Master Thesis

Erasmus School of Economics

MSc Data Science and Marketing Analytics

Towards Fair ML: A Monitoring Approach

Monitoring discriminatory bias in ML models using Statistical Process Control and Sequential Analysis procedures

D. M. (Douwe) Kruyt

498342

Supervisor: prof. dr. Dennis Fok

Second Assessor: prof. dr. Martijn de Jong

Company Supervisor: Fabian de Bont

Date final version: 02-08-2023

Abstract

Due to the widespread application and dependence on Machine Learning (ML) models it is imperative to ensure that models provide fair outcomes towards various segments of society. This study aims to investigate how deployed ML models, with continuous streams of new data, can be monitored for discriminatory bias to guarantee the quick discovery and eventual correction of the bias, thereby reducing the potential for discrimination. By leveraging monitoring techniques from the fields of statistical process control and sequential analysis, this research presents a novel approach being the first study to create a comprehensive monitoring system specifically for discriminatory bias. To achieve this, the following methods are employed: the Shewhart chart, cumulative sum control chart (CUSUM), exponentially weighted moving average control chart (EWMA), sequential probability ratio t-test (SPRT-t) and sequential Bayesian factor testing (SBF). The effectiveness of the methods is tested by measuring their performance over a variety of thirteen differently biased settings, using three different measurement metrics and a varying measurement window for one of them. The biased settings are created through synthetic generation of datasets, in order to guarantee controlled and reproducible conditions for evaluating the methods' performance across a diverse range of scenarios. The findings suggest that three methods – the CUSUM, SPRT-t and SBF – are particularly effective. Additionally, this research advocates for the integration of various methods into a comprehensive monitoring system, enhancing its robustness and nuance by leveraging their combined strengths rather than relying on a single approach.

Keywords: *Fairness, Machine Learning, Discrimination, Monitoring, Process Control*

Table of Contents

List of Abbreviations

Abbreviation	Meaning	Page
ML	Machine Learning	1
SPC	Statistical Process Control	16
SA	Sequential Analysis	17
NHST	Null Hypothesis Statistical Testing	17
CUSUM	Cumulative Sum (control chart)	16
EWMA	Exponentially Weighted Moving Average (control chart)	16
SBF	Sequential Bayesian Factor	18
SPRT	Sequential Probability Ratio Test	18
SPRT-t	Sequential Probability Ratio Test t-test	27
GS	Group Sequential (testing)	17
IC	In Control	17
OC	Out Control	17
FPR	False Positive Rate	31
FP	False Positive	7
TP	True Positive	7
FN	False Negative	7
TN	True Negative	7
LCL	Lower Control Limit	23
UCL	Upper Control limit	23

1 Introduction

1.1 Discrimination within Machine Learning models

With the digitization of the modern world and the increasing amount of data available, human kind is becoming increasingly dependent on machines. Research conducted by the International Data Corporation illustrates that the amount of stored data will increase from 118 zettabytes in 2023 to 149 zettabytes in 2024 (Rydning, 2022). This increase in data creation and consumption is expected to only increase further in the future. Industry leaders are applying Machine Learning (ML) models throughout all facets of society, from marketing to national defense operations, in order to process and optimally gain insights from these large quantities of data (Shrestha & Das, 2022). Despite the promising potential of these models to offer efficient solutions to complex issues, they can often exhibit biases and unjust treatment towards certain groups or individuals.

Fairness within ML algorithms is a topic receiving increased amounts of attention over the past couple of years (Shrestha & Das, 2022). Due to their extensive use even marginal biases can have large-scale impacts. Considering the complexity of ML models and the challenges associated with untangling the mechanics behind their decision-making systems, ensuring the fairness of their outputs becomes an intricate task. Because of the deep rooted nature of discrimination within human society, these historical biases can trickle into the automated systems through the data. ML models trained on this data often pick up on these biases within the data and can amplify them in their internal computations. Depending on where and how these models are being used, they can inadvertently magnify and reinforce the systemic biases within society.

