapping. The sparsity desideratum is explicitly mentioned by the authors. They state that ”it is often argued that sparser solutions are desirable as they emphasize fewer changes (in explanations) or fewer variables to act upon (in recommendations) and are thus more interpretable for the individual” [47].

The work summarises existing literature on plausibility in three categories:

1. Domain-consistency, restricts the counterfactual instance to the range of admissible values for the domain of features.
2. Density-consistency, focuses on likely states in the (empirical) distribution of features.

3. Prototypical-consistency, selects counterfactual instances that are either directly present in the dataset or close to a prototypical example.

Lastly, the authors mention robustness in the context of the interplay of recourse and ethical Machine Learning (ML). They state that ”giving the right of recourse to individuals should not be considered in a vacuum and independently of the effect that providing explanations/recommendations may have on other stakeholders (e.g., model deployer and regulators), or in relation to other desirable properties (e.g., fairness, security, privacy, robustness), broadly referred to as ethical ML”. When discussing the current literature describing robustness, previously described works [4, 3, 49, 38] are presented among other studies.

5

Related Work

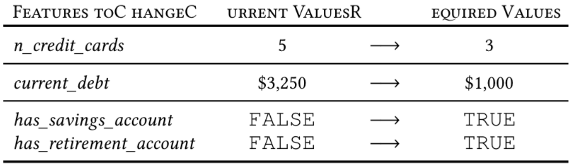
In this chapter we will delve into the existing literature on data imbalance and fairness in the field of recourse and CEs. First, we discuss the role of data imbalance in the context of CE. Next, we look into a collection of studies that evaluate fairness and propose a framework or method for fair recourse/CEs. Here, the goal is to find out how fairness is evaluated by looking at the utilised metrics and the experimental setup.

5.1. Data Imbalance in CE

To the best of our knowledge there exists only one study that compares the quality of CEs under imbalanced datasets. This is the work of Li et al. [51]. The study investigates the effect of class imbalance on counterfactual explanations generated for churn prediction datasets. The authors compare different counterfactual generation methods on different datasets in the churn prediction domain. They conduct experiments on artificially imbalanced datasets and on real-world datasets that are inherently imbalanced. To evaluate the quality of the counterfactuals, the authors report validity, distance, sparsity and a metric which they refer to as credibility and is similar to plausibility metrics in other CE literature. The study concludes by saying that it ”experimentally proves that there are obvious differences in the success rate of finding a counterfactual explanation, the distance between counterfactual explanation and the original instance (i.e. proximity), the proportion of feature change (i.e. sparsity), and the degree of proximity support (i.e. credibility) with the original instance in different instance locations and unbalanced data sets” [51]. However, after our own analysis of the research we believe that these conclusions were too easily drawn. In some cases, the authors draw conclusions based on values that we were not able to find in the reported results.

5.2. Fairness in Recourse and CE

In the work by Ustun et al. [81], the authors state that their tools will allow a range of stakeholders to answer a list of questions. One of these questions is: ”Are there disparaties in recourse between subgroups in the target population?”. The authors propose that fairness can be evaluated with two approaches: cost of recourse and flipsets. The cost of recourse is defined by the authors as a function of percentile shifts per feature. The benefit of this approach is stated as follows: ”Unlike standard



Features toC hangeC	urrent ValuesR		equired Values
n_credit_cards	5	⟶	3
current_debt	$3,250	⟶	$1,000
has_savings_account	FALSE	⟶	TRUE
has_retirement_account	FALSE	⟶	TRUE

Figure 5.1: Example of a flipset for a person who is denied credit by a classification model [81].

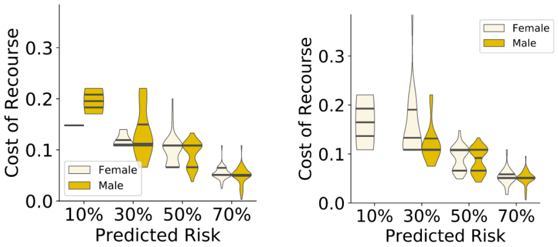
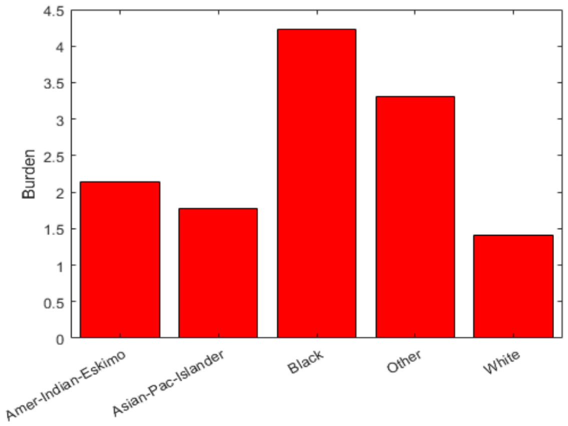


Figure 5.2: Overview of recourse disparities between males and females in the target population. On the top row, we plot the distribution of the cost of recourse for males and females based on their predicted risk and true label: we plot the cost for individuals where =+1 (left) and = 1 (right) [81].

Person	Feature	FitnessM	FitnessU
1	Race	0.63	0.87
2	Gender	0.41	0.62
3	Race	0.81	0.81

(a) Fitness values when race and gender attributes are muted (FitnessM) and unmuted (FitnessU) for three people [75].



(b) Burden on different groups belonging to a particular race in the UCI adult dataset [75].

Figure 5.3

Euclidean distance metrics, cost functions based on percentiles do not depend on the scale of features, and account for the distribution of features in the target population." [81]. Flipsets, on the other hand, are essentially a collection of actions that a person has to take to flip their predicted label (example in Figure 5.1). To assess the disparities in recourse between subgroups, one can compare the cost of recourse and the flipsets between subgroups and see whether differences can be found.

Ustun et al. present an example of an assessment by training a logistic regression model on the `german` dataset [44] and generating individual recourse. In the experiment, the training of the model is done on the dataset without the `gender` feature to assess the disparities between male and female subgroups. The model predicts a risk percentage between 0 and 100% where individuals with a risk percentage above 50% are labeled in the $y_i = -1$ class. The results of the cost of recourse comparison is given in Figure 5.2. The authors conclude that the cost of recourse can vary between males and females in the target population. Furthermore, the authors present the flipsets of a male and female individual with identical true labels and similar predicted risks. Again, the authors state that disparities between the subgroups can be found.

In [75], the authors state that machine learning models need to be rigorously audited for fairness, robustness, transparency, and interpretability. They propose a model-agnostic approach which addresses these issues in unison called CERTIFAI (Counterfactual Explanations for Robustness, Transparency, Interpretability, and Fairness of Artificial Intelligence models). The approach is able to audit fairness from an individual's perspective and from a group's perspective. Both times, the fairness is assessed using the distance from the individual to the counterfactual. In the case of individual fairness, it is evaluated by comparing the fitness of a counterfactual in a 'muted' and 'unmuted' scenario. Fitness refers to the inverse of the distance, meaning a higher fitness equals to a counterfactual closer to the individual. Muting in this case refers to making sure that certain can (unmuted) or can not (muted) be changed. For example, if one would like to calculate the individual fairness based on a gender attribute, they can compare the fitness of the counterfactual generated when gender is muted and unmuted. If the fitness values differ, the individual could claim that the audited model is unfair. To get the group fairness of a specific group, the average distance from an individual in the group to its counterfactual is calculated which the authors refer to as the burden.

An example of a fairness evaluation is given by performing an experiment on the `UCI adult` dataset [8]. Figure 5.3a shows the results for the individual fairness evaluation of three different individuals. Here, the features `race` and `gender` are tested for fairness. The group fairness is evaluated by creating groups based on the `race` feature and comparing the burden of each group. The results are shown in Figure 5.3b.

6

Methodology

This chapter details the methodology employed throughout this work. First, in Section 6.1 we give an overview of the components (datasets, classification models, and counterfactual generators) which are necessary for the experiments. Section 6.2 delves into the design of specific experiments, linking them back to the research questions and hypotheses defined in Section 1.1. This section also discusses the technical components chosen for the experiments with justifications for their selection. Lastly, in Section 6.3 we will present the evaluation metrics and discuss how they are evaluated.

6.1. Experimental Setup

The setup comprises several components: datasets, classification models, and counterfactual generation methods. The following sections will provide a concise overview and explanation of each component. Section 6.2 details the selection of components for each experiment. We deviate from an exhaustive, cartesian product approach by strategically choosing components that directly address the corresponding research question guiding the experiment. Justifications for these targeted selections will be presented within Section 6.2 as well.

6.1.1. Datasets

To conduct the experiments, we will use multiple datasets that belong to the credit-risk domain and the MNIST dataset [50], commonly used for image recognition tasks. We specifically chose the credit-risk domain because it is typically studied in the context of CE and AR. Moreover, the visual aspect of the MNIST dataset allows for better comparisons of generated counterfactuals. A summary of which credit-risk datasets are found in related work is shown in Table 6.1. Based on the usage of datasets in the related work, we have decided to conduct experiments on the German Credit [44], Credit Default [89] and GMSC [17] datasets. Additionally, we have added the Adult [8] dataset because it is particularly interesting to use in the experiments that aim to answer Research Question 4.

Descriptions of the datasets and why they are relevant for our research are given below:

- MNIST: This dataset consists of 70,000 grayscale images of handwritten digits, each resized to

Table 6.1: Summary of credit-risk datasets found in related work.

Dataset	Research Papers
Home Equity Line of Credit (HELOC)[1]	[28]
German Credit [44]	[32] [36] [81]
Credit Default [89]	[32] [36] [45] [81]
HMDA [43]	[36]
Give Me Some Credit (GMSC) [17]	[32] [81]
Propublica[2]	[32]

28x28 pixels. Each pixel value ranges from 0 (black) to 1 (white), representing the pixel intensity. Due to its widespread adoption, standard classification models achieve high accuracy on MNIST. This allows us to focus on counterfactual generation rather than optimizing model performance. Additionally, the visual aspect of the dataset makes it a relevant dataset for our research because it allows for a visual comparison of generated counterfactuals. This visual comparison can also be found in related work, namely in [22], [74], [75] and [83].

- German Credit: This dataset is widely used for credit risk assessment tasks. It comprises 1,000 data points, each representing an individual described by 20 attributes related to their loan application and financial situation. Each data point is labeled as either ”good credit” (0) or ”bad credit” (1), signifying the loan repayment risk associated with the individual. The dataset comes in two forms, one with categorical features and one with solely numerical features. For our research we will work with the numerical form because it makes the preprocessing of the data simpler. The dataset is particularly useful for our research because of its manageable size, allowing for efficient experimentation.
- Credit Default: This dataset is commonly used for investigating credit card default prediction. It consists of approximately 30,000 data points, each representing a credit card customer in Taiwan. Each data point is characterized by 23 attributes, including demographic information, credit card usage statistics, and payment history. Additionally, each data point is labeled as either ”paying customer” (0) or ”defaulted” (1), indicating the customer’s credit card repayment behavior.
- GMSC: This is a benchmark dataset used for credit-risk classification tasks. The dataset consists of 150,000 data points, each representing an individual and characterized by various attributes related to their banking information. The data points are labeled as either ”good credit” (0) or ”bad credit” (1), signifying the loan repayment risk associated with the individual. Moreover, the dataset is imbalanced with only 7% of the data belonging to the ”bad credit” class. This makes GMSC a relevant dataset for our research as we investigate the impact of class imbalance and balancing techniques on counterfactual generation.
- Adult: This widely-used dataset is employed for research on income prediction. It comprises approximately 48,000 data points, each representing an individual described by 14 attributes, including demographic information, education level, work history, and occupation. Each data point is labeled as either ”less than $50K” or ”greater than $50K” annual income. The Adult dataset presents a valuable resource for our research, particularly in the context of the experiments described in Section 6.2.4. Specifically, the dataset’s inclusion of various social features (e.g., education, occupation) facilitates the creation of a negatively biased group. Nevertheless, it is also useful for other experiments because it resembles a real-world situation in the financial domain for which CE’s are interesting.

Data Preprocessing

We keep the MNIST dataset how it comes out-of-the-box because it is already standardised in such a way that all the pixel values are between 0 and 1. For the other datasets we employ the same data preprocessing. We standardise the feature values by fitting an `MLJModels.Standardizer`[3] on each feature column. The code in Listing 6.1 displays the Julia code to perform the preprocessing.

Listing 6.1: Preprocessing using an MLJBase transformer.

```julia
transformer = MLJModels.Standardizer(; count=true)
mach = MLJBase.fit!(MLJBase.machine(transformer, df[:, DataFrames.Not(:target)]))
X = MLJBase.transform(mach, df[:, DataFrames.Not(:target)])
X = Matrix(X)
X = permutedims(X)
```

6.1.2. Classification Models

Based on the datasets presented in Section 6.1.1, we will utilise three different classification models for our research. Two of these are common models in the ML/AI space, namely a Multi-Layer Perceptron (MLP) and a Random Forest model. The last model is called a NeuroTree which combines a Neural Network with a Decision Tree and has been implemented for Julia in [20]. A detailed description of how

[3] https://alan-turing-institute.github.io/MLJ.jl/v0.5/built_in_transformers/#MLJModels.Standardizer

this classification model works can be found in Section 2.4. The following list presents the relevance of these classification models to our research:

- Multi-layer perceptron (MLP): A multi-layer perceptron (MLP) is employed as the primary classification model for the MNIST dataset. Notably, while the MLP achieves satisfactory performance on MNIST which has been investigated in [6], [16] and [67], the research in [12] and [26] shows that it may not be the optimal choice for tabular datasets in the credit risk domain.
- Random Forest: This is a model that is known to generally perform well on tabular datasets in the credit risk domain. Both [12] and [26] present this with various experiments. The down-side of the Random Forest model is that it is not differentiable which limits the options of counterfactual generators since most state-of-the-art generators rely on differentiable classification models.
- NeuroTree: The NeuroTree model combines the differentiability of neural networks with the performance of decision trees on tabular data[4].

6.1.3. Counterfactual Generators

The employed counterfactual generators can be categorized into two categories: gradient-based and non-gradient-based. The gradient-based generators are:

- Wachter [86]: This generator performs gradient descent in the feature space. It generates counterfactuals that achieve the desired outcome whilst making minimal changes to the original data, thus keeping the recourse cost low.
- Greedy [74]: This generator aims to generate unambiguous and realistic counterfactuals by using the predictive uncertainty of the classification model. More specifically, the authors talk about two types of uncertainty: epistemic and aleatoric uncertainty. The research suggests that focusing on counterfactuals where the classifier has low uncertainty of both types leads to more unambiguous and realistic results. To achieve this, they extend the Jacobian-based Saliency Map Attack (JSMA) [63] to find which feature changes result in the best possible counterfactual.
- REVISE [45]: This generator performs gradient descent on the latent space, which is learned through a generative model, instead of the feature space. According to the research, this approach has two advantages:

  1. The generated counterfactuals will be realistic because they will follow the data-generation process encoded in the latent space.
  2. The latent space is a compressed version of the feature space which means the generation process is less costly.

  However, a disadvantage of this generator is the need for a well-specified generative model that can learn the latent embeddings accurately. In our research, we will use a Variational Auto-Encoder (VAE) for this task.

The non-gradient-based generator is:

- FeatureTweak [79]: This generator is specifically designed to generate counterfactuals for binary Random Forest classifiers. It examines the various trees in the forest and tries find which feature tweak results in flipping the prediction of most trees. This will eventually make sure that the classifier will predict the opposite label of the factual.

To facilitate the analysis of the decision threshold's impact and effectively address Research Question 5, we differentiate between generators based on their decision threshold values. This is achieved by appending a suffix of ”-0.5” or ”-0.95” to the generator name throughout the research. For instance, ”REVISE-0.5” denotes the REVISE generator employing a decision threshold of 0.5.

6.2. Experiments

This section details the experimental design employed to address the research questions outlined in Section 1.1. Each subsection focuses on a specific research question and describes the corresponding

[4] A comparison of the NeuroTree implementation with other classification models on various benchmark datasets can be found at `https://github.com/Evovest/MLBenchmarks.jl`

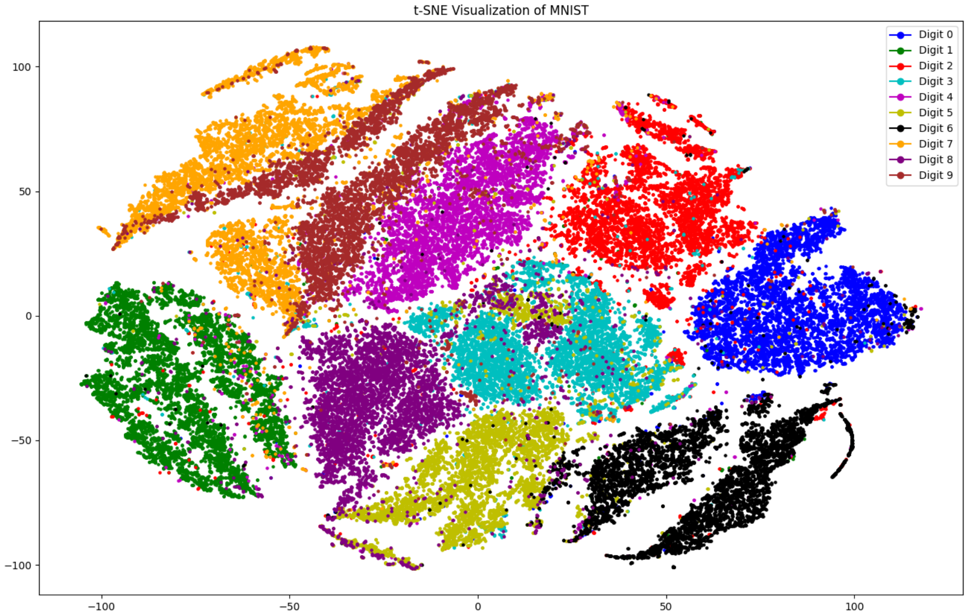


Figure 6.1: Visualisation of the MNIST dataset using t-SNE for dimensionality reduction

experiments designed to investigate it. We present the chosen experimental setup for each experiment, justifying the selection of specific datasets, classification models, and counterfactual generation methods. As described earlier, Research Question 5 does not have a dedicated set of experiments that helps to answer it. Instead, it will be answered throughout the other experiments.

To ensure consistency across all experiments, a uniform approach is adopted. We randomly sample one hundred data points from the factual class. For each data point, a counterfactual is generated within the target class. To mitigate the effects of randomness in data point selection on the overall outcomes, this process is cross-validated five times. During cross-validation, we split the data in a training and test set corresponding to a 80/20 split. The training set is used for all the experimental steps up until the counterfactual generation step and the test set is used to generate counterfactuals. Section 6.3 elaborates on the evaluation methodology employed to assess the quality of the generated counterfactuals.

6.2.1. Inter-class Distance

This experiment addresses Research Question 1:

How does inter-class distance impact the quality of generated counterfactuals?

To address this question, we require a dataset with classes exhibiting varying inter-class distances. We assess the suitability of the MNIST handwritten digit classification dataset [50]. A common technique to analyse this dataset is the t-SNE dimension reduction technique [82]. With this technique we can compress the high-dimensional MNIST data into two dimensions and plot the results.

As observed in Figure 6.1, inter-class distances are not uniform. The separation between classes ”0” and ”1” appears larger compared to the separation between ”7” and ”9”. This aligns with our intuitive understanding of digit similarity: ”7”, and ”9” share visual characteristics, while ”0” and ”1” are visually distinct. These varying inter-class distances make the MNIST dataset well-suited to investigate the impact of inter-class distance on counterfactual quality.

All the other datasets that have been described in 6.1.1 are not suitable for investigating the impact of inter-class distance on counterfactual quality due to their limited number of classes (two). Consequently,

Table 6.2: Hyperparameters used to train the MLP model on the MNIST dataset.

Hyperparameter	Value
Epochs	100
Hidden layers	1
Hidden dimensions	32
Activation function	ReLU
Optimiser	Adam

Table 6.3: Hyperparameters used to train the VAE on the MNIST dataset.

Hyperparameter	Value
Epochs	100
Learning rate	0.0001
Latent dimensions	28
Hidden dimensions	50
Optimiser	Adam

they lack the inter-class distance variation necessary for our analysis.

As detailed in Section 6.1, Multi-Layer Perceptrons (MLPs) achieve good performance on the MNIST dataset. For this research we have trained the MLP with the parameters shown in Table 6.2. This led to a test set accuracy of 97.6% with 5-fold cross-validation. Additionally, Wachter, Greedy, and REVISE generators are all differentiable, making them compatible with the MLP classifier and therefore useful for this experiment. This means that we also have to train a VAE to serve as the generative model for the REVISE generator. The hyperparameters of this model can be found in 6.3. This leads to the following experimental steps:

1. Distance Measurement: We will calculate the pairwise distances between all digits in the MNIST dataset. These distances will be calculated by taking the Euclidean distances between the averages of each class. They will then be analyzed alongside the evaluation metrics to identify potential correlations with counterfactual quality.
2. Counterfactual Generation: Counterfactuals will be generated for each digit pair in the MNIST dataset using the Wachter, Greedy, and REVISE generators. The MLP model will be employed for classification.

6.2.2. Data Imbalance

This experiment addresses Research Question 2:

What is the effect of data imbalance on the quality of counterfactuals?

Evaluating the impact of data imbalance requires a dataset with both imbalanced and balanced distributions such that we can compare the counterfactuals and understand what the effect of data imbalance is. While the GMSC dataset exhibits inherent imbalance, it lacks a balanced counterpart for comparison. In Section 6.2.3, we will see how it, however, is useful for finding out what the effect of balancing techniques is on the quality of counterfactuals. Conversely, the German Credit, Credit Default and Adult datasets offer the flexibility to be manipulated for controlled imbalance. The following steps will be taken to perform this experiment:

1. Introduce Imbalance: Undersampling is employed to introduce imbalance in each of the datasets. The imbalance is introduced in such a way that the resulting class ratio is equal to $90/10$.
2. Train Models: We train a Random Forest and a NeuroTree model on the original and imbalanced versions of each dataset for classification. For the REVISE generator, we train a VAE to serve as the generative model.
3. Counterfactual Generation: Counterfactuals will be generated for all datasets using the Wachter, Greedy, REVISE and FeatureTweak generators. We will use the NeuroTree and Random Forest models for classification.

Table 6.4: The sizes of the original and imbalanced versions of the datasets.

Dataset	Original Size	Imbalanced Size
German Credit	800	622
Credit Default	4000	2222
Adult	26049	21973

Table 6.5: Hyperparameters used train the Random Forest on the data imbalance datasets.

Dataset		Amount of trees	Maximum depth
German Credit	Original	100	50
	Imbalanced	25	25
Credit Default	Original	50	25
	Imbalanced	10	25
Adult	Original	100	50
	Imbalanced	100	50

The following sections describe steps 1 and 2 of this experiment in more detail.

Introduce Imbalance

As described above, we use undersampling to introduce imbalance in each of the datasets. This means that we randomly undersample the positive class such that the desired class ratio of $90/10$ is achieved. Table 6.4 gives an overview of the datasets and their sizes.

Train Models

As described in Section 6.1.2, MLPs do not perform well on tabular data such as the datasets in the credit-risk domain. Random Forest models, on the other hand, have better performance on these types of datasets. The hyperparameters used to train the Random Forest model and the performance results can be found in Table 6.5 and Table 6.6, respectively. Because each dataset has a different size and amount of features, we perform hyperparameter tuning on each dataset.

In the experimental setup in Section 6.1, we discussed how the NeuroTree model can solve the issues of a Random Forest not being differentiable. Additionally, the team behind this model has benchmarked it with various datasets and it has proven to perform similarly to other state-of-the-art classification models[5]. We present the hyperparameters of the NeuroTree in Table 6.7 and the performance of the model in Table 6.8. For the NeuroTree we found with hyperparameter tuning that the same set of parameters performs best on all the datasets therefore the table only presents one set of parameters.

Lastly, for the REVISE generator we have to train a VAE to serve as the generative model. The hyperparameters for training the VAE are shown in Table 6.9.

6.2.3. Balancing Techniques

These experiments address Research Question 3:

How do balancing techniques employed on imbalanced datasets affect the quality of counterfactuals?

With the previous experiments, we aim to find out what the effect is of imbalance on the quality of counterfactuals. As described in Chapter 3, to mitigate the effect of imbalance on the prediction results in classification problems, balancing techniques have become an important part of machine learning. In

[5] `https://github.com/Evovest/MLBenchmarks.jl`

Table 6.6: Performance of the Random Forest on the data imbalance datasets.

Dataset	Accuracy		F1 Score		AUC	
	Original	Imbalanced	Original	Imbalanced	Original	Imbalanced
German Credit	0.743 ± 0.03	0.906 ± 0.01	0.458 ± 0.04	0.186 ± 0.07	0.827 ± 0.02	0.941 ± 0.01
Credit Default	0.78 ± 0.01	0.902 ± 0.01	0.77 ± 0.01	0.163 ± 0.04	0.85 ± 0.01	0.937 ± 0.01
Adult	0.928 ± 0.01	0.943 ± 0.00	0.929 ± 0.01	0.644 ± 0.01	0.977 ± 0.00	0.983 ± 0.00

Table 6.7: Hyperparameters used to train the NeuroTree model on the data imbalance datasets.

Hyperparameter	Value
Epochs	10
Amount of trees	64
Maximum depth	5

Table 6.8: Performance of the NeuroTree model on the data imbalance datasets.

Dataset	Accuracy		F1 Score		AUC	
	Original	Imbalanced	Original	Imbalanced	Original	Imbalanced
German Credit	0.755 ± 0.01	0.905 ± 0.01	0.53 ± 0.02	0.19 ± 0.05	0.83 ± 0.01	0.949 ± 0.01
Credit Default	0.719 ± 0.01	0.902 ± 0.00	0.704 ± 0.01	0.314 ± 0.02	0.797 ± 0.01	0.937 ± 0.00
Adult	0.878 ± 0.00	0.916 ± 0.00	0.883 ± 0.00	0.405 ± 0.07	0.948 ± 0.00	0.974 ± 0.00

this research we opt to use SMOTE, RUS and a hybrid method which combines these techniques and we will refer to as SMOTE/RUS. By answering the research question above, we want to see whether balancing techniques can have a positive effect on the counterfactual generation process. In order for us to do this, we will make use of the datasets and models that have been introduced in Section 6.2.2. We will apply the aforementioned balancing techniques to these imbalanced datasets to find out what the effect is on the quality of the counterfactuals. This leads to the following experimental design:

1. Balance Datasets: We take the imbalanced datasets and balance them using SMOTE, RUS and SMOTE/RUS.
2. Train Models: We train both the Random Forest and the NeuroTree model on the newly balanced datasets. For the REVISE generator, we train a VAE to serve as the generative model.
3. Counterfactual Generation: We generate counterfactuals for the imbalanced and balanced datasets and compare them. The Greedy, Wachter and REVISE generator are used in combination with the NeuroTree model and the FeatureTweak generator is used in combination with the Random Forest.

The following sections describe steps 1 and 2 in more detail.

Balance Datasets

To balance the datasets, we take each of the imbalanced datasets from the previous experiments and apply SMOTE, RUS or a combination of SMOTE/RUS to balance them. Additionally, as mentioned before, the GMSC dataset is also suitable for these experiments. We refer to the original GMSC dataset as imbalanced because it is inherently imbalanced. Even though the class ratio of the GMSC dataset is approximately 93/7, which is not the same as the class ratio of 90/10 found in the manually imbalanced dataset, we keep the dataset as it is. When balancing the datasets we always want to create the same class ratio. For the SMOTE technique we apply a desired class ratio of 50/50. In the case of RUS, we undersample the majority class such that the class ratio becomes 75/25. When both techniques are applied together, we first undersample such that the class ratio becomes 75/25 and then apply SMOTE which leads to a 50/50 class ratio. Table 6.10 gives an overview of all the datasets that we will generate counterfactuals for in these experiments.

Table 6.9: Hyperparameters used to train the VAE on the data imbalance datasets.

Hyperparameter	Value
Epochs	200
Learning rate	0.0001
Latent dimensions	5
Hidden dimensions	2
Optimiser	Adam

Table 6.10: An overview of the dataset sizes used for the balancing techniques experiments.

Dataset	Imbalanced Size	SMOTE Size	RUS Size	SMOTE/RUS Size
German Credit	622	800	248	372
Credit Default	2222	4000	888	1332
Adult	26049	39551	8789	13183
GMSC	96136	178904	26736	40104

Table 6.11: Accuracy of the Random Forest and NeuroTree model on the balanced datasets.

Model	Dataset	Imbalanced	SMOTE	RUS	SMOTE/RUS
Random Forest	German Credit	0.904 ± 0.00	0.819 ± 0.01	0.873 ± 0.01	0.876 ± 0.01
	Credit Default	0.908 ± 0.00	0.867 ± 0.03	0.898 ± 0.00	0.871 ± 0.01
	Adult	0.942 ± 0.00	0.930 ± 0.00	0.913 ± 0.00	0.886 ± 0.00
	GMSC	0.934 ± 0.00	0.851 ± 0.01	0.906 ± 0.00	0.821 ± 0.01
NeuroTree	German Credit	0.905 ± 0.01	0.783 ± 0.03	0.842 ± 0.03	0.710 ± 0.04
	Credit Default	0.902 ± 0.00	0.799 ± 0.01	0.884 ± 0.00	0.810 ± 0.00
	Adult	0.916 ± 0.00	0.808 ± 0.00	0.891 ± 0.00	0.805 ± 0.01
	GMSC	0.931 ± 0.00	0.774 ± 0.02	0.924 ± 0.00	0.779 ± 0.03

Train Models

To train the models for this experiment we utilise the same method as for the Imbalance Experiments described in Section 6.2.2. The hyperparameters employed for training the Random Forest and NeuroTree models on the balanced datasets remain consistent with those used for the original datasets, as documented in Table 6.5 and Table 6.7, respectively. Furthermore, a VAE is trained on the balanced datasets using the hyperparameters specified in Table 6.9. The classification performance of these models is evaluated using accuracy, F1 score, and AUC, which are presented in Table 6.11, Table 6.12, and Table 6.13, respectively.

6.2.4. Negatively Biased Groups

This experiment addresses Research Question 4:

How does the quality of counterfactuals for subgroups in the data which the classifier is biased towards compare to other data points in the same class?

To address this, we first define bias as a higher probability of being classified as the negative class compared to the overall dataset. For example, if our classifier predicts the negative class for 25% of the dataset and there exists a subgroup of data points for which the classifier predicts the negative class 75% of the time, then we can say that the classifier is biased for that subgroup. We identify a subgroup by specifying boundaries for a set of features, e.g., all data points with ”age < 35”. We then compare the quality of counterfactuals generated for the subgroup against those generated for other points within the same class. For this experiment, we utlise the Adult dataset as described earlier in 6.1.1. This leads to the following experimental design:

1. Define Subgroup: We will define a subgroup based on features that according to a SHAP analysis have a high likelihood of getting classified as the positive label.

Table 6.12: F1 Score of the Random Forest and NeuroTree model on the balanced datasets.

Model	Dataset	Imbalanced	SMOTE	RUS	SMOTE/RUS
Random Forest	German Credit	0.146 ± 0.07	0.294 ± 0.05	0.325 ± 0.05	0.337 ± 0.03
	Credit Default	0.277 ± 0.03	0.397 ± 0.04	0.408 ± 0.03	0.386 ± 0.02
	Adult	0.637 ± 0.01	0.646 ± 0.01	0.611 ± 0.01	0.564 ± 0.01
	GMSC	0.268 ± 0.03	0.345 ± 0.01	0.431 ± 0.02	0.346 ± 0.01
NeuroTree	German Credit	0.190 ± 0.05	0.303 ± 0.03	0.318 ± 0.09	0.308 ± 0.02
	Credit Default	0.314 ± 0.02	0.315 ± 0.01	0.442 ± 0.02	0.367 ± 0.00
	Adult	0.405 ± 0.01	0.448 ± 0.01	0.514 ± 0.02	0.456 ± 0.01
	GMSC	0.269 ± 0.04	0.308 ± 0.02	0.396 ± 0.00	0.309 ± 0.02

Table 6.13: AUC of the Random Forest and NeuroTree model on the balanced datasets.

Model	Dataset	Imbalanced	SMOTE	RUS	SMOTE/RUS
Random Forest	German Credit	0.744 ± 0.03	0.722 ± 0.04	0.737 ± 0.07	0.732 ± 0.02
	Credit Default	0.773 ± 0.02	0.759 ± 0.03	0.772 ± 0.02	0.755 ± 0.02
	Adult	0.927 ± 0.01	0.930 ± 0.01	0.923 ± 0.01	0.918 ± 0.00
	GMSC	0.849 ± 0.01	0.821 ± 0.02	0.853 ± 0.03	0.840 ± 0.01
NeuroTree	German Credit	0.949 ± 0.01	0.875 ± 0.02	0.931 ± 0.02	0.787 ± 0.03
	Credit Default	0.937 ± 0.00	0.864 ± 0.02	0.932 ± 0.00	0.826 ± 0.00
	Adult	0.973 ± 0.00	0.914 ± 0.00	0.962 ± 0.00	0.908 ± 0.01
	GMSC	0.971 ± 0.00	0.815 ± 0.02	0.966 ± 0.00	0.810 ± 0.03

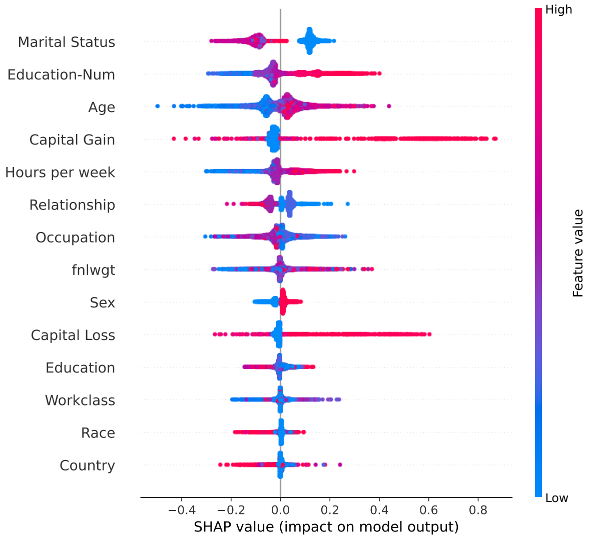


Figure 6.2: SHAP analysis of the Adult dataset.

2. Create Biased Classifier: Based on the subgroup, we will modify the dataset to make sure that the classifier becomes negatively biased. This will also be tested by predicting the labels of newly generated data points that belong to the subgroup.
3. Counterfactual Generation: Due to a limitation of time, we will only generate counterfactuals with the Greedy, Wachter and REVISE generators and the NeuroTree model as classifier.

The following sections describe steps 1 and 2 in more detail.

Define Subgroup

We perform a SHAP analysis [55] on the Adult dataset to identify features with high positive importance values for the positive class. The SHAP framework assigns each feature an importance value for a particular prediction, allowing us to visualize how each feature value influences the prediction outcome. Figure 6.2 presents the SHAP analysis results for the Adult dataset.

In Figure 6.2, we observe a clear positive importance value when the feature `Marital Status` has a low value (corresponding to `Never-married` in the original dataset). Similarly, a high value in the `Education-Num` feature (corresponding to `Bachelors, Masters, Prof-school, and Doctorate`) also exhibits positive importance. Based on these features, we define a subgroup where 80% of the data points belong to the positive class, serving as a starting point for our biased classifier.

Create Biased Classifier

To induce bias towards the defined subgroup, we remove all subgroup data points with negative labels. This enforces a positive association between the subgroup and the positive class within the modified dataset, which is subsequently split into training and testing sets. Bias evaluation is conducted in two ways:

1. Predicting labels of test set subgroup data points: This results in 100% classification as the positive class.
2. Predicting labels of newly generated subgroup data points: To further assess bias, we generate 100 data points with values that belong to the subgroup for the defining features. Random values based on the mean and standard deviation of each feature are used for the remaining features. Our biased classifier predicts the negative class for approximately 70% of these generated data points, significantly higher than the overall dataset's 20% negative classification rate. This confirms the classifier's bias towards the subgroup.

6.3. Metrics and Evaluation

This section will first give an overview of all the evaluation metrics which we assess to answer the research questions. Secondly, we will discuss how we aim to evaluate the metrics both visually and statistically.

6.3.1. Metrics

Our approach leverages a combination of established metrics and novel approaches. Firstly, we will discuss the established metrics in the CounterfactualExplanations.jl package [1]. These are the metrics that are readily available in the package:

1. validity: assesses whether the generated counterfactual is accurately classified as the target class by the classification model.
2. distance: the distance between the factual and the counterfactual in the feature space.
3. redundancy: the proportion of features that remain unchanged between the original data point and the counterfactual.