There are ample situations where deployed ML models were discovered to produce discriminating outputs. Many of these situations arise in settings that are directly tied to marketing efforts or areas that have great implications for marketing strategies. Take for example a hiring tool developed by Amazon, which discriminated against women for technical roles. Or facial recognition software, where of the 189 facial recognition developers evaluated, more than 50% of the algorithms employed were showing signs of discriminatory bias. Also in credit decisions AI systems have been used and found to discriminate between applicants. These examples highlight a crucial point for marketers: while automated systems and ML have the ability to improve processes and offer insightful data, they also run the risk of amplifying already-existing social inequities (Wadsworth, Vera, & Piech, 2018). The ethical use of these tools is essential for marketers to really understand and serve markets in all of their diversity and to prevent reputational harm and legal consequences.

1.2 Defining the problem

Due to the widespread application and dependence on ML models it is imperative to ensure that models provide fair outcomes towards various segments of society. Disappointingly, very little ML models have (internal or external) safeguards in place to ensure fairness towards various groups, nor do most developers factor in such considerations when building and deploying the models (Loureiro, et al., 2023).

With the increasing attention of fairness in ML models there has been a surge in research done on the topic over the past couple of years (Shrestha & Das, 2022). Despite the significant contributions of these scientific publications and research in revealing insights into minority biases in ML and Artificial Intelligence systems, little research has been conducted on the identification and monitoring of discriminatory bias in deployed ML models. The vast majority of research on ML discrimination focusses on mitigation of bias over stationary datasets. However, real life ML scenarios include dynamic data and continuously running models.

This outlines a vital problem: there are seemingly no systems to monitor discriminatory bias in deployed ML models in real-time, flagging biased results as they occur. Considering the 'black box' nature of many ML models and the common practice of marketers relying on third-party consultants to develop these models, it becomes crucial to not only acknowledge the necessity of fairness in ML models but also prioritize the active monitoring of these models once they are deployed. This guarantees that discriminatory outputs from ML models can be quickly discovered and corrected, in turn reducing the potential for discrimination. While rectifying these biases is a complex challenge in itself, the first step in tackling this problem lies in reliably detecting them. In response to this problem, this research aims to answer the following research question:

> *How can discriminatory bias be monitored effectively in deployed Machine Learning models to prevent unfair predictions from being used, considering the continuous streaming of new data?*

This study contributes to existing literature by proposing a system that can be used to monitor deployed ML models for discriminatory outputs on dynamic datasets, or deployed models. As all research up-to-date focusses on the identification of discriminatory bias in stationary settings, there is a need for research on developing a system that allows developers and data-scientists to effectively monitor fairness, across diverse scenarios and applications. This study contributes in the following ways: firstly, it presents a set of monitoring techniques from the statistical field, that have not been previously applied in this context, demonstrating their potential for monitoring ML systems. Secondly, it showcases the efficacy of these

diverse techniques in various biased scenarios. Lastly, it advocates an effective and comprehensive method-combining monitoring approach, which enables a nuanced assessment of discriminatory bias.

This paper is structured as follows: to begin, a literature review introduces the fundamental concepts of fair ML and explores relevant research findings essential for this study. Subsequently, the literature review explores various monitoring methodologies that will be utilized in the later stages of this research. Afterwards, the research methodology will be outlined, including an explanation of the datasets used in this research. Next, the results will be presented and the primary findings will be highlighted. Last, these results will be discussed and the conclusions and recommendations for further research will be deliberated upon.

2 Literature Review

This literature review consists of a substantive literature review and a methodological literature review. The substantive review will look into how ML models capture and often amplify discriminatory bias. This is done by investigating the different sources of discriminatory bias in ML models. In the methodological review several aspects of the monitoring of discriminatory bias are covered. First, different statistical definitions of discriminatory bias and corresponding metrics for the quantification of discriminatory bias will be discussed. Second, several researches will be introduced that have attempted to either audit or monitor discriminatory bias in ML models. Last, monitoring techniques originating from the statistical literature will be showcased.