In the literature that was discussed in Chapter 4, we found that counterfactuals are often evaluated on validity, distance and sparsity. Therefore, the already implemented metrics of the CounterfactualExplanations.jl package are suitable and will be used in this research.

Additionally, we found that plausability is another important way of comparing the quality of counterfactuals. It assesses how closely the counterfactual resembles other data points within the target class. We discussed that various research has tried to specify this metric in more detail [30, 45, 68, 74]. However, there is no consensus yet on how to exactly measure it. In this research, we will start by looking at how the authors of [2] have defined implausability. We choose to look at this definition because to our understanding it is the latest research on implausibility for CE's. This means that the authors incorporate the findings of other research surrounding this topic. They specifically focused on the work by Guidotti et al [30]. In this research, an implausibility metric is proposed that that measures the distance of the counterfactual from its nearest neighbour in the target class. In [2], however, the metric is defined by the average distance of a sampled set of nearest neighbours. The authors propose that this way they avoid the risk that the nearest neighbour of the counterfactual itself is not plausible. Equation (6.1) presents the metric's mathematical notation.

$$\text{impl}(x', X_{y^+}) = \frac{1}{|X_{y^+}|} \sum_{x \in X_{y^+}} \text{distance}(x', x) \tag{6.1}$$

- X_{y^+} is the set of randomly sampled data points in the target class.
- x' represents the counterfactual.
- distance is some form of distance metric.

The implausability metric can be implemented using two approaches:

1. Label-based: taking the sampled points based on the original labels.
2. Prediction-based: taking the sampled points based on the predictions of the classification model

[2] utilises the label-based approach to ensure that the generated counterfactuals comply with the true and unobserved data-generating process (DGP). However, a limitation is that the underlying data

distribution does not reflect how the classifier predicts new data, especially if the model struggles to accurately predict the target class. Therefore, we introduce the prediction-based approach which reveals if generated counterfactuals align with the model's perception of the target class. However, it is susceptible to the model's performance limitations. A poorly performing classifier could lead to misleading implausibility scores for counterfactuals.

As described in Chapter 4 when discussing the robustness desideratum, Laugel et al. [49] state that it is difficult to evaluate this desideratum under its current definition. Therefore, we opt to leave robustness out of the evaluation metrics.

Lastly, we introduce a new metric specifically for the inter-class distance experiments. We introduce a relative distance metric which addresses the influence of inter-class distance on the standard distance metric. This metric normalizes the distance between the factual and counterfactual data points by dividing it by the average distance between the factual and target class. This normalization effectively removes the bias introduced by inherent differences between classes.

We prioritize the evaluation of validity for counterfactuals. Distance, redundancy and implausibility metrics are meaningless for invalid counterfactuals. Therefore, we establish validity before proceeding to other evaluations.

7 Results

7.1. Inter-class Distance

In this section we will discuss the results of the experiments that were described in Section 6.2.1 and we try to answer Research Question 1. We analyze the results using two complementary approaches: quantitative evaluation based on metric values and qualitative assessment of the visually-represented counterfactuals. We begin by examining the quantitative results in Figure 7.1, which compares various metrics with inter-class distance. As outlined in Section 6.2, we generate 100 counterfactuals per factual-target pair and perform 5-fold cross-validation. The figures depict lines fitted to the average values obtained across cross-validation runs.

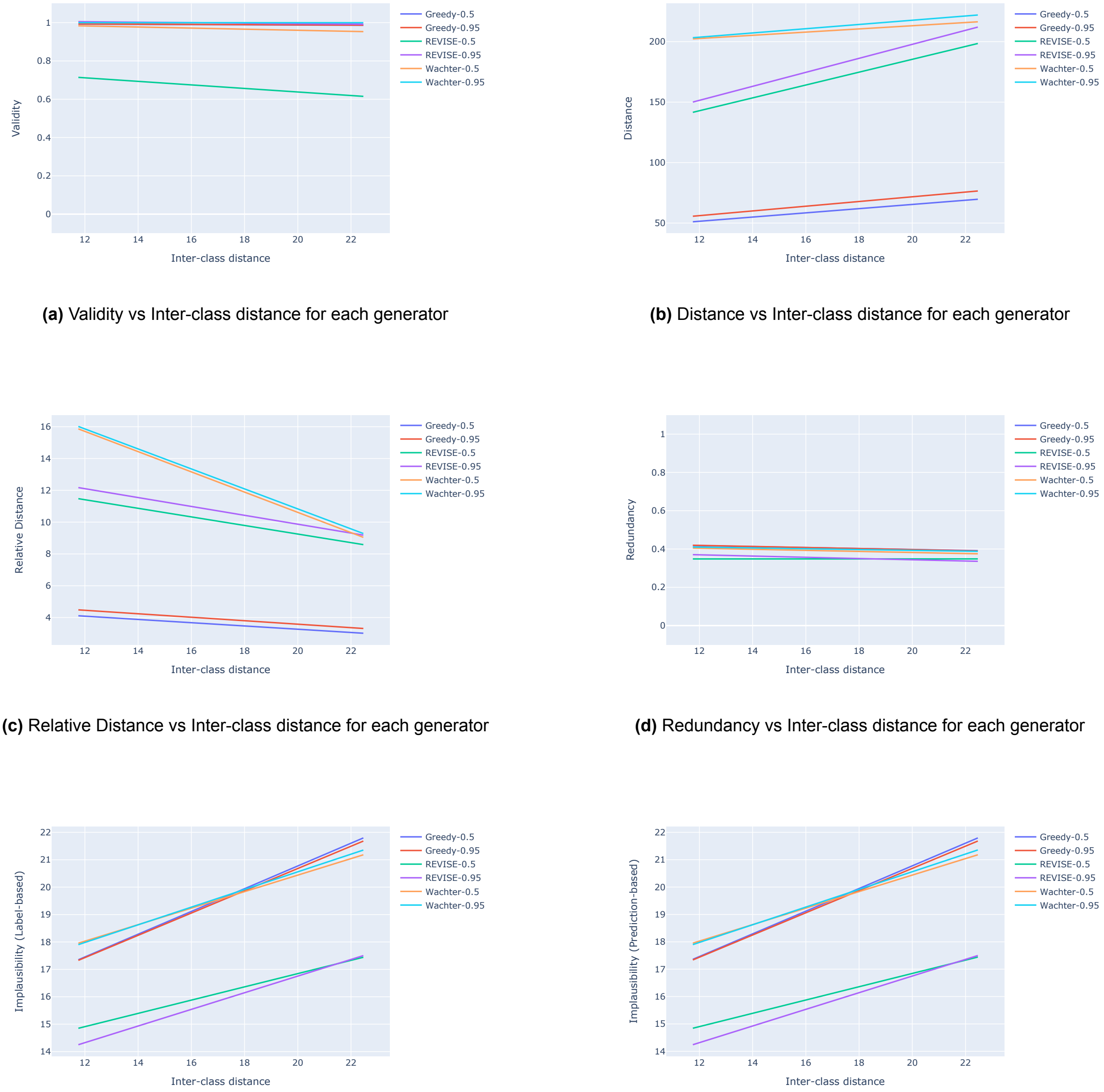
We begin by analyzing the validity metric in Figure 7.1a. This figure demonstrates that all generators, with the exception of the REVISE-0.5 generator, achieve high validity scores. Notably, the REVISE-0.5 generator is the only one clearly affected by inter-class distance, exhibiting a decrease in validity as the distance increases. The other generators are minimally impacted by the change in inter-class distance maintaining very high, near-perfect validity scores across various inter-class distances.

Next, we examine the distance and relative distance metrics presented in Figure 7.1b and Figure 7.1c, respectively. As expected, the distance scores increase with a larger inter-class distance, although differences in the actual distance values between the generators are observed. The results for the relative distance metric demonstrate a decrease as the inter-class distance increases. When considered in combination with the distance scores, this indicates that while the distance between factual and counterfactual increases, it does not increase at the same pace as the inter-class distance does.

For the redundancy metric, presented in Figure 7.1d, the conclusion can be drawn swiftly: the inter-class distance has a negligible effect on redundancy for all generators. We can also try to explain why the redundancy seems to be centered around the value of 0.4. In a 28x28 image, 40% of the pixels is equal to approximately 313 pixels. Notably, the three outermost layers encompass precisely 310 pixels. This observation suggests that, on average, these outer layers represent black pixels across various digits. Consequently, these pixels are likely left unchanged during the counterfactual generation process, contributing to the redundancy values consistently hovering around 0.4.

Lastly, Figure 7.1e and Figure 7.1f present the results for the implausibility label-based metric and implausibility prediction-based metric, respectively. We observe a near-identical pattern in the scores for both metrics. This can be explained by revisiting the performance of the MLP model on this dataset, where it achieved a test set accuracy of 97.6%. This implies an almost perfect classification of the different classes. Consequently, both implausibility metrics are highly similar because the labels and predictions are nearly identical. We further observe an increase in implausibility for all generators as the inter-class distance increases, even though the implausibility is lower for the REVISE generators. This, when combined with the validity scores, suggests that while the various generators can still generate valid counterfactuals with a larger inter-class distance, these counterfactuals are less plausible.

We proceed with a qualitative evaluation by visually inspecting counterfactuals. For each MNIST class



(a) Validity vs Inter-class distance for each generator

(b) Distance vs Inter-class distance for each generator

(c) Relative Distance vs Inter-class distance for each generator

(d) Redundancy vs Inter-class distance for each generator

(e) Implausibility (label-based) vs Inter-class distance for each generator

(f) Implausibility (prediction-based) vs Inter-class distance for each generator

Figure 7.1: A comparison of the CE metrics versus the Inter-class Distance.

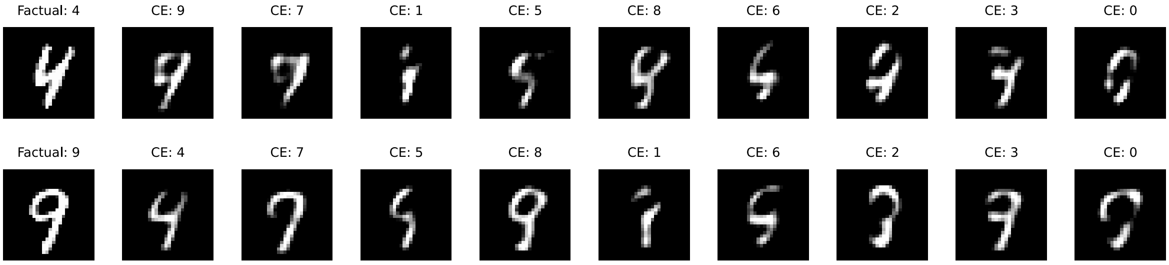


Figure 7.2: Visual representations of counterfactuals generated by the REVISE-0.5 generator.

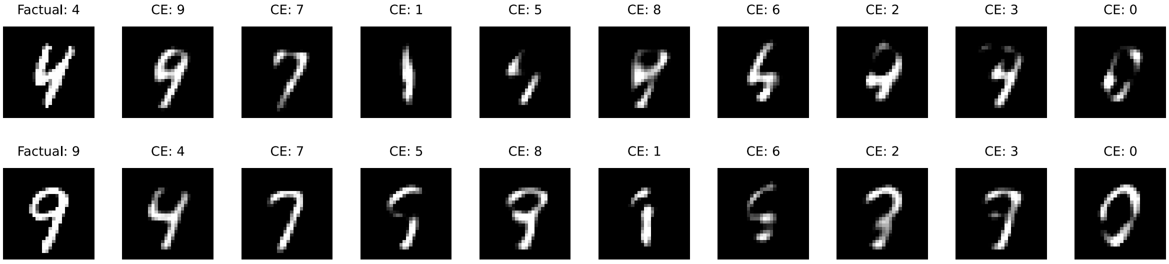


Figure 7.3: Visual representations of counterfactuals generated by the REVISE-0.95 generator.

digit, we generate and present a visual representation of a counterfactual produced by each generator. These counterfactuals are ordered by their inter-class distance from the factual digit, allowing us to observe how visual quality changes with increasing distance. To ensure conciseness, a representative subset of these visualizations is presented in this section; the complete set can be found in Appendix A.

First we will analyse the counterfactuals generated by the REVISE generators. Figure 7.2 and Figure 7.3 depict counterfactuals for digits 4 and 9, generated by the REVISE-0.5 and REVISE-0.95 generators, respectively. We observe that counterfactuals closer to the factual label (shown to the left in the figures) exhibit higher visual quality, particularly those generated by REVISE-0.95. These counterfactuals are readily classifiable by humans, demonstrating their plausibility. However, as we move towards larger inter-class distances (rightward in the figures), the digits become increasingly malformed and implausible, aligning with our prior quantitative assessment.

Next, we examine visual representations of the counterfactuals generated by the Greedy and Wachter generators which can be found in Figure 7.4, Figure 7.5, Figure 7.6 and Figure 7.7. We immediately observe a generally higher level of implausibility in these counterfactuals compared to those generated by REVISE generators. This is again consistent with the quantitative results. However, these visualizations do not conclusively confirm or refute the trend of increasing implausibility with inter-class distance observed in Figures 7.1e and 7.1f. This is because all counterfactuals appear equally implausible, making visual distinction between them challenging.

Our experiments, designed to investigate Research Question 1, demonstrate that the impact of inter-class distance varies across different metrics. Notably, validity and redundancy metrics exhibit minimal sensitivity to changes in inter-class distance. As anticipated, the distance metric increases with larger inter-class distances, demonstrating a positive correlation. However, the relative distance metric reveals a slower increase in the distance between factual data and counterfactuals compared to the inter-class distance itself. Finally, and most interestingly, inter-class distance has a clear negative effect on implausibility. This results in counterfactuals that are valid but appear increasingly implausible.

Additionally, we can make some initial conclusions regarding Research Question 5. With respect to the effect of the decision threshold on the impact of inter-class distance, we can state that it is minimal. Except for the REVISE generator, which showed a clear difference only for the validity, all the other generators did not show any major difference when the decision threshold was changed. This holds for both the quantitative and the qualitative assessment.

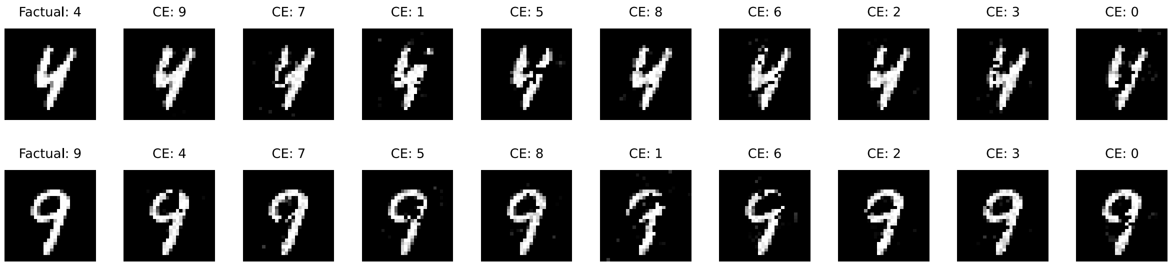


Figure 7.4: Visual representations of counterfactuals generated by the Greedy-0.5 generator.

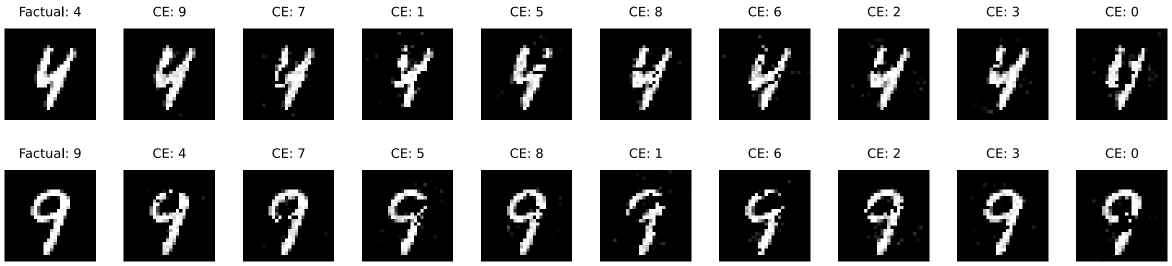


Figure 7.5: Visual representations of counterfactuals generated by the Greedy-0.95 generator.

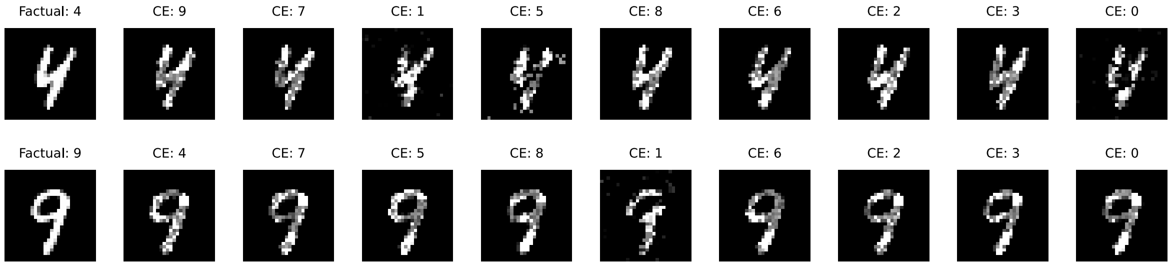


Figure 7.6: Visual representations of counterfactuals generated by the Wachter-0.5 generator.

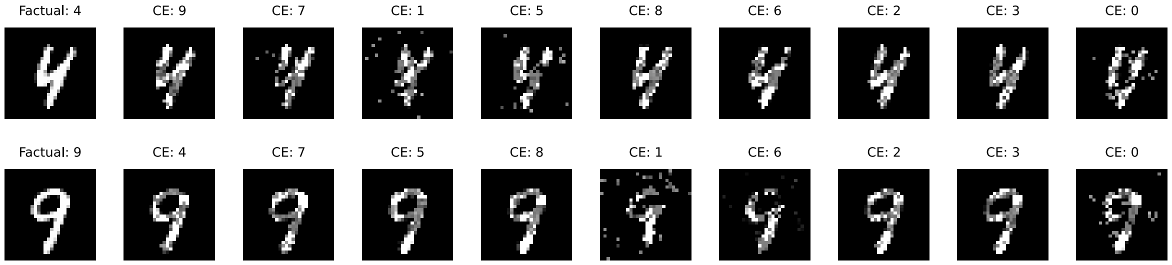


Figure 7.7: Visual representations of counterfactuals generated by the Wachter-0.95 generator.

Table 7.1: The metric scores for the data imbalance experiments.

Data	Model	Generator	Validity ↑		Distance ↓		Implausibility (Label-based) ↓		Implausibility (Prediction-based) ↓	
			Original	Imbalanced	Original	Imbalanced	Original	Imbalanced	Original	Imbalanced
German Credit	NeuroTree	Wachter-0.5	0.992 ± 0.01	0.274 ± 0.06	8.691 ± 0.20	3.444 ± 0.38	6.33 ± 0.10	5.812 ± 0.06	6.334 ± 0.10	2.094 ± 0.36
		Wachter-0.95	0.67 ± 0.09	0.58 ± 0.07	3.646 ± 0.40	6.6 ± 1.17	6.091 ± 0.08	7.451 ± 0.56	6.039 ± 0.09	6.453 ± 0.37
		Greedy-0.5	1.00 ± 0.00	1.00 ± 0.00	2.885 ± 0.08	2.196 ± 0.12	6.263 ± 0.10	6.445 ± 0.26	6.275 ± 0.10	6.47 ± 0.33
		Greedy-0.95	1.00 ± 0.00	1.00 ± 0.00	10.729 ± 0.15	5.545 ± 0.38	6.202 ± 0.08	5.839 ± 0.52	6.201 ± 0.14	5.914 ± 0.47
		REVISE-0.5	1.00 ± 0.00	1.00 ± 0.00	14.059 ± 0.30	17.512 ± 2.39	4.618 ± 0.01	4.593 ± 0.12	4.618 ± 0.01	4.593 ± 0.12
		REVISE-0.95	0.98 ± 0.01	1.00 ± 0.00	15.086 ± 0.69	20.371 ± 1.54	4.575 ± 0.05	4.548 ± 0.43	4.537 ± 0.02	4.535 ± 0.44
	Random Forest	FeatureTweak-0.5	0.93 ± 0.01	1.00 ± 0.00	3.553 ± 0.35	5.59 ± 0.61	6.259 ± 0.72	7.066 ± 1.03	6.215 ± 0.76	7.033 ± 0.88
Credit Default	NeuroTree	Wachter-0.5	0.997 ± 0.01	0.903 ± 0.03	11.097 ± 0.13	9.705 ± 0.07	6.386 ± 0.20	7.481 ± 0.38	6.411 ± 0.14	7.217 ± 0.08
		Wachter-0.95	0.98 ± 0.25	0.76 ± 0.08	11.73 ± 4.77	9.445 ± 2.95	6.478 ± 0.41	6.048 ± 0.38	6.542 ± 0.50	5.808 ± 0.35
		Greedy-0.5	0.937 ± 0.11	1.00 ± 0.00	3.181 ± 0.65	2.939 ± 0.26	5.941 ± 0.12	6.444 ± 0.53	5.99 ± 0.16	6.439 ± 0.51
		Greedy-0.95	1.00 ± 0.00	1.00 ± 0.00	3.507 ± 2.82	4.72 ± 0.97	5.937 ± 0.19	6.322 ± 0.13	5.993 ± 0.21	6.403 ± 0.14
		REVISE-0.5	0.997 ± 0.01	0.997 ± 0.01	13.055 ± 0.09	13.617 ± 1.27	4.216 ± 0.13	5.2 ± 0.93	4.133 ± 0.15	5.151 ± 0.90
		REVISE-0.95	0.98 ± 0.05	1.00 ± 0.00	15.117 ± 1.26	13.307 ± 1.00	4.161 ± 0.13	4.156 ± 0.20	4.153 ± 0.07	4.168 ± 0.19
	Random Forest	FeatureTweak-0.5	0.81 ± 0.02	0.97 ± 0.07	4.424 ± 0.75	3.1 ± 0.51	5.547 ± 0.91	5.758 ± 0.88	5.476 ± 0.92	5.676 ± 1.01
Adult	NeuroTree	Wachter-0.5	0.977 ± 0.01	0.953 ± 0.02	6.454 ± 0.20	5.515 ± 0.27	5.397 ± 0.09	5.594 ± 0.14	5.371 ± 0.11	5.541 ± 0.15
		Wachter-0.95	0.91 ± 0.01	0.87 ± 0.05	5.647 ± 0.39	5.414 ± 0.23	5.311 ± 0.29	5.346 ± 0.26	5.282 ± 0.29	5.334 ± 0.26
		Greedy-0.5	0.98 ± 0.01	0.957 ± 0.03	2.048 ± 0.16	2.156 ± 0.17	5.394 ± 0.13	5.57 ± 0.23	5.369 ± 0.13	5.509 ± 0.22
		Greedy-0.95	0.95 ± 0.01	0.947 ± 0.03	4.283 ± 0.43	4.591 ± 0.08	5.336 ± 0.03	5.594 ± 0.31	5.318 ± 0.03	5.538 ± 0.30
		REVISE-0.5	0.993 ± 0.01	1.00 ± 0.00	8.434 ± 0.54	9.148 ± 0.37	3.961 ± 0.13	3.811 ± 0.05	3.948 ± 0.16	3.775 ± 0.07
		REVISE-0.95	0.993 ± 0.01	1.00 ± 0.00	9.643 ± 0.40	10.386 ± 0.72	3.998 ± 0.20	3.568 ± 0.05	3.989 ± 0.21	3.555 ± 0.04

7.2. Data Imbalance

This section delves into the results of the data imbalance experiments outlined in Section 6.2.2, aiming to address Research Question 2:

What is the effect of data imbalance on the quality of counterfactuals?

We analyze the results for each dataset individually, initially focusing on specific counterfactual generation methods before drawing broader conclusions across all methods. Additionally, we examine the influence of the decision threshold (Research Question 5) within this context. To make this section more concise, we have left out the redundancy metric because all values were either equal or very close to 0.

7.2.1. German Credit

First of all, we take a look at the results for the German Credit dataset. We prioritise the validity metric, as low validity renders other metrics less meaningful (as described in Section 6.3). Greedy, REVISE, and FeatureTweak generators show high validity scores for both the original and imbalanced datasets. Wachter-0.5 shows high validity on the original dataset but severely drops for the imbalanced dataset, potentially indicating that the generators has trouble with handling class imbalance. Interestingly, the implausibility metrics of Wachter-0.5 deviate when comparing label-based and prediction-based calculation methods. This aligns with our expectation that the classifier, specifically in the case of an imbalanced dataset, might perceive the counterfactual as plausible (low prediction-based score), while it remains factually implausible (high label-based score). Wachter-0.95 performs moderately on both datasets, with low validity scores limiting the interpretability of other metrics.

For the Greedy, REVISE, and FeatureTweak generators, the distance metric results are mixed. REVISE and FeatureTweak show larger distances for the imbalanced dataset, suggesting a potential decrease in counterfactual quality. Conversely, Greedy exhibits a shorter distance, indicating potentially better quality. Implausibility generally appears similar between both datasets, with the imbalanced dataset occasionally showing higher values, potentially indicating decreased quality.

In general, the German Credit dataset reveals that Wachter generators struggle with validity under data imbalance, while other methods maintain high performance. Implausibility seems relatively unaffected, and the distance metric shows mixed results. This could be partially attributed to the relatively small size of the German Credit dataset, where the randomness inherent in selecting data points for counterfactual generation might play a more critical role than desired. Therefore, in the context of the German Credit data, data imbalance does not appear to impact the overall quality of generated counterfactuals in a major way.

7.2.2. Credit Default

Next, we discuss the results for the Credit Default dataset. Similar to the German Credit dataset, the Credit Default dataset reveals a clear drop in validity for the Wachter generator under data imbalance. Interestingly, the FeatureTweak generator exhibits a contrasting behavior, with a low validity score for the original dataset but a high score for the imbalanced dataset.

The distance metric generally shows lower values for the imbalanced dataset. This can be partially explained by the inherent effect of class imbalance on classification boundaries. When a dataset is imbalanced, the classifier tends to favor the majority class, shifting the decision boundary closer to the minority class. Consequently, counterfactuals classified as the majority class might appear closer to the original data point. However, we expect this shift to be accompanied by increased implausibility, indicating that these counterfactuals are not necessarily of good quality. Specifically, label-based implausibility should rise as the classifier perceives the counterfactual surrounded by misclassified minority class data points. This trend is observed for some generators, including Wachter-0.5, both Greedy variants, and REVISE-0.5. Other generators show either similar implausibility results across datasets or a decrease in implausibility for the imbalanced dataset. Additionally, the high similarity between both implausibility variants deviates from our initial expectations.

Overall, data imbalance seems to have varying effects on different generators in the Credit Default dataset. Similar to the German Credit analysis, the Wachter generators struggle with validity under imbalance. While implausibility generally increases for the imbalanced dataset for other generators, the distance metric shows a decrease, potentially due to the aforementioned shift in the classification boundary.

7.2.3. Adult

Lastly, we analyse the results of the experiments performed on the Adult dataset. The Adult dataset analysis reveals minimal impact of data imbalance on counterfactual quality across all generators. All methods maintain high validity scores, with only minor differences observed, particularly for the Wachter generators.

Regarding the distance metric, the Wachter generators exhibit a decrease in distance for the imbalanced dataset, potentially indicating closer counterfactuals. Conversely, the other generators show an increase in distance. However, these changes are not large enough to draw definitive conclusions.

Implausibility metrics remain remarkably similar across both datasets and generator variants. This suggests that data imbalance does not affect the perceived plausibility of generated counterfactuals in this context.

Generally speaking, the Adult dataset analysis indicates that data imbalance has the least noticeable effect on counterfactual quality compared to the other datasets. While minor variations exist in the distance metric, the overall impact is negligible. Combined with the consistent implausibility scores, we can conclude that data imbalance does not seem to drastically influence the quality of counterfactuals generated for the Adult dataset.

7.2.4. Conclusion

Combining the findings across all datasets, we can conclude that data imbalance has a minimal effect on the quality of generated counterfactuals, as addressed in Research Question 2. While some specific combinations of generators and datasets exhibit slight variations in certain metrics, these changes lack consistent patterns and are not considerable enough to definitively attribute them to data imbalance itself. Therefore, we can generally state that data imbalance does not substantially impact the quality of counterfactuals in this study.

An additional consideration is the direction of counterfactual generation. In our experiments, we focused on generating counterfactuals for the minority class towards the majority class. This leverages the knowledge of the majority class, which is typically well-represented in the data. Future work could explore the mirrored scenario, where counterfactuals are generated from the majority class to the minority class. This poses a potentially greater challenge, as the generator would need to learn a more nuanced understanding of the minority class from a limited number of samples. This is further discussed in Section 9.1.

Furthermore, our analysis of the decision threshold's influence revealed a clear impact on the Wachter generator. As the decision threshold increases, data imbalance affects this generator more noticeably. However, such a clear relationship was not observed for other generators. Overall, the decision

Table 7.2: The validity and distance scores for the balancing techniques experiments.

Data	Model	Generator	Validity ↑				Distance ↓			
			Imbalanced	SMOTE	RUS	SMOTE/RUS	Imbalanced	SMOTE	RUS	SMOTE/RUS
German Credit	NeuroTree	Wachter-0.5	0.274 ± 0.06	1.00 ± 0.00	0.946 ± 0.03	0.91 ± 0.01	3.444 ± 0.38	9.518 ± 0.19	7.62 ± 0.36	8.26 ± 0.18
		Wachter-0.95	0.517 ± 0.07	0.807 ± 0.177	0.33 ± 0.217	0.357 ± 0.13	5.369 ± 1.17	8.291 ± 1.43	3.213 ± 1.05	3.502 ± 0.50
		Greedy-0.5	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	2.196 ± 0.12	2.219 ± 0.07	2.515 ± 0.06	2.616 ± 0.04
		Greedy-0.95	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	5.811 ± 0.38	4.763 ± 0.39	10.706 ± 0.08	9.316 ± 0.68
		REVISE-0.5	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	17.512 ± 2.39	14.285 ± 0.88	14.959 ± 1.01	14.487 ± 0.70
		REVISE-0.95	1.00 ± 0.00	0.99 ± 0.02	1.00 ± 0.00	0.987 ± 0.02	20.08 ± 1.54	15.902 ± 0.91	15.182 ± 1.70	14.736 ± 0.84
	Random Forest	FeatureTweak-0.5	1.00 ± 0.00	0.32 ± 0.11	0.78 ± 0.15	0.56 ± 0.1	5.59 ± 0.55	2.828 ± 0.25	3.278 ± 0.29	3.206 ± 0.32
Credit Default	NeuroTree	Wachter-0.5	0.903 ± 0.03	1.00 ± 0.00	0.957 ± 0.01	0.547 ± 0.04	9.705 ± 0.74	10.787 ± 0.38	11.453 ± 0.33	4.263 ± 0.17
		Wachter-0.95	0.715 ± 0.08	0.773 ± 0.12	0.425 ± 0.11	0.62 ± 0.24	8.45 ± 2.95	5.402 ± 1.39	2.693 ± 0.06	3.759 ± 0.79
		Greedy-0.5	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.87 ± 0.23	2.939 ± 0.26	2.707 ± 0.01	2.876 ± 0.18	3.526 ± 1.33
		Greedy-0.95	1.00 ± 0.00	1.00 ± 0.00	0.98 ± 0.03	0.963 ± 0.06	5.836 ± 0.97	4.564± 0.50	10.324 ± 1.09	7.529 ± 2.01
		REVISE-0.5	0.997 ± 0.01	0.973 ± 0.03	1.00 ± 0.00	0.987 ± 0.02	13.617 ± 1.27	14.859 ± 0.72	12.508 ± 1.25	13.169 ± 0.69
		REVISE-0.95	1.00 ± 0.00	0.973 ± 0.05	0.973 ± 0.03	0.717 ± 0.12	14.316 ± 1.00	17.947 ± 1.61	14.387 ± 0.89	13.569 ± 1.03
	Random Forest	FeatureTweak-0.5	0.97 ± 0.01	0.14 ± 0.01	0.77 ± 0.05	0.22 ± 0.03	3.1 ± 0.32	2.982 ± 0.21	3.166 ± 0.34	3.025 ± 0.28
Adult	NeuroTree	Wachter-0.5	0.953 ± 0.02	0.987 ± 0.01	0.96 ± 0.01	0.987 ± 0.02	5.515 ± 0.27	6.448 ± 0.26	6.084 ± 0.08	6.315 ± 0.08
		Wachter-0.95	0.87 ± 0.05	0.883 ± 0.04	0.587 ± 0.07	0.767 ± 0.13	5.414 ± 0.23	6.124 ± 0.36	2.901 ± 0.18	4.833 ± 0.60
		Greedy-0.5	0.957 ± 0.03	0.993 ± 0.01	0.987 ± 0.01	0.993 ± 0.01	2.156 ± 0.17	1.177 ± 0.04	1.723 ± 0.07	1.797 ± 0.124
		Greedy-0.95	0.937 ± 0.03	0.983 ± 0.01	0.98 ± 0.01	0.993 ± 0.01	4.591 ± 0.08	3.263 ± 0.11	5.831 ± 0.85	3.843 ± 0.18
		REVISE-0.5	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.02	9.148 ± 0.37	8.081 ± 0.46	8.589 ± 0.63	8.17 ± 0.48
		REVISE-0.95	1.00 ± 0.00	0.987 ± 0.02	0.99 ± 0.02	1.00 ± 0.00	10.368 ± 0.72	8.797 ± 0.38	10.391 ± 0.46	9.631 ± 0.23
GMSC	NeuroTree	Wachter-0.5	0.93 ± 0.02	0.993 ± 0.01	0.967 ± 0.01	0.99 ± 0.02	3.68 ± 0.04	3.91 ± 0.233	4.203 ± 0.09	4.067 ± 0.32
		Wachter-0.95	0.9 ± 0.03	0.993 ± 0.01	0.557 ± 0.24	0.843 ± 0.18	3.688 ± 0.10	4.578 ± 0.53	2.764 ± 1.29	3.563 ± 1.24
		Greedy-0.5	0.933 ± 0.02	0.99 ± 0.01	0.97 ± 0.03	0.987 ± 0.01	1.609 ± 0.14	1.155 ± 0.08	1.409 ± 0.22	1.163 ± 0.06
		Greedy-0.95	0.933 ± 0.05	0.993 ± 0.01	0.977 ± 0.02	0.993 ± 0.01	2.8 ± 0.5	1.683 ± 0.09	9.482 ± 0.41	2.197 ± 0.22
		REVISE-0.5	0.96 ± 0.05	0.963 ± 0.06	0.97 ± 0.02	0.993 ± 0.01	3.941 ± 0.31	4.216 ± 0.75	4.851 ± 0.21	4.117 ± 0.09
		REVISE-0.95	0.95 ± 0.03	0.92 ± 0.06	0.978 ± 0.02	0.81± 0.03	4.709 ± 0.43	6.023 ± 0.96	5.207 ± 0.16	4.377 ± 0.21

Table 7.3: The implausibility scores for the balancing techniques experiments.