2.1 Substantive literature review

2.1.1 Sources of discriminatory bias

The prevalence of ML on decision-making processes is on the rise, and thereby increasingly influencing people's lives. When problems with the data or the development process arise, ML algorithms frequently have unforeseen implications since they generalize existing data patterns to previously unexplored data. These effects have been seen in predictive policing and face recognition for instance. Often these unforeseen implications involve 'discriminatory bias'. As defined by Buolamwini and Gebru (2018), discriminatory bias is the unfair treatment or unequal representation of certain demographic groups due to biased training data or algorithms. Throughout this research, 'discriminatory bias' and 'fairness' will be used interchangeably, both referring to the same concept, though from contrasting angles.

Frequently, anomalous data—defined as an unexpected or potentially damaging aspect of the data—is held responsible for these undesirable results (Suresh & Guttag, 2019). However, the data itself is influenced by a number of factors, such as measurement mistakes and the historical background, hence the issues are not brought on solely by the data. After the data has been acquired there are several phases within the ML pipelines which can exacerbate these issues and lead to unintended consequences. Understanding these various sources is important because it enables a comprehension of how bias is generated, why it persists and the complex interplay of factors contributing to discriminatory bias in ML systems.

There is ample literature on different frameworks and alternatives to mapping the different sources of discriminatory bias in the ML operational process. The majority revolve around a similar set of sources of bias, which are all captured in one of the most renowned frameworks proposed by Suresh and Guttag (2019):

2.1.1.1 Historical bias

Historical bias arises when a model produces undesirable outcomes, even though the data is measured and sampled in an accurate manner. As discrimination is prevalent in society, data can be representative of a population, but may still be biased. When ML models are trained on this data, these biases will be represented or even amplified. As a result certain identity groups can suffer representational harm. For instance, should image search results for 'CEO' reflect the fact that just 5% of fortune 500 CEOs are women? Even though the evidence correctly depicts the world, it can be interpreted as unfair (Zarya, 2018).

2.1.1.2 Representation bias

When specific regions of the input space are underrepresented, representation bias develops, which results in a less reliable model for minority populations. This might happen when sampling techniques only cover a small percentage of the population or the population of interest has changed. Representation bias can be mathematically described by the divergence between the true data distribution P and the sampled distribution $\hat{P}$. If $D(\hat{P} \,||\, P) \geq 0$, where D is a measure of divergence, representation bias may occur. When the sampled distribution $\hat{P}$ under-samples a specific group, the learned function f will have higher uncertainty for new (x, y) pairs from that group. For instance, only 1% and 2.1% of the photos in ImageNet, a popular image dataset often used for training image recognition models, are from China and India respectively. As a result, classifiers trained on this dataset perform much worse on images that are originating from these countries as compared to other (western) countries (Shankar, et al., 2020). This occurred in Amazon's facial recognition software which showed significantly worse performance on certain ethnicities.

2.1.1.3 Measurement bias

Measurement bias occurs when the input data used to train a model is systematically incorrect or incomplete, leading to biased results. Using sub-optimally representative features and labels, that don't effectively capture the information needed for accurate modeling, causes measurement bias. This results in differing levels of measurement inaccuracy between groups. This might happen when the described classification goal is oversimplified or when there are differences in the quality or granularity of the data between groups. Mathematically, this can be expressed as the difference between the true label y and its proxy $\tilde{y}$. If the relationship between y and $\tilde{y}$ is systematically different across groups, measurement bias occurs. For instance, arrest rates are utilized in predictive policing apps to gauge crime rates, which increases the likelihood of false positives for some communities due to inconsistent enforcement methods

across neighborhoods (Dressel & Farid, 2018). Measurement bias and societal bias are related as societal bias (historical bias) can influence the way that data is collected (measurement bias). Conversely, the presence of measurement bias can further perpetuate historical biases.

2.1.1.4 Aggregation bias

Aggregation bias arises when a one-size-fits-all model is used for groups with different conditional distributions, $p(Y|X)$, X representing the features and Y the labels. Suppose we have groups A and B with different relationships between X and Y. If a model is trained with the assumption that $p(Y \mid X)$ is consistent across both groups, it might result in suboptimal performance for one or both groups. One model might not be the greatest fit for any group, even if they are all equally represented in the training data. For example, diabetic patients have known disparities across ethnicities making a single model suboptimal for addressing the diverse needs of different subpopulations (Giffen et al., 2022).