Data	Model	Generator	Implausibility (Label-based) ↓				Implausibility (Prediction-based) ↓			
			Imbalanced	SMOTE	RUS	SMOTE/RUS	Imbalanced	SMOTE	RUS	SMOTE/RUS
German Credit	NeuroTree	Wachter-0.5	5.812 ± 0.06	6.418 ± 0.06	6.201 ± 0.07	6.27 ± 0.09	2.094 ± 0.36	6.453 ± 0.06	6.235 ± 0.09	6.244 ± 0.07
		Wachter-0.95	6.816 ± 0.56	6.524 ± 0.08	6.41 ± 0.45	6.136 ± 0.29	6.171 ± 0.37	6.534 ± 0.57	6.297 ± 0.57	6.128 ± 0.32
		Greedy-0.5	6.445 ± 0.36	6.308 ± 0.2	6.721 ± 0.54	6.257 ± 0.12	6.47 ± 0.33	6.35 ± 0.15	6.749 ± 0.5	6.315 ± 0.9
		Greedy-0.95	6.207 ± 0.52	6.37 ± 0	6.844 ± 0.24	6.451 ± 0.15	6.246 ± 0.47	6.406 ± 0.05	6.879 ± 0.2	6.458 ± 0.27
		REVISE-0.5	4.593 ± 0.12	4.614 ± 0.1	4.705 ± 0.01	4.599 ± 0.02	4.593 ± 0.12	4.629 ± 0.06	4.695 ± 0.04	4.584 ± 0.08
		REVISE-0.95	4.829 ± 0.43	4.702 ± 0.11	4.674 ± 0.01	4.605 ± 0.04	4.813 ± 0.44	4.694 ± 0.14	4.664 ± 0.04	4.577 ± 0.12
	Random Forest	FeatureTweak-0.5	7.066 ± 0.62	6.134 ± 0.44	6.34 ± 0.28	6.168 ± 0.23	7.033 ± 0.69	5.994 ± 0.19	6.202 ± 0.25	6.151 ± 0.32
Credit Default	NeuroTree	Wachter-0.5	7.481 ± 0.2	6.212 ± 0.11	6.766 ± 0.2	6.221 ± 0.09	7.217 ± 0.08	6.23 ± 0.1	6.713 ± 0.17	6.223 ± 0.08
		Wachter-0.95	6.522 ± 0.38	5.747 ± 0.17	5.802 ± 0.06	5.771 ± 0.25	6.226 ± 0.35	5.717 ± 0.21	5.352 ± 0.23	5.716 ± 0.42
		Greedy-0.5	6.444 ± 0.53	5.612 ± 0.26	6.312 ± 0.07	5.596 ± 0.22	6.439 ± 0.51	5.616 ± 0.21	6.342 ± 0.06	5.658 ± 0.23
		Greedy-0.95	6.452 ± 0.13	5.793 ± 0.23	6.736 ± 0.4	5.8 ± 0.08	6.474 ± 0.14	5.807 ± 0.27	6.772 ± 0.43	5.853 ± 0.1
		REVISE-0.5	5.2 ± 0.93	4.83 ± 0.24	4.638 ± 0.15	4.383 ± 0.13	5.151 ± 0.9	4.788 ± 0.27	4.636 ± 0.15	4.34 ± 0.06
		REVISE-0.95	4.232 ± 0.2	4.607 ± 0.33	4.581 ± 0.28	4.413 ± 0.32	4.224 ± 0.19	4.535 ± 0.27	4.505 ± 0.26	4.318 ± 0.36
	Random Forest	FeatureTweak-0.5	5.758 ± 0.25	5.47 ± 0.21	5.621 ± 0.19	5.599 ± 0.33	5.676 ± 0.18	5.028 ± 0.27	5.532 ± 0.14	5.211 ± 0.11
Adult	NeuroTree	Wachter-0.5	5.594 ± 0.14	5.162 ± 0.13	5.054 ± 0.05	4.982 ± 0.13	5.541 ± 0.15	5.284 ± 0.14	5.066 ± 0.07	5.08 ± 0.13
		Wachter-0.95	5.346 ± 0.26	4.861 ± 0.05	4.985 ± 0.11	4.849 ± 0.04	5.334 ± 0.26	4.912 ± 0.05	4.896 ± 0.14	4.833 ± 0.09
		Greedy-0.5	5.57 ± 0.23	4.787 ± 0.04	5.025 ± 0.13	4.886 ± 0.11	5.509 ± 0.22	4.97 ± 0.03	5.056 ± 0.13	5.057 ± 0.09
		Greedy-0.95	5.594 ± 0.31	4.823 ± 0.05	4.94 ± 0.18	4.835 ± 0.09	5.538 ± 0.3	4.95 ± 0.03	4.944 ± 0.17	4.946 ± 0.08
		REVISE-0.5	3.811 ± 0.05	3.903 ± 0.25	3.649 ± 0.03	3.739 ± 0.11	3.775 ± 0.07	3.98 ± 0.25	3.673 ± 0.04	3.847 ± 0.12
		REVISE-0.95	3.568 ± 0.05	3.796 ± 0.21	3.647 ± 0.03	3.65 ± 0.11	3.555 ± 0.04	3.848 ± 0.23	3.63 ± 0.04	3.681 ± 0.13
GMSC	NeuroTree	Wachter-0.5	6.08 ± 0.25	3.211 ± 0.3	4.019 ± 0.5	3.335 ± 0.3	5.91 ± 0.27	3.179 ± 0.28	3.992 ± 0.48	3.236 ± 0.3
		Wachter-0.95	5.254 ± 0.87	3.331 ± 0.11	4.439 ± 0.25	3.912 ± 0.23	5.152 ± 0.83	3.255 ± 0.11	4.668 ± 0.1	3.791 ± 0.21
		Greedy-0.5	5.959 ± 0.76	3.496 ± 0.59	4.517 ± 1.08	3.496 ± 0.75	5.843 ± 0.72	3.448 ± 0.6	4.463 ± 1.1	3.415 ± 0.74
		Greedy-0.95	5.404 ± 1.9	3.2 ± 0.24	4.641 ± 0.98	3.481 ± 0.22	5.239 ± 1.77	3.169 ± 0.24	4.591 ± 0.95	3.432 ± 0.24
		REVISE-0.5	4.129 ± 2.27	2.666 ± 0.18	3.491 ± 0.64	2.502 ± 0.01	4.006 ± 2.14	2.616 ± 0.15	3.388 ± 0.59	2.406 ± 0.05
		REVISE-0.95	4.498 ± 1.23	3.752 ± 0.49	3.084 ± 0.55	2.532 ± 0.36	4.329 ± 1.11	3.734 ± 0.52	3.008 ± 0.55	2.49 ± 0.39

threshold's influence on the effect of data imbalance appears limited.

7.3. Balancing Techniques

In this section we will analyse the results of the experiments that try to answer Research Question 3:

How do balancing techniques employed on imbalanced datasets affect the quality of counterfactuals?

Similar to the data imbalance analysis, we discuss the results for each dataset individually before drawing broader conclusions. As this research focuses on the general impact of balancing techniques rather than specific methods, we will analyze these techniques collectively, only highlighting results from individual techniques if deemed necessary. Additionally, we examine the influence of the decision threshold, as discussed previously (Research Question 5). The results are presented in two tables: table 7.2 summarises validity and distance metrics, while table 7.3 presents the implausibility metrics.

7.3.1. German Credit

We begin by analysing the counterfactuals generated for the German Credit dataset. As with previous analyses, we prioritise the validity metric. Greedy and REVISE generators demonstrate excellent performance when balancing techniques are applied. Wachter-0.5 exhibits a large increase in validity compared to the imbalanced dataset, while Wachter-0.95 performs better with SMOTE but shows lower validity for other techniques. Conversely, FeatureTweak experiences a decrease in validity for all balancing techniques compared to the imbalanced dataset. Due to this low validity, we will not focus on FeatureTweak for this dataset.

The distance metric reveals clear patterns for Greedy and REVISE generators. Greedy generators maintain similar distances with SMOTE and experience an increase with RUS or SMOTE/RUS. REVISE generators show a decrease in distance for all balancing techniques compared to the imbalanced dataset. The results for Wachter generators are less clear. Wachter-0.95's low validity makes drawing conclusions from other metrics difficult. Wachter-0.5 exhibits an increase in distance with balancing techniques, which might be interpreted negatively. However, considering the validity increase, it is possible that valid but further away counterfactuals are preferable to invalid closer ones.

Implausibility metrics are largely unaffected by balancing techniques for most generators. The only notable change is observed for Wachter-0.5, which shows a reversed pattern compared to the data imbalance experiments, with implausibility increasing to the original dataset level. Additionally, the FeatureTweak generator exhibits some differences in implausibility, but due to its low validity, these results are inconclusive.

To conclude, balancing techniques have a minor impact on validity and implausibility for most generators in the German Credit dataset. However, their effect on the distance metric varies depending on the generator used, with either positive or negative changes observed.

7.3.2. Credit Default

Similar to the German Credit dataset, the Credit Default data reveals consistent trends in validity scores. Wachter-0.95 maintains low performance overall, with further decreases observed for RUS and SMOTE/RUS balancing techniques. FeatureTweak also performs poorly, replicating the results observed in the German Credit analysis. Consequently, we will again exclude FeatureTweak from further discussion in this dataset.

For Greedy and REVISE generators, the impact of balancing techniques on validity appears minimal. While a slight decrease in validity is observed for the SMOTE/RUS dataset across both generators, the overall effect is minimal. Wachter-0.5 exhibits a similar trend, suggesting that balancing techniques do not drastically impact the validity of generated counterfactuals in this context.

Analyzing the distance metric proves challenging to draw definitive conclusions. Focusing on generators with consistently high validity (Greedy and REVISE), the results appear inconsistent. SMOTE seems to reduce the distance for Greedy generators, while other techniques increase the value. Conversely, REVISE generators show the opposite pattern, with SMOTE increasing the distance and other techniques reducing or maintaining it.

Implausibility metrics, however, indicate a potential positive effect of balancing techniques. Across all generators, the techniques either have no impact or lead to a reduction in implausibility. While the differences are not substantial, they suggest a potential benefit of balancing techniques in improving counterfactual quality.

In conclusion, the Credit Default dataset analysis suggests a potentially positive impact of balancing techniques on counterfactual quality, primarily through the observed reduction in implausibility. Validity scores remain relatively stable, and the distance metric shows mixed results, making it difficult to draw solid conclusions.

7.3.3. Adult

The Adult dataset analysis reveals generally high or similar validity scores for all generators compared to the imbalanced dataset, with one exception. Applying RUS to the Wachter-0.95 generator results in a drastic drop in validity.

The distance metric shows an increase for Wachter generators when balancing techniques are applied, except for the aforementioned RUS-Wachter-0.95 case, which we disregard due to its low validity. Conversely, Greedy generators exhibit a general decrease in distance with balancing techniques. REVISE generators show either a decrease or no clear difference in distance compared to the imbalanced dataset.

Implausibility metrics display a consistent trend across all generators: balancing techniques either reduce or maintain the implausibility level compared to the imbalanced dataset. While not all observed reductions in implausibility are large, the general trend indicates that balancing techniques could be beneficial for enhancing counterfactual quality.

To summarise, the Adult dataset analysis aligns with the Credit Default analysis, suggesting a generally positive impact of balancing techniques on counterfactual quality. Validity remains stable, implausibility generally decreases, and the distance metric shows mixed results, but the overall trend points towards improvement.

7.3.4. GMSC

Lastly, we present the analysis of the GMSC dataset. Similar to the Adult dataset, validity scores for the GMSC dataset remain generally high, with the recurring exception of Wachter-0.95 when applying RUS. This consistent observation suggests a potential limitation of the Wachter-0.95 generator when dealing with undersampled data. The heavy undersampling of the majority class during RUS might leave insufficient training data for the classifier, hindering its ability to generate valid counterfactuals for new data points. However, the lack of a similar effect on other generators necessitates further investigation.

The distance metric again exhibits mixed results. The exceptionally high distance value for Greedy-0.95 with RUS is particularly noteworthy. This could be attributed to the increased randomness introduced by random undersampling in RUS, potentially leading to outlier results. However, the consistently high distance and low standard deviation in this case suggest a more substantial influence. Further exploration is necessary to find out what is exactly happening. Regarding other distance metrics, Wachter generators show a general increase (excluding the low-validity RUS case), while Greedy generators remain relatively stable except for the outlier mentioned above. REVISE generators also seem to show an increase, although the trend is less clear.

Implausibility metrics align with the Credit Default and Adult datasets, demonstrating a decrease with balancing techniques. This effect appears to be most clearly visible in the GMSC dataset, however. This can potentially be attributed to the large size of the GMSC dataset. This could reduce the randomness induced by the experiments, thus leading to the clearest results.

Our conclusion of the GMSC analysis mirrors previous findings. Validity remains largely unaffected except for specific cases involving RUS and Wachter-0.95. Distance metrics offer mixed results. Implausibility consistently improves with balancing techniques, suggesting a potential benefit for counterfactual quality in this context.

7.3.5. Conclusion

Across all datasets, balancing techniques exhibited a nuanced impact on counterfactual quality. While validity remained mostly stable and distance metrics showed mixed results, a consistent decrease in implausibility suggests potential improvement in the perceived quality of generated counterfactuals.

As concluded for the data imbalance experiments in Section 7.2, changing the direction of counterfactual generation could also be an interesting approach for future work that considers applying balancing techniques in the field of CEs. This is discussed in more detail in Section 9.1.

Moreover, looking at the impact of the decision threshold to answer Research Question 5, we can see that generally the impact of the decision threshold is minimal. However, when looking more closely at the distance and implausibility metrics, we find that a higher decision threshold results in clearer differences between imbalanced datasets and datasets that have been balanced. This is in line with our expectations, which were that changing the decision threshold would result in larger effects on the quality of counterfactuals.

7.4. Negatively Biased Subgroups

Table 7.4 presents the results of the experiment addressing Research Question 4:

Table 7.4: The implausibility scores for the negatively biased subgroup experiments.

Generator	**Validity ↑**		**Distance ↓**		**Implausibility (Label-based) ↓**		**Implausibility (Prediction-based) ↓**	
	Original	Biased	Original	Biased	Original	Biased	Original	Biased
Wachter-0.5	0.983 ± 0.01	0.91 ± 0.05	6.088 ± 0.19	6.617 ± 0.37	5.1 ± 0.29	5.959 ± 0.26	5.075 ± 0.27	5.92 ± 0.28
Wachter-0.95	0.893 ± 0.02	0.867 ± 0.04	5.27 ± 0.09	5.859 ± 0.33	4.984 ± 0.16	5.899 ± 0.29	4.955 ± 0.14	5.865 ± 0.28
Greedy-0.5	0.983 ± 0.02	0.943 ± 0.03	1.829 ± 0.09	3.054 ± 0.34	4.99 ± 0.06	5.946 ± 0.13	4.978 ± 0.06	5.909 ± 0.14
Greedy-0.95	0.97 ± 0.02	0.96 ± 0.03	3.943 ± 0.33	5.311 ± 0.18	4.998 ± 0.23	5.592 ± 0.27	4.963 ± 0.22	5.562 ± 0.26
REVISE-0.5	1 ± 0.00	0.997 ± 0.00	8.303 ± 0.33	10.075 ± 0.38	3.867 ± 0.09	3.878 ± 0.06	3.836 ± 0.06	3.844 ± 0.08
REVISE-0.95	1 ± 0.00	1 ± 0.00	9.42 ± 0.21	11.662 ± 0.2	4.007 ± 0.04	4.159 ± 0.03	3.975 ± 0.03	4.12 ± 0.03

How does the quality of counterfactuals for subgroups in the data which the classifier is biased towards compare to other data points in the same class?

We begin by examining the validity metric. Across all generators, the difference between the original and biased datasets remains negligible. However, the distance metric reveals a clear trend: the distance increases for the biased subgroup when generating counterfactuals with all generators. Furthermore, this difference between original and biased groups generally amplifies as the decision threshold increases. For instance, the REVISE generator exhibits a distance difference of 1.772 between the two groups at a decision threshold of 0.5, which increases to 2.242 at a threshold of 0.95.

The implausibility metric also shows clear differences between the original and biased groups when using the Wachter and Greedy generators. However, this clear difference is not apparent for the REVISE generator, potentially due to its generally lower implausibility values compared to the others. This suggests that REVISE might be better at generating plausible counterfactuals, mitigating the negative impact on the biased subgroup.

To sum up, counterfactuals generated for the biased subgroup exhibit lower quality compared to the general data points. While the validity remains unchanged, both distance and implausibility metrics show degradation for the biased group. Furthermore, increasing the decision threshold has an impact on this effect specifically for the distance between factual and counterfactual.

8
Conclusion

This research investigated the impact of several aspects on the quality of counterfactual explanations. We explored the influence of inter-class distance, data imbalance, balancing techniques, and the presence of biased subgroups. We can summarise our key findings as follows:

- **Inter-class distance:** Our findings confirmed our initial hypothesis, demonstrating that increasing inter-class distance leads to a decrease in counterfactual quality. Specifically the plausibility of explanations was an issue when inter-class distance increased.
- **Data imbalance:** Contrary to our initial hypothesis, data imbalance did not significantly impact counterfactual quality. While some specific cases showed an effect, it was inconsistent and lacked a clear positive or negative direction. Further research is necessary to draw more definitive conclusions.
- **Balancing Techniques:** While the effect was minimal, we observed a general positive impact of balancing techniques on counterfactual plausibility. This aligns with our hypothesis and suggests that balancing can contribute to improved quality of counterfactuals for minority classes. Additionally, our findings support the notion that balancing techniques, known to improve classification performance, do not negatively impact counterfactual generation.
- **Biased Subgroups:** Our research supported the hypothesis that biased subgroups experience lower counterfactual quality compared to the overall data points. This highlights the need for further development in counterfactual generation methods to ensure fair treatment across all subgroups, especially those already negatively affected by the classifier.
- **Decision Threshold:** Our initial hypothesis that higher decision thresholds amplify the observed effects was not fully supported. While specific cases, such as biased subgroups, exhibited amplification, it was not a consistent trend across all experiments.

9 Discussion

This chapter delves into observations and limitations encountered during the research, along with potential avenues for future work that build upon the present study.

Beyond the scope of our primary research objectives and questions, several noteworthy findings emerged. Notably, the implausibility metrics exhibited minimal variation between the label-based and prediction-based approaches throughout the study. As discussed in Section 7.1, this could be attributed to the MLP model's exceptional performance on the MNIST dataset. While the NeuroTree and Random Forest models performed significantly worse on imbalanced datasets, the implausibility metrics remained virtually identical. Currently, the metric involves randomly sampling target class points and calculating their distance to the counterfactual. An intriguing alternative would be to sample target class points specifically near the counterfactual, since this proximity to the decision boundary is where the classification model is most likely to misclassify instances. Here, we expect a more pronounced difference between label-based and prediction-based implausibility. Another limitation of our work is that the size of the neighbourhood that we compare the counterfactual to, stays equal across all dataset sizes. This approach is inherently flawed, as smaller datasets need a smaller neighborhood to avoid averaging over the entire dataset's distance.

While analyzing credit-risk datasets, we observed a general decrease in standard deviations within the data imbalance and balancing technique experiments as the dataset size increased. This aligns with the established notion that larger datasets yield more consistent results. However, it is crucial to acknowledge that small datasets like German Credit and Credit Default might not be suitable for data imbalance experiments. Future investigations should prioritize datasets like Adult and GMSC.

Furthermore, limitations were encountered within the negatively biased subgroup experiments, particularly concerning the biased classifier. When trying to create a biased model we found that the Random Forest model exhibited stronger signs of bias than the NeuroTree model. However, due to time constraints, counterfactual generation for this model was not feasible. Exploring this avenue could potentially reveal even more pronounced discrepancies in counterfactual quality between the biased group and the remaining data.

Another limitation arises from the handling of categorical features, prevalent in credit-risk datasets. Our current approach utilises a standard integer encoding, which might not be optimal. This encoding could lead the classification model and counterfactual generator to misinterpret the feature values as rankings rather than distinct categorical values. Additionally, the counterfactual generation process disregards these categorical values, resulting in counterfactuals with values that do not correspond to specific integer encodings. Future research building upon this work should consider these limitations when evaluating counterfactual quality.

9.1. Future Work

We propose three distinct research directions that hold promise for advancing the field of counterfactual explanations:

- **Evaluating Counterfactual Quality through Distribution Analysis:** This approach involves analyzing the distributions of generated counterfactuals and comparing them via their mean or a distribution comparison technique like Maximum Mean Discrepancy (MMD).
- **Investigating Feature Mutability and Bias:** This research area delves into the relationship between feature mutability and counterfactual quality, particularly within the context of negatively biased subgroups.
- **Experimenting with Data Imbalance Ratios and Balancing Techniques:** This direction involves exploring the impact of varying data imbalance ratios and different parameter settings for balancing techniques on the quality of generated counterfactuals. Moreover, as discussed in both Section 7.2 and Section 7.3, changing the direction of the counterfactual explanation might also provide interesting insights.

9.1.1. Counterfactual Distribution Analysis

Our research primarily focused on measuring a counterfactual metric, averaging it across a large number of counterfactuals, and comparing it to another group. Instead of relying on such proxy metrics, an intriguing alternative lies in analyzing the actual distributions of the generated counterfactuals and their positioning within the feature space. This approach directly compares the counterfactuals themselves rather than relying on intermediary metrics. If discrepancies emerge in the counterfactual distributions, it suggests potential quality concerns as counterfactuals from different groups occupy distinct regions in the feature space. A simple comparison method involves analyzing the average counterfactual of each group. Another interesting approach might be to use the MMD metric which has been proposed in [29]. This metric has been proven to work well when comparing distributions and might therefore be interesting in comparing counterfactual distributions as well.

9.1.2. Feature Mutability

The relationship between feature mutability and counterfactual quality presents another interesting research area. For certain features like gender or race, the ability to mutate these values is crucial, as individuals cannot change them in reality. In the context of biased classifiers, this approach could reveal insights into the fairness of the counterfactual generation process. If significant quality discrepancies exist between generators capable of mutating these features and those that cannot, it suggests that these features are an important part of the counterfactual process, which is undesirable. This could further extend our negatively biased subgroup experiment. By disabling the mutation of features known to be associated with classifier bias, we expect to observe larger differences in counterfactual explanation quality, further highlighting potential fairness issues.

9.1.3. Data Imbalance and Balancing Techniques

A comprehensive future investigation could involve applying our research findings to datasets with varying levels of data imbalance and different balancing technique parameters. We anticipate that more severe class imbalances might lead to more pronounced differences in counterfactual quality compared to our observations. Additionally, optimizing balancing techniques for optimal classification model performance, followed by a comparison of the resulting counterfactual quality, could provide valuable insights into this topic. Regarding the direction of the counterfactual explanations, initial experiments show that generating counterfactuals from majority to minority classes leads to differences in quality compared to CEs for balanced datasets. Further research is necessary to find out whether these differences are impactful and whether these situations are applicable to real world scenarios.

References

[1] Patrick Altmeyer, Arie van Deursen, and Cynthia C. S. Liem. “Explaining Black-Box Models through Counterfactuals”. In: *JuliaCon Proceedings* 1.1 (Aug. 2023). arXiv:2308.07198 [cs], p. 130. ISSN: 2642-4029. DOI: `10.21105/jcon.00130`. URL: `http://arxiv.org/abs/2308.07198` (visited on 03/26/2024).

[2] Patrick Altmeyer et al. *Faithful Model Explanations through Energy-Constrained Conformal Counterfactuals*. arXiv:2312.10648 [cs]. Dec. 2023. DOI: `10.48550/arXiv.2312.10648`. URL: `http://arxiv.org/abs/2312.10648` (visited on 03/26/2024).

[3] David Alvarez Melis and Tommi Jaakkola. “Towards Robust Interpretability with Self-Explaining Neural Networks”. In: *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc., 2018. URL: `https://proceedings.neurips.cc/paper_files/paper/2018/hash/3e9f0fc9b2f89e043bc6233994dfcf76-Abstract.html` (visited on 05/20/2024).

[4] David Alvarez-Melis and Tommi S. Jaakkola. *On the Robustness of Interpretability Methods*. arXiv:1806.08049 [cs, stat]. June 2018. DOI: `10.48550/arXiv.1806.08049`. URL: `http://arxiv.org/abs/1806.08049` (visited on 05/13/2024).

[5] André Artelt et al. “Evaluating Robustness of Counterfactual Explanations”. In: *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*. Dec. 2021, pp. 01–09. DOI: `10.1109/SSCI50451.2021.9660058`. URL: `https://ieeexplore.ieee.org/abstract/document/9660058` (visited on 05/08/2024).

[6] Alejandro Baldominos, Yago Saez, and Pedro Isasi. “A Survey of Handwritten Character Recognition with MNIST and EMNIST”. en. In: *Applied Sciences* 9.15 (Jan. 2019). Number: 15 Publisher: Multidisciplinary Digital Publishing Institute, p. 3169. ISSN: 2076-3417. DOI: `10.3390/app9153169`. URL: `https://www.mdpi.com/2076-3417/9/15/3169` (visited on 04/02/2024).

[7] Alejandro Barredo Arrieta et al. “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI”. In: *Information Fusion* 58 (June 2020), pp. 82–115. ISSN: 1566-2535. DOI: `10.1016/j.inffus.2019.12.012`. URL: `https://www.sciencedirect.com/science/article/pii/S1566253519308103` (visited on 05/31/2024).

[8] Barry Becker and Ronny Kohavi. *Adult*. UCI Machine Learning Repository. DOI: `https://doi.org/10.24432/C5XW20`. 1996.

[9] Cigdem Beyan and Robert Fisher. “Classifying imbalanced data sets using similarity based hierarchical decomposition”. In: *Pattern Recognition* 48.5 (May 2015), pp. 1653–1672. ISSN: 0031-3203. DOI: `10.1016/j.patcog.2014.10.032`. URL: `https://www.sciencedirect.com/science/article/pii/S003132031400449X` (visited on 06/04/2024).

[10] Siddharth Bhatore, Lalit Mohan, and Y. Raghu Reddy. “Machine learning techniques for credit risk evaluation: a systematic literature review”. en. In: *Journal of Banking and Financial Technology* 4.1 (Apr. 2020), pp. 111–138. ISSN: 2524-7964. DOI: `10.1007/s42786-020-00020-3`. URL: `https://doi.org/10.1007/s42786-020-00020-3` (visited on 05/08/2024).

[11] Paula Branco, Luís Torgo, and Rita P. Ribeiro. “A Survey of Predictive Modeling on Imbalanced Domains”. en. In: *ACM Computing Surveys* 49.2 (June 2017), pp. 1–50. ISSN: 0360-0300, 1557-7341. DOI: `10.1145/2907070`. URL: `https://dl.acm.org/doi/10.1145/2907070` (visited on 06/04/2024).

[12] Iain Brown and Christophe Mues. “An experimental comparison of classification algorithms for imbalanced credit scoring data sets”. In: *Expert Systems with Applications* 39.3 (Feb. 2012), pp. 3446–3453. ISSN: 0957-4174. DOI: `10.1016/j.eswa.2011.09.033`. URL: `https://www.sciencedirect.com/science/article/pii/S095741741101342X` (visited on 05/01/2024).

[13] Nicola Capuano et al. “A Methodology based on Commonsense Knowledge and Ontologies for the Automatic Classification of Legal Cases”. In: *Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14)*. WIMS ’14. New York, NY, USA: Association for Computing Machinery, June 2014, pp. 1–6. ISBN: 978-1-4503-2538-7. DOI: 10.1145/2611040.2611048. URL: https://dl.acm.org/doi/10.1145/2611040.2611048 (visited on 01/19/2024).

[14] Rich Caruana et al. “Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission”. en. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Sydney NSW Australia: ACM, Aug. 2015, pp. 1721–1730. ISBN: 978-1-4503-3664-2. DOI: 10.1145/2783258.2788613. URL: https://dl.acm.org/doi/10.1145/2783258.2788613 (visited on 06/02/2024).

[15] N. V. Chawla et al. “SMOTE: Synthetic Minority Over-sampling Technique”. en. In: *Journal of Artificial Intelligence Research* 16 (June 2002), pp. 321–357. ISSN: 1076-9757. DOI: 10.1613/jair.953. URL: https://www.jair.org/index.php/jair/article/view/10302 (visited on 05/01/2024).

[16] Dan Claudiu Cireşan et al. “Deep Big Multilayer Perceptrons for Digit Recognition”. en. In: *Neural Networks: Tricks of the Trade: Second Edition*. Ed. by Grégoire Montavon, Geneviève B. Orr, and Klaus-Robert Müller. Berlin, Heidelberg: Springer, 2012, pp. 581–598. ISBN: 978-3-642-35289-8. DOI: 10.1007/978-3-642-35289-8_31. URL: https://doi.org/10.1007/978-3-642-35289-8_31 (visited on 04/02/2024).

[17] Will Cukierski Credit Fusion. *Give Me Some Credit*. 2011. URL: https://kaggle.com/competitions/GiveMeSomeCredit.

[18] Javier Del Ser et al. “On generating trustworthy counterfactual explanations”. In: *Information Sciences* 655 (Jan. 2024), p. 119898. ISSN: 0020-0255. DOI: 10.1016/j.ins.2023.119898. URL: https://www.sciencedirect.com/science/article/pii/S0020025523014834 (visited on 06/02/2024).

[19] Misha Denil and Thomas Trappenberg. “Overlap versus Imbalance”. en. In: *Advances in Artificial Intelligence*. Ed. by Atefeh Farzindar and Vlado Keŝelj. Berlin, Heidelberg: Springer, 2010, pp. 220–231. ISBN: 978-3-642-13059-5. DOI: 10.1007/978-3-642-13059-5_22.

[20] Jeremie Desgagne-Bouchard and Patrick Altmeyer. *Evovest/NeuroTreeModels.jl: v1.3.0*. Apr. 2024. DOI: 10.5281/zenodo.11011884. URL: https://doi.org/10.5281/zenodo.11011884.

[21] *Detection of colon cancer based on microarray dataset using machine learning as a feature selection and classification techniques | Request PDF*. URL: https://www.researchgate.net/publication/342304511_Detection_of_colon_cancer_based_on_microarray_dataset_using_machine_learning_as_a_feature_selection_and_classification_techniques (visited on 05/10/2024).

[22] Amit Dhurandhar et al. “Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives”. In: *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc., 2018. URL: https://proceedings.neurips.cc/paper_files/paper/2018/hash/c5ff2543b53f4cc0ad3819a36752467b-Abstract.html (visited on 03/26/2024).

[23] Ricardo Dominguez-Olmedo, Amir H. Karimi, and Bernhard Schölkopf. “On the Adversarial Robustness of Causal Algorithmic Recourse”. en. In: *Proceedings of the 39th International Conference on Machine Learning*. ISSN: 2640-3498. PMLR, June 2022, pp. 5324–5342. URL: https://proceedings.mlr.press/v162/dominguez-olmedo22a.html (visited on 05/24/2024).

[24] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. “Explainable artificial intelligence: A survey”. In: *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. May 2018, pp. 0210–0215. DOI: 10.23919/MIPRO.2018.8400040. URL: https://ieeexplore.ieee.org/abstract/document/8400040 (visited on 06/02/2024).

[25] Maria Fox, Derek Long, and Daniele Magazzeni. *Explainable Planning*. en. arXiv:1709.10256 [cs]. Sept. 2017. URL: http://arxiv.org/abs/1709.10256 (visited on 06/02/2024).

[26] Archana Gahlaut, Tushar, and Prince Kumar Singh. “Prediction analysis of risky credit using Data mining classification models”. In: *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. July 2017, pp. 1–7. DOI: 10.1109/ICCCNT.2017.8203982. URL: https://ieeexplore.ieee.org/abstract/document/8203982 (visited on 04/02/2024).

[27] Bryce Goodman and Seth Flaxman. “European Union Regulations on Algorithmic Decision Making and a ”Right to Explanation””. English. In: *AI Magazine* 38.3 (2017). Num Pages: 50-57 Place: La Canada, United States Publisher: Association for the Advancement of Artificial Intelligence Section: Articles, pp. 50–57. ISSN: 07384602. URL: https://www.proquest.com/docview/1967052651/abstract/62FD7D06EDBE4F2BPQ/1 (visited on 01/19/2024).

[28] Rory Mc Grath et al. *Interpretable Credit Application Predictions With Counterfactual Explanations*. arXiv:1811.05245 [cs]. Nov. 2018. DOI: 10.48550/arXiv.1811.05245. URL: http://arxiv.org/abs/1811.05245 (visited on 03/26/2024).

[29] Arthur Gretton et al. “A kernel two-sample test”. In: *The Journal of Machine Learning Research* 13.1 (2012), pp. 723–773.

[30] Riccardo Guidotti. “Counterfactual explanations and how to find them: literature review and benchmarking”. en. In: *Data Mining and Knowledge Discovery* (Apr. 2022). ISSN: 1573-756X. DOI: 10.1007/s10618-022-00831-6. URL: https://doi.org/10.1007/s10618-022-00831-6 (visited on 03/26/2024).

[31] Riccardo Guidotti et al. “A Survey of Methods for Explaining Black Box Models”. In: *ACM Computing Surveys* 51.5 (Aug. 2018), 93:1–93:42. ISSN: 0360-0300. DOI: 10.1145/3236009. URL: https://dl.acm.org/doi/10.1145/3236009 (visited on 05/29/2024).

[32] Vivek Gupta et al. *Equalizing Recourse across Groups*. arXiv:1909.03166 [cs, stat]. Sept. 2019. DOI: 10.48550/arXiv.1909.03166. URL: http://arxiv.org/abs/1909.03166 (visited on 01/19/2024).

[33] Taehyun Ha, Sangwon Lee, and Sangyeon Kim. “Designing Explainability of an Artificial Intelligence System”. en. In: *Proceedings of the Technology, Mind, and Society*. Washington DC USA: ACM, Apr. 2018, pp. 1–1. ISBN: 978-1-4503-5420-2. DOI: 10.1145/3183654.3183683. URL: https://dl.acm.org/doi/10.1145/3183654.3183683 (visited on 06/02/2024).

[34] Haibo He et al. “ADASYN: Adaptive synthetic sampling approach for imbalanced learning”. In: *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. Hong Kong, China: IEEE, June 2008, pp. 1322–1328. ISBN: 978-1-4244-1820-6. DOI: 10.1109/IJCNN.2008.4633969. URL: http://ieeexplore.ieee.org/document/4633969/ (visited on 06/04/2024).

[35] Guo Haixiang et al. “Learning from class-imbalanced data: Review of methods and applications”. In: *Expert Systems with Applications* 73 (May 2017), pp. 220–239. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2016.12.035. URL: https://www.sciencedirect.com/science/article/pii/S0957417416307175 (visited on 05/01/2024).

[36] Aparajita Haldar, Teddy Cunningham, and Hakan Ferhatosmanoglu. “RAGUEL: Recourse-Aware Group Unfairness Elimination”. In: *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. CIKM ’22. New York, NY, USA: Association for Computing Machinery, Oct. 2022, pp. 666–675. ISBN: 978-1-4503-9236-5. DOI: 10.1145/3511808.3557424. URL: https://dl.acm.org/doi/10.1145/3511808.3557424 (visited on 03/26/2024).

[37] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. “Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning”. en. In: *Advances in Intelligent Computing*. Ed. by De-Shuang Huang, Xiao-Ping Zhang, and Guang-Bin Huang. Berlin, Heidelberg: Springer, 2005, pp. 878–887. ISBN: 978-3-540-31902-3. DOI: 10.1007/11538059_91.

[38] Leif Hancox-Li. “Robustness in machine learning explanations: does it matter?” In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* ’20. New York, NY, USA: Association for Computing Machinery, Jan. 2020, pp. 640–647. ISBN: 978-1-4503-6936-7. DOI: 10.1145/3351095.3372836. URL: https://dl.acm.org/doi/10.1145/3351095.3372836 (visited on 05/13/2024).

[39] Paul R. Harper. “A review and comparison of classification algorithms for medical decision making”. In: *Health Policy* 71.3 (Mar. 2005), pp. 315–331. ISSN: 0168-8510. DOI: 10.1016/j.healthpol.2004.05.002. URL: https://www.sciencedirect.com/science/article/pii/S016885100400096X (visited on 01/19/2024).

[40] Hassan Hassan et al. “ASSESSMENT OF ARTIFICIAL NEURAL NETWORK FOR BATHYMETRY ESTIMATION USING HIGH RESOLUTION SATELLITE IMAGERY IN SHALLOW LAKES: CASE STUDY EL BURULLUS LAKE.” In: *International Water Technology Journal* 5 (Dec. 2015).

[41] Hossein Hassani et al. “A review of data mining applications in crime”. en. In: *Statistical Analysis and Data Mining: The ASA Data Science Journal* 9.3 (2016), pp. 139–154. ISSN: 1932-1872. DOI: 10.1002/sam.11312. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sam.11312 (visited on 05/08/2024).

[42] *High-level summary of the AI Act | EU Artificial Intelligence Act*. en-US. URL: https://artificialintelligenceact.eu/high-level-summary/ (visited on 05/29/2024).

[43] *HMDA - Home Mortgage Disclosure Act*. URL: https://ffiec.cfpb.gov/ (visited on 04/28/2024).

[44] Hans Hofmann. *Statlog (German Credit Data)*. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5NC77. 1994.

[45] Shalmali Joshi et al. *Towards Realistic Individual Recourse and Actionable Explanations in Black-Box Decision Making Systems*. arXiv:1907.09615 [cs, stat]. July 2019. DOI: 10.48550/arXiv.1907.09615. URL: http://arxiv.org/abs/1907.09615 (visited on 01/19/2024).

[46] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. “Algorithmic Recourse: from Counterfactual Explanations to Interventions”. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’21. New York, NY, USA: Association for Computing Machinery, Mar. 2021, pp. 353–362. ISBN: 978-1-4503-8309-7. DOI: 10.1145/3442188.3445899. URL: https://dl.acm.org/doi/10.1145/3442188.3445899 (visited on 01/19/2024).

[47] Amir-Hossein Karimi et al. *A survey of algorithmic recourse: definitions, formulations, solutions, and prospects*. arXiv:2010.04050 [cs, stat]. Mar. 2021. DOI: 10.48550/arXiv.2010.04050. URL: http://arxiv.org/abs/2010.04050 (visited on 04/19/2024).