2.1.1.5 Evaluation bias

When the benchmark data for an algorithm doesn't accurately reflect the target population, evaluation bias develops. Evaluation bias is related to the discrepancy between the evaluation data distribution Q and the target population distribution P. If $D(Q \,||\, P) \geq 0$, where D is a measure of divergence, evaluation bias occurs. The specific measurements used to report performance can mask subgroup underperformance or differences in mistake kinds, which can exacerbate evaluation bias. For instance, representation bias occurred in facial analysis algorithms which underperformed on dark-skinned females. The algorithms showed a good overall model performance, but benchmarks failed to identify and correct the unfair outcomes (Ryu et al., 2018). Its relation to representation bias is hence that representation bias causes models to be unfair or inaccurate towards certain groups, whereas evaluation biases masks these issues by suggesting the model is performing good, even though in reality this is not the case.

2.1.1.6 Additional bias sources

In addition to the sources identified by Suresh and Guttag (2019) there are many approaches considering additional discriminatory bias sources. As identified by van Giffen et al. (2022), the most pertinent additions to this framework are deployment bias and feedback bias. Deployment bias originates when the ML model is used and interpreted in a different context than it was built for (Mehrabi et al., 2019; Olteanu et al., 2019). For instance, ML models capturing certain risk assessments to forecast the likelihood that a criminal will commit a future crime. When in reality, the model is being applied to different situations, such as deciding how long offenders' sentences should be. Feedback bias on the other hand is when the

outcome of a ML model influences the training data such that a small bias can be reinforced by a feedback loop. For instance, if a piece of content receives a high ranking in a rating algorithm based on how many times it has been clicked, this will alter its positioning and promotion, which will result in additional clicks (Mehrabi et al., 2019; Olteanu et al., 2019).

2.2 Methodological literature review

2.2.1 Statistical Definitions of Fairness

A crucial question one must consider is how 'discriminatory bias' or 'fairness' is defined in ML environments. Anti-discrimination laws in most nations forbid unfair treatment of people based on certain sensitive features such as gender, religion or race (VII of the Civil Rights Act, 1964). Often these laws use two distinct concepts to express fairness: disparate treatment and disparate impact (Barocas & Selbst, 2016). Disparate treatment occurs when the decisions are (partly) based on the subject's sensitive attribute, and it has disparate impact if its outcomes disproportionally hurt or benefit people with certain sensitive features. The challenge is however, that these two definitions of fairness are too abstract to mathematically quantify. Although, many papers suggest different sets of metrics and different applications of these metrics, as will be discussed in section 2.2.2, the basis of most of these approaches are based on the different metrics of the confusion matrix: the false positive (FP), false negative (FN), true positive (TP) and true negative (TN). These metrics however don't serve as theoretical concepts themselves for achieving fairness. In this section, six concepts that serve as a theoretical foundation for achieving fairness, will be discussed. These concepts form the basis for the majority of the different fairness metrics (Gajane & Pechenizkiy, 2018; Chen et al., 2023). Note, that fairness metrics provide a way to operationalize the goals set out by these six concepts.

2.2.1.1 Fairness through unawareness

Fairness through unawareness constitutes excluding the sensitive attribute as a feature in the training data. This concept is consistent with disparate treatment, which assumes exclusion of the sensitive attribute. Mathematically, fairness through unawareness can be expressed by stating that a model's predictions should be independent of the sensitive attribute. If $\hat{Y}$ is the predicted binary outcome, A is a binary indicator of the sensitive attribute, the definition is as follows

$$P(\hat{Y}|A) = P(\hat{Y}). \quad (1)$$

One way to validate this concept could be to check for the independence between the output and the sensitive attributes. Chi-square tests of Independence or a Fisher's exact test can be used for this.

Although the approach is easy to understand, it does come with its limitations as it fails to account for existing biases in the data, and can therefore still lead to unfair outcomes. Put simply, if there are (many) highly correlated features that are proxies of the sensitive attribute, removing the sensitive attribute will not have the desired impact (Gajane & Pechenizkiy, 2018). Additionally, oftentimes it can be objectively relevant to include the sensitive attribute, such as in the healthcare domain.

Opposing the fairness through unawareness (philosophical) school is the fairness through awareness school, whom advocate including the sensitive attribute. The remaining concepts elaborated upon in the remainder of this section pertain to the latter.

2.2.1.2 Demographic Parity

Demographic Parity, also commonly referred to as Independence or Statistical Parity requires the output to have the same distribution of predictions for each group defined by the sensitive attribute. To put it differently, the probability of a positive outcome should be the same for all groups, irrespective of their underlying distribution (Gajane & Pechenizkiy, 2018). Mathematically Demographic Parity, assuming binary classification, can be defined as follows

$$P\big(\hat{Y} = 1 \big| A = 0\big) = P\big(\hat{Y} = 1 \big| A = 1\big). \qquad (2)$$

To statistically test demographic parity it must be tested if the probability of a positive outcome is the same across groups. This could be done using a Chi-square test or Fisher's exact test. Alternatively, a Z-test or a t-test could be used to compare the proportions of positive outcomes in different groups.

Demographic Parity also has its flaws. In some cases, it may provide a significant percentage of false positives or false negatives for some groups because it fails to account for the true distribution of the target variable in each group. Additionally, demographic parity may not be appropriate when the objective is to maximize accuracy or minimize another measure of error.

2.2.1.3 Equalized Odds

Equalized Odds, commonly referred to as Separation or Positive Rate Parity, was first introduced by Hardt et al. (2016) and further elaborated upon by Zafar et al. (2017). It requires the algorithm to have equal FP rates and equal FN rates for each group defined by the sensitive attribute. This can be expressed as

$$P\big(\hat{Y} = 1 \big| Y = 0, A = 0\big) = P(\hat{Y} = 1 | Y = 0, A = 1) \qquad (3)$$

and

$$P(\hat{Y} = 0 | Y = 1, A = 0) = P(\hat{Y} = 0 | Y = 1, A = 1) \quad (4)$$

where Y represents the ground truth label, equation *3* ensures equal false positives across groups and *4* ensures equal false negatives across groups. The advantage of Equalized Odds compared to Demographic Parity is that it provides incentive to reduce errors uniformly across all groups. A limitation of the Equalized Odds is that it often comes at expense of the model accuracy. To satisfy the notion, a model may need to purposefully make incorrect predictions, in turn negatively affecting the accuracy.

To test for equalized odds tests as a Z-test or a t-test could be used to compare the occurrences and proportions of FPs or FNs across various groups.

2.2.1.4 Predictive Rate Parity

Predictive Rate Parity is a fairness criterion that requires classifiers to have equal Positive Predictive Values (PPV) and equal Negative Predictive Values (NPV) across groups defined by the sensitive attribute. Hence, given the predicted outcome and the sensitive attribute, the likelihood of a true outcome should be the same for all groups (Zafar et al., 2017).

Mathematically Predictive Rate Parity can be defined as

$$P(\hat{Y} = 1 | Y = 1, A = 1) = P(\hat{Y} = 1 | Y = 1, A = 0) \quad (5)$$

and

$$P(\hat{Y} = 0 | Y = 0, A = 1) = P(\hat{Y} = 0 | Y = 0, A = 0) \quad (6)$$

where equation *5* represents the equal PPV and *6* represents equal NPV. What Predictive Rate Parity fundamentally ensures is similar accuracy for all groups with respect to both positive and negative predictions. The flaw however is similar to that of Equality of Opportunity in that it can often be at the expense of accuracy. Testing for Predictive Rate Parity can be done using a Z- or t-test, similar to equalized odds.