[48] Todd Kulesza et al. “Principles of Explanatory Debugging to Personalize Interactive Machine Learning”. en. In: *Proceedings of the 20th International Conference on Intelligent User Interfaces*. Atlanta Georgia USA: ACM, Mar. 2015, pp. 126–137. ISBN: 978-1-4503-3306-1. DOI: 10.1145/2678025.2701399. URL: https://dl.acm.org/doi/10.1145/2678025.2701399 (visited on 06/02/2024).

[49] Thibault Laugel et al. *Issues with post-hoc counterfactual explanations: a discussion*. arXiv:1906.04774 [cs, stat]. June 2019. DOI: 10.48550/arXiv.1906.04774. URL: http://arxiv.org/abs/1906.04774 (visited on 05/13/2024).

[50] Y. Lecun et al. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11 (Nov. 1998). Conference Name: Proceedings of the IEEE, pp. 2278–2324. ISSN: 1558-2256. DOI: 10.1109/5.726791. URL: https://ieeexplore.ieee.org/abstract/document/726791 (visited on 01/19/2024).

[51] Yuanyuan Li et al. “Counterfactual learning in customer churn prediction under class imbalance”. In: *Proceedings of the 2023 6th International Conference on Big Data Technologies*. ICBDT ’23. New York, NY, USA: Association for Computing Machinery, Dec. 2023, pp. 96–102. DOI: 10.1145/3627377.3627392. URL: https://dl.acm.org/doi/10.1145/3627377.3627392 (visited on 04/02/2024).

[52] Zachary C. Lipton. “The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery.” en. In: *Queue* 16.3 (June 2018), pp. 31–57. ISSN: 1542-7730, 1542-7749. DOI: 10.1145/3236386.3241340. URL: https://dl.acm.org/doi/10.1145/3236386.3241340 (visited on 06/02/2024).

[53] Victoria López et al. “An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics”. In: *Information Sciences* 250 (Nov. 2013), pp. 113–141. ISSN: 0020-0255. DOI: 10.1016/j.ins.2013.07.007. URL: https://www.sciencedirect.com/science/article/pii/S0020025513005124 (visited on 06/04/2024).

[54] Octavio Loyola-González et al. “Study of the impact of resampling methods for contrast pattern based classifiers in imbalanced databases”. In: *Neurocomputing* 175 (Jan. 2016), pp. 935–947. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2015.04.120. URL: https://www.sciencedirect.com/science/article/pii/S0925231215015908 (visited on 06/04/2024).

[55] Scott M Lundberg and Su-In Lee. “A Unified Approach to Interpreting Model Predictions”. In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html (visited on 05/01/2024).

[56] Tambiama Madiega. “Artificial intelligence act”. In: *European Parliament: European Parliamentary Research Service* (2021).

[57] Divyat Mahajan, Chenhao Tan, and Amit Sharma. *Preserving Causal Constraints in Counterfactual Explanations for Machine Learning Classifiers*. arXiv:1912.03277 [cs, stat]. June 2020. DOI: 10.48550/arXiv.1912.03277. URL: http://arxiv.org/abs/1912.03277 (visited on 01/19/2024).

[58] Saumitra Mishra et al. *A Survey on the Robustness of Feature Importance and Counterfactual Explanations*. arXiv:2111.00358 [cs]. Jan. 2023. DOI: 10.48550/arXiv.2111.00358. URL: http://arxiv.org/abs/2111.00358 (visited on 05/20/2024).

[59] Christoph Molnar. *Interpretable Machine Learning*. en. Google-Books-ID: jBm3DwAAQBAJ. Lulu.com, 2020. ISBN: 978-0-244-76852-2.

[60] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. “Explaining machine learning classifiers through diverse counterfactual explanations”. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* ’20. New York, NY, USA: Association for Computing Machinery, Jan. 2020, pp. 607–617. ISBN: 978-1-4503-6936-7. DOI: 10.1145/3351095.3372850. URL: https://dl.acm.org/doi/10.1145/3351095.3372850 (visited on 01/19/2024).

[61] Krystyna Napierala and Jerzy Stefanowski. “Types of minority class examples and their influence on learning classifiers from imbalanced data”. en. In: *Journal of Intelligent Information Systems* 46.3 (June 2016), pp. 563–597. ISSN: 1573-7675. DOI: 10.1007/s10844-015-0368-1. URL: https://doi.org/10.1007/s10844-015-0368-1 (visited on 06/04/2024).

[62] *NeuroTree - A differentiable tree operator for tabular data | NeuroTreeModels*. URL: https://evovest.github.io/NeuroTreeModels.jl/dev/design (visited on 05/13/2024).

[63] Nicolas Papernot et al. “The Limitations of Deep Learning in Adversarial Settings”. In: *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*. Mar. 2016, pp. 372–387. DOI: 10.1109/EuroSP.2016.36. URL: https://ieeexplore.ieee.org/abstract/document/7467366 (visited on 03/21/2024).

[64] Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. “On Counterfactual Explanations under Predictive Multiplicity”. en. In: *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*. ISSN: 2640-3498. PMLR, Aug. 2020, pp. 809–818. URL: https://proceedings.mlr.press/v124/pawelczyk20a.html (visited on 06/02/2024).

[65] Martin Pawelczyk et al. “Probabilistically Robust Recourse: Navigating the Trade-offs between Costs and Robustness in Algorithmic Recourse”. en. In: Sept. 2022. URL: https://openreview.net/forum?id=sC-PmTsiTB (visited on 05/24/2024).

[66] Judea Pearl. *Causality*. en. Google-Books-ID: f4nuexsNVZIC. Cambridge University Press, Sept. 2009. ISBN: 978-0-521-89560-6.

[67] Dabal Pedamonti. *Comparison of non-linear activation functions for deep neural networks on MNIST classification task*. arXiv:1804.02763 [cs, stat]. Apr. 2018. DOI: 10.48550/arXiv.1804.02763. URL: http://arxiv.org/abs/1804.02763 (visited on 04/02/2024).

[68] Rafael Poyiadzi et al. “FACE: Feasible and Actionable Counterfactual Explanations”. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. AIES ’20. New York, NY, USA: Association for Computing Machinery, Feb. 2020, pp. 344–350. ISBN: 978-1-4503-7110-0. DOI: 10.1145/3375627.3375850. URL: https://dl.acm.org/doi/10.1145/3375627.3375850 (visited on 04/02/2024).

[69] Ronaldo C. Prati, Gustavo E. A. P. A. Batista, and Diego F. Silva. “Class imbalance revisited: a new experimental setup to assess the performance of treatment methods”. en. In: *Knowledge and Information Systems* 45.1 (Oct. 2015), pp. 247–270. ISSN: 0219-3116. DOI: 10.1007/s10115-014-0794-3. URL: https://doi.org/10.1007/s10115-014-0794-3 (visited on 05/30/2024).

[70] Pramila Rani et al. “An empirical study of machine learning techniques for affect recognition in human–robot interaction”. en. In: *Pattern Analysis and Applications* 9.1 (May 2006), pp. 58–69. ISSN: 1433-755X. DOI: 10.1007/s10044-006-0025-y. URL: https://doi.org/10.1007/s10044-006-0025-y (visited on 06/02/2024).

[71] Protection Regulation. “Regulation (EU) 2016/679 of the European Parliament and of the Council”. In: *Regulation (eu)* 679 (2016), p. 2016.

[72] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “”Why Should I Trust You?”: Explaining the Predictions of Any Classifier”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’16. New York, NY, USA: Association for Computing Machinery, Aug. 2016, pp. 1135–1144. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939778. URL: https://dl.acm.org/doi/10.1145/2939672.2939778 (visited on 01/19/2024).

[73] Cynthia Rudin. “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead”. en. In: *Nature Machine Intelligence* 1.5 (May 2019). Publisher: Nature Publishing Group, pp. 206–215. ISSN: 2522-5839. DOI: 10.1038/s42256-019-0048-x. URL: https://www.nature.com/articles/s42256-019-0048-x (visited on 05/29/2024).

[74] Lisa Schut et al. “Generating Interpretable Counterfactual Explanations By Implicit Minimisation of Epistemic and Aleatoric Uncertainties”. en. In: *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*. ISSN: 2640-3498. PMLR, Mar. 2021, pp. 1756–1764. URL: https://proceedings.mlr.press/v130/schut21a.html (visited on 03/20/2024).

[75] Shubham Sharma, Jette Henderson, and Joydeep Ghosh. “CERTIFAI: Counterfactual Explanations for Robustness, Transparency, Interpretability, and Fairness of Artificial Intelligence models”. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. arXiv:1905.07857 [cs, stat]. Feb. 2020, pp. 166–172. DOI: 10.1145/3375627.3375812. URL: http://arxiv.org/abs/1905.07857 (visited on 01/19/2024).

[76] Naeem Siddiqi. *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*. en. John Wiley & Sons, June 2012. ISBN: 978-1-118-42916-7.

[77] Dylan Slack et al. “Counterfactual Explanations Can Be Manipulated”. In: *Advances in Neural Information Processing Systems*. Vol. 34. Curran Associates, Inc., 2021, pp. 62–75. URL: https://proceedings.neurips.cc/paper/2021/hash/009c434cab57de48a31f6b669e7ba266-Abstract.html (visited on 05/20/2024).

[78] Muhammad Atif Tahir et al. “A Multiple Expert Approach to the Class Imbalance Problem Using Inverse Random under Sampling”. en. In: *Multiple Classifier Systems*. Ed. by Jón Atli Benediktsson, Josef Kittler, and Fabio Roli. Berlin, Heidelberg: Springer, 2009, pp. 82–91. ISBN: 978-3-642-02326-2. DOI: 10.1007/978-3-642-02326-2_9.

[79] Gabriele Tolomei et al. “Interpretable Predictions of Tree-based Ensembles via Actionable Feature Tweaking”. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’17. New York, NY, USA: Association for Computing Machinery, Aug. 2017, pp. 465–474. ISBN: 978-1-4503-4887-4. DOI: 10.1145/3097983.3098039. URL: https://dl.acm.org/doi/10.1145/3097983.3098039 (visited on 03/20/2024).

[80] Divya Tomar and Sonali Agarwal. “A survey on Data Mining approaches for Healthcare”. en. In: *International Journal of Bio-Science and Bio-Technology* 5.5 (Oct. 2013), pp. 241–266. ISSN: 22337849, 22337849. DOI: 10.14257/ijbsbt.2013.5.5.25. URL: http://gvpress.com/journals/IJBSBT/vol5_no5/25.pdf (visited on 05/08/2024).

[81] Berk Ustun, Alexander Spangher, and Yang Liu. “Actionable Recourse in Linear Classification”. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* ’19. New York, NY, USA: Association for Computing Machinery, Jan. 2019, pp. 10–19. ISBN: 978-1-4503-6125-5. DOI: 10.1145/3287560.3287566. URL: https://dl.acm.org/doi/10.1145/3287560.3287566 (visited on 01/19/2024).

[82] Laurens Van der Maaten and Geoffrey Hinton. “Visualizing data using t-SNE.” In: *Journal of machine learning research* 9.11 (2008). ISSN: 1532-4435.

[83] Arnaud Van Looveren and Janis Klaise. “Interpretable Counterfactual Explanations Guided by Prototypes”. en. In: *Machine Learning and Knowledge Discovery in Databases. Research Track*. Ed. by Nuria Oliver et al. Cham: Springer International Publishing, 2021, pp. 650–665. ISBN: 978-3-030-86520-7. DOI: 10.1007/978-3-030-86520-7_40.

[84] Sahil Verma et al. *Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review*. arXiv:2010.10596 [cs, stat]. Nov. 2022. URL: http://arxiv.org/abs/2010.10596 (visited on 04/19/2024).

[85] Giulia Vilone and Luca Longo. “Notions of explainability and evaluation approaches for explainable artificial intelligence”. In: *Information Fusion* 76 (Dec. 2021), pp. 89–106. ISSN: 1566-2535. DOI: 10.1016/j.inffus.2021.05.009. URL: https://www.sciencedirect.com/science/article/pii/S1566253521001093 (visited on 05/31/2024).

[86] Sandra Wachter, Brent Mittelstadt, and Chris Russell. “Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR”. en. In: *SSRN Electronic Journal* (2017). ISSN: 1556-5068. DOI: 10.2139/ssrn.3063289. URL: https://www.ssrn.com/abstract=3063289 (visited on 01/19/2024).

[87] Hui-Xin Wang et al. “Smoking and the Occurence of Alzheimer’s Disease: Cross-Sectional and Longitudinal Data in a Population-based Study”. In: *American Journal of Epidemiology* 149.7 (Apr. 1999), pp. 640–644. ISSN: 0002-9262. DOI: 10.1093/oxfordjournals.aje.a009864. URL: https://doi.org/10.1093/oxfordjournals.aje.a009864 (visited on 06/02/2024).

[88] Feiyu Xu et al. “Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges”. en. In: *Natural Language Processing and Chinese Computing*. Ed. by Jie Tang et al. Cham: Springer International Publishing, 2019, pp. 563–574. ISBN: 978-3-030-32236-6. DOI: 10.1007/978-3-030-32236-6_51.

[89] I-Cheng Yeh. *Default of Credit Card Clients*. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C55S3H. 2016.

[90] Ligang Zhou. “Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods”. In: *Knowledge-Based Systems* 41 (Mar. 2013), pp. 16–25. ISSN: 0950-7051. DOI: 10.1016/j.knosys.2012.12.007. URL: https://www.sciencedirect.com/science/article/pii/S095070511200353X (visited on 06/04/2024).

A

Inter-class Distance Visuals

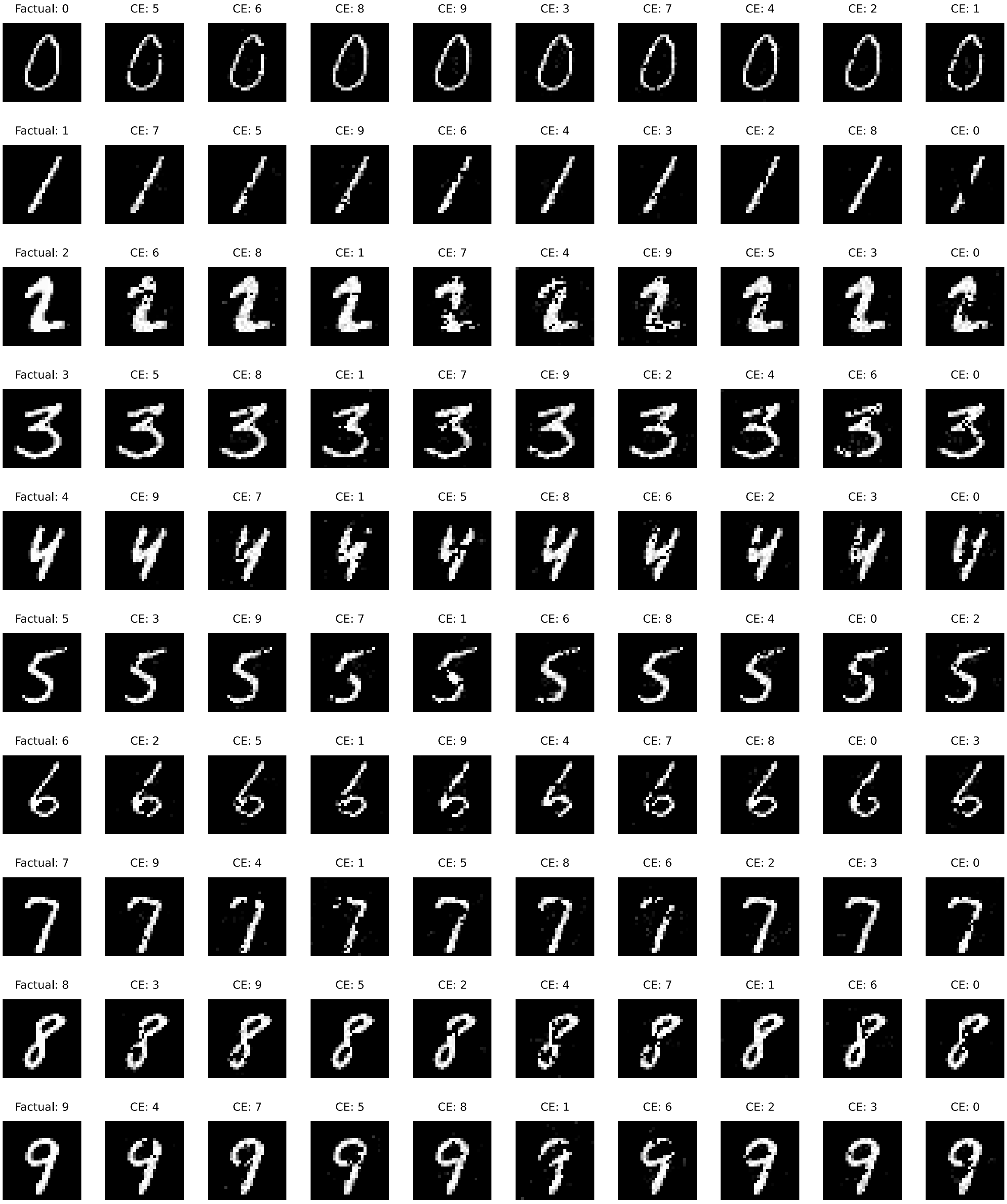


Figure A.1: Counterfactuals generated with the Greedy-0.5 generator. For each factual, the counterfactuals are ordered by inter-class distance.

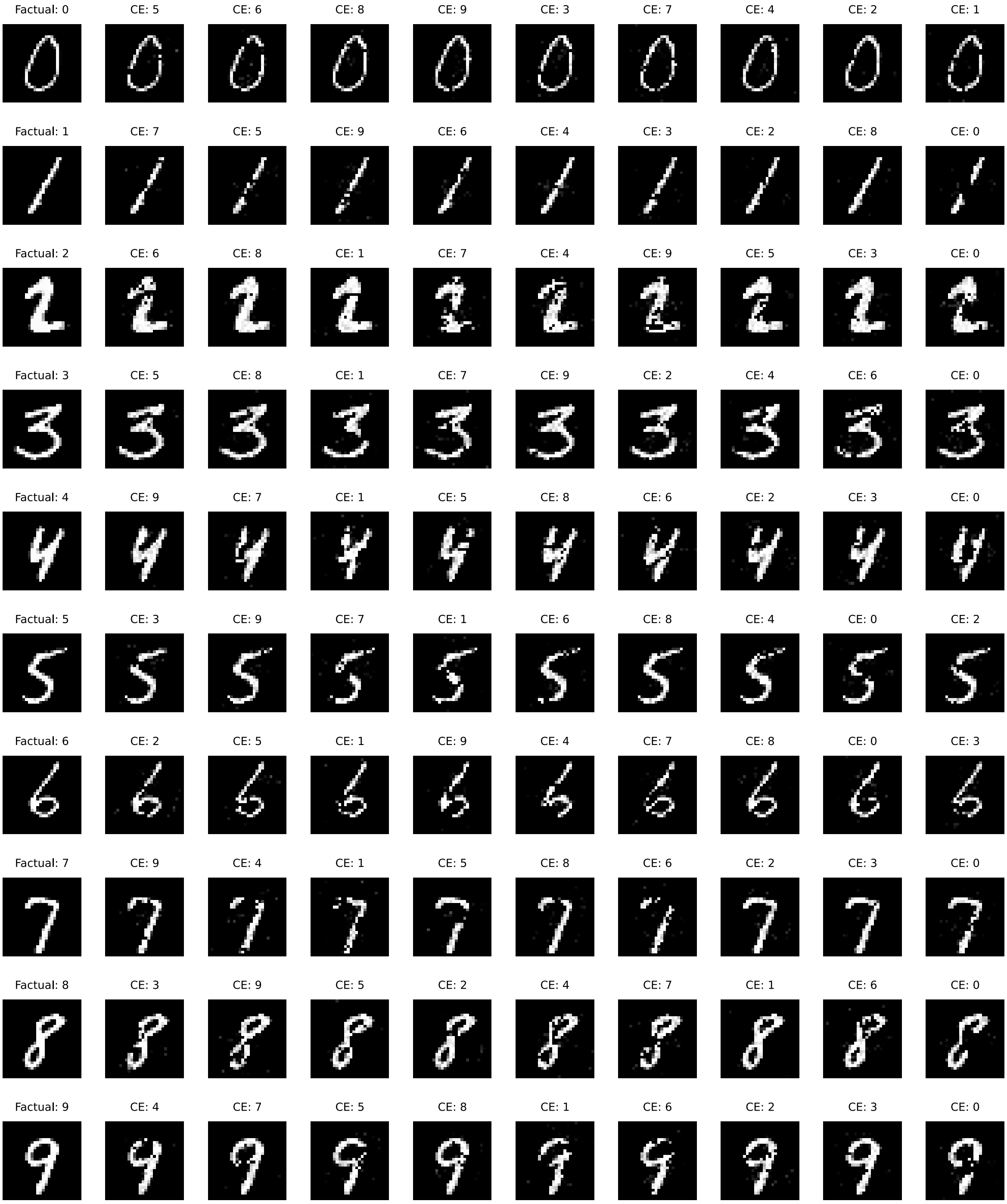


Figure A.2: Counterfactuals generated with the Greedy-0.95 generator. For each factual, the counterfactuals are ordered by inter-class distance.

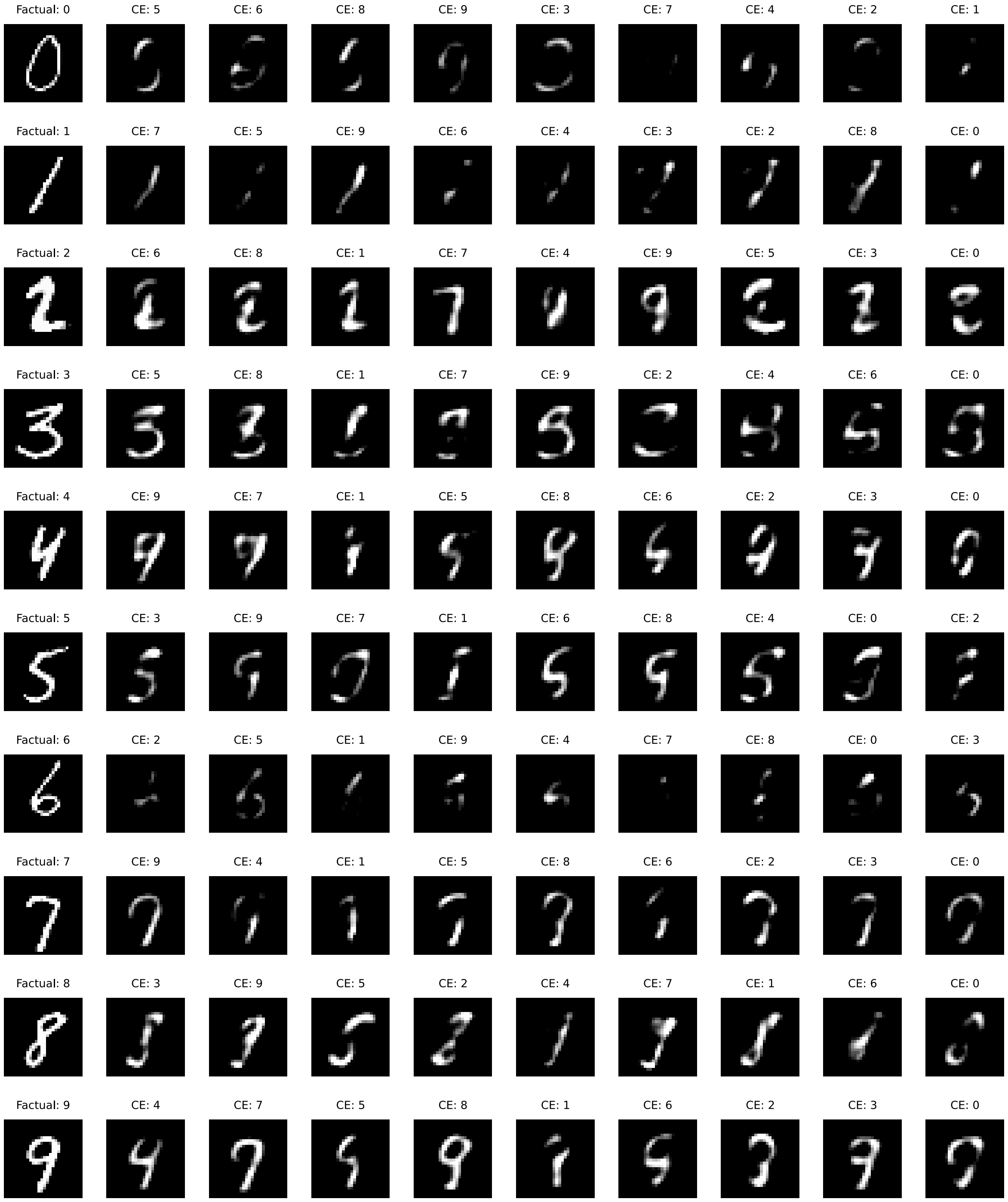


Figure A.3: Counterfactuals generated with the REVISE-0.5 generator. For each factual, the counterfactuals are ordered by inter-class distance.

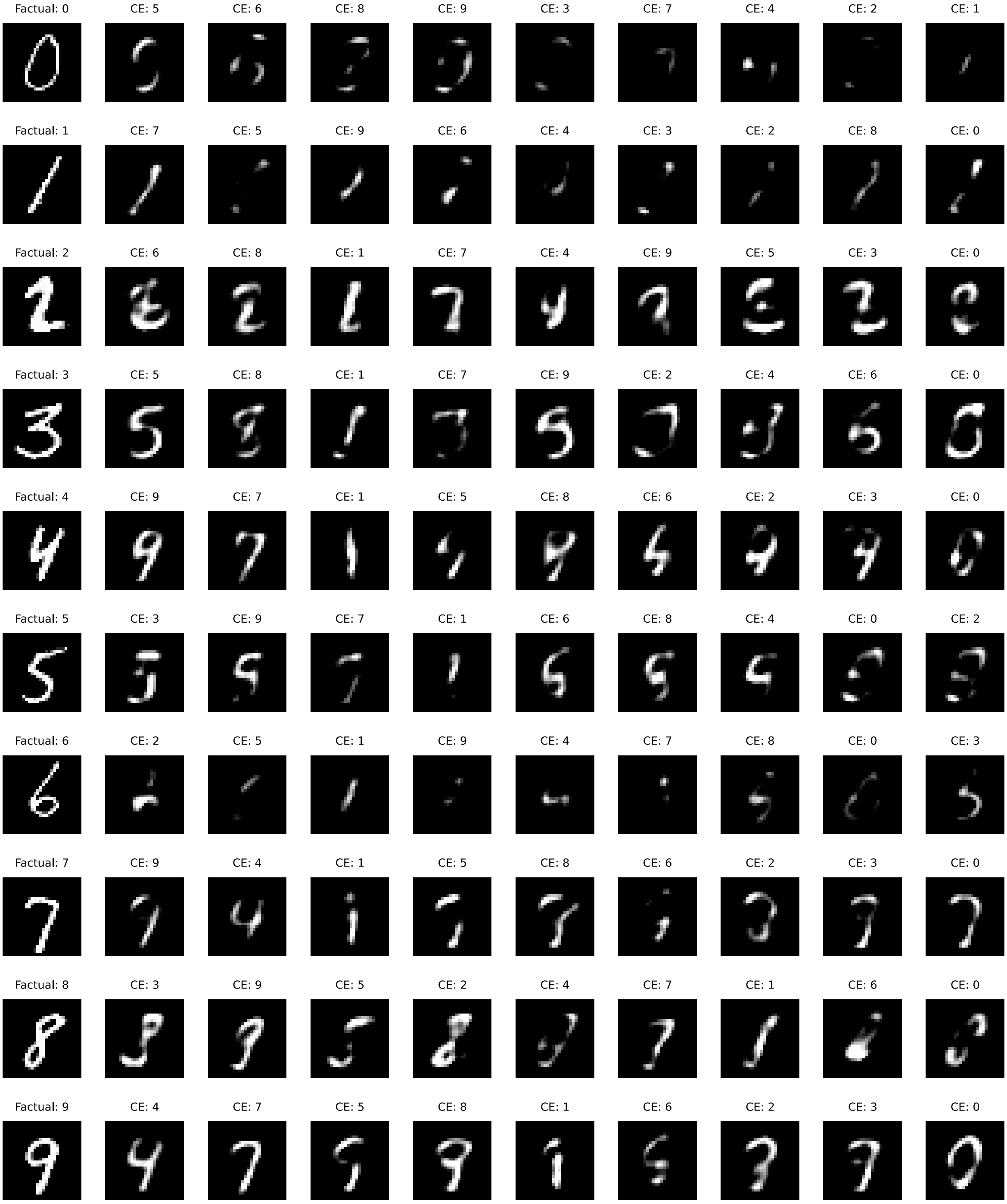


Figure A.4: Counterfactuals generated with the REVISE-0.95 generator. For each factual, the counterfactuals are ordered by inter-class distance.

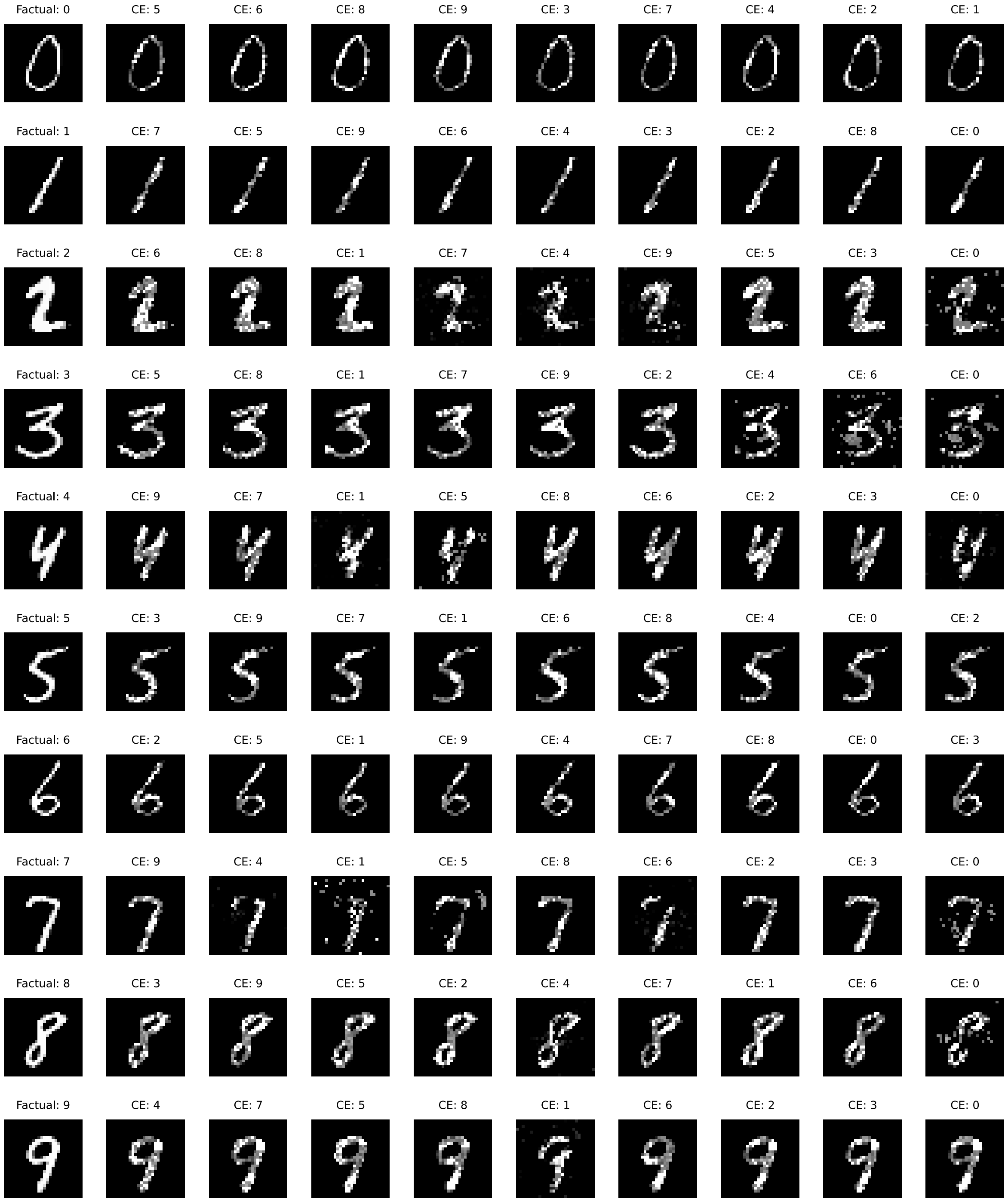


Figure A.5: Counterfactuals generated with the Wachter-0.5 generator. For each factual, the counterfactuals are ordered by inter-class distance.

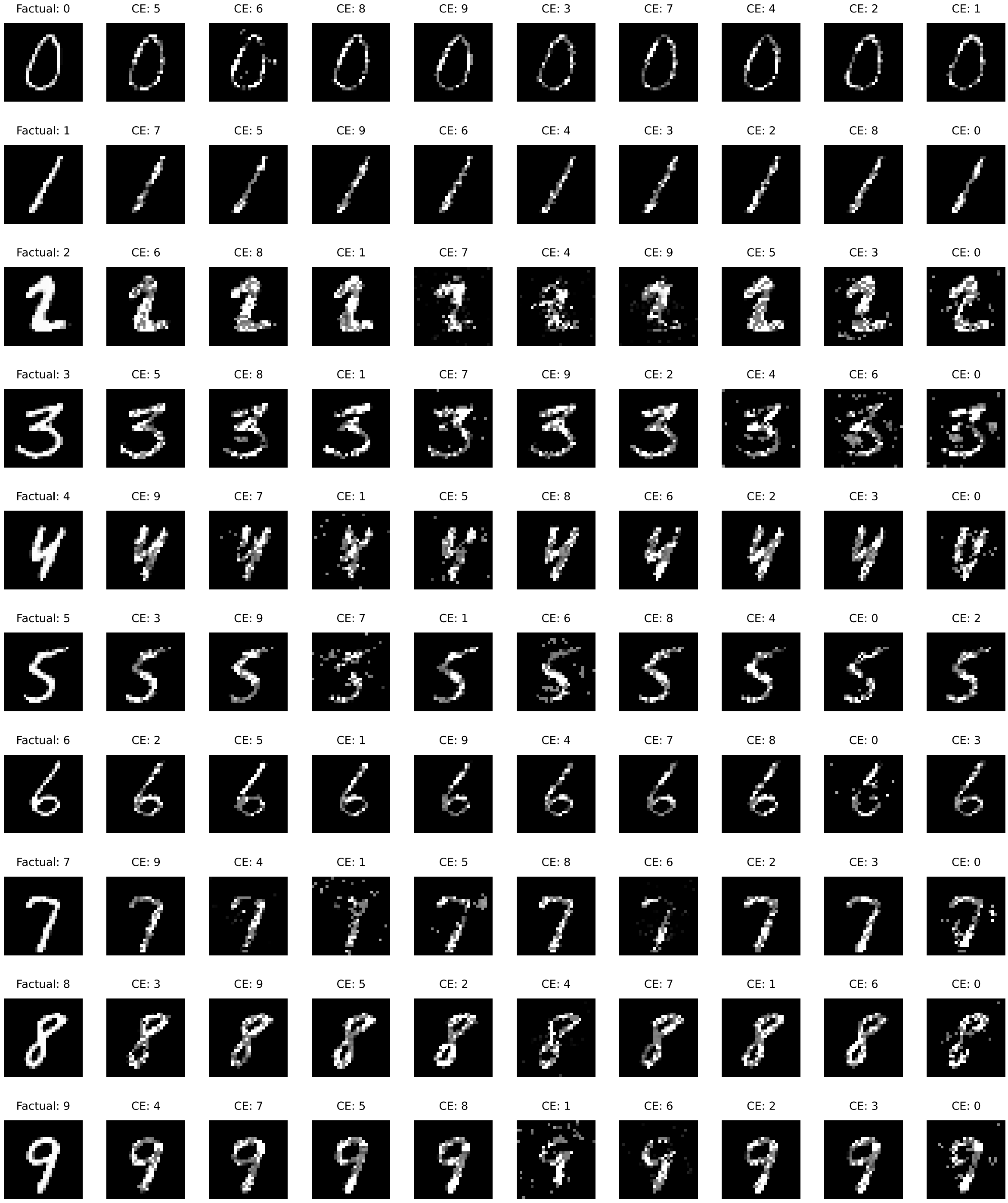


Figure A.6: Counterfactuals generated with the Wachter-0.95 generator. For each factual, the counterfactuals are ordered by inter-class distance.