2.2.1.5 Individual Fairness

Individual Fairness, as first proposed in one of the most influential papers in the fairness field by Dwork et al. (2012), takes on a different angle than the previous definitions. As the name implies Individual

Fairness focusses on equal treatment on the individual level, as opposed to the group level which was the case for the previously discussed metrics. The criterion ensures that similar individuals are treated similarly by the algorithm, regardless of their group membership or sensitive attributes.

Individual Fairness states that the distance between probabilities of certain outcomes assigned to two individuals (i.e. the similarity of treatment) should not be larger than the similarity distance between the individuals (i.e. the similarity of individuals). It works by measuring similarity of individuals and similarity of outcomes, representing these two measurements in a similar format (as similarly scaled distance metrics) and constraining how far apart they can be.

A limitation to Individual Fairness is that it is often difficult to determine what an appropriate metric is for measuring similarity between individuals, and moreover the ability to accurately measure this similarity (Kim et al., 2018). Besides, Individual Fairness may often conflict with other fairness criteria or overall model performance. Testing for individual fairness is an intricate and complex issue and an area that is undergoing extensive research (Yeom & Fredrikson, 2020).

2.2.1.6 Counterfactual fairness

As first introduced by Russel et al. (2017), Counterfactual Fairness focusses on ensuring that the algorithm's decisions (would) remain the same, if a sensitive attribute were to be changed *ceteris paribus.* Counterfactual Fairness thus examines if an individual's output would change if their sensitive attribute would be different. Predictor $\hat{Y}$ is counterfactually fair if

$$P\left(\hat{Y}_{A\leftarrow a} = y \,\middle|\, X = x, A = a\right) = P\left(\hat{Y}_{A\leftarrow a'} = y \,\middle|\, X = x, A = a\right). \qquad (7)$$

For all observed features x such that they only differ in their sensitive attribute A with values a and a', with predicted outcome y (Russel, Kusner, & Loftus, 2017). The equation is hence saying that for the actual individual and for the counterfactual individual, the probabilities of a particular outcome y should be the same. Changing the sensitive attribute from a to a' should therefore not change the probability of the outcome.

Relating this back to the previous metrics, Fairness through unawareness struggles to capture the entire scope of discriminatory bias due to correlations among features. Demographic Parity, Equalized Odds and Predictive Rate Parity, are all observational fairness criteria limiting their ability to identify and address the root causes of observed disparities. Individual Fairness, presents its own challenge in determining an appropriate measure of similarity between individuals. Counterfactual Fairness in turn solves all these issues by directly interrogating the causal impact of sensitive attributes on outcomes.

By exploring hypothetical scenarios through a causal graph, counterfactual fairness helps to understand what would have happened if the sensitive attribute were different, thereby offering a nuanced approach to identify biases. Achieving counterfactual fairness can however be challenging in practice, as it often requires modeling the causal relationships between the sensitive attributes, other input features and the outcome. This may entail applying methods from causal inference, such as structural causal models or interventions.

2.2.1.7 The Impossibility theorem of fairness

The different notions of fairness are all context dependent and may very well conflict with one another, or with the overall model performance. A theory that expresses this is the Impossibility Theorem of Fairness. As introduced by Jon Kleinberg et al. in his paper '*Inherent Trade-Offs in the Fair Determination of Risk Scores'* (2017), The Impossibility theorem of Fairness states that any two of the three criteria 2 – 4 are mutually exclusive. The authors prove that it is, in all non-degenerative cases, impossible to satisfy all three fairness criteria. The authors use the following arguments:

(1) Demographic Parity versus Predictive Rate Parity

If A is dependent on Y, then either Demographic Parity is satisfied or Predictive Rate Parity is satisfied, but both cannot be achieved simultaneously.

$$A \not\perp Y \text{ and } A \perp Y \mid \hat{Y}, \text{then } A \not\perp \hat{Y} \qquad (8)$$

(2) Demographic Parity versus Equalized Odds

If A is dependent on Y and $\hat{Y}$ is dependent on Y, then either Demographic Parity is satisfied or Equalized Odds is satisfied, but achieving both at the same time is not possible.

$$\hat{Y} \perp A \text{ and } \hat{Y} \perp A|Y, \text{then either } A \perp Y \text{ or } \hat{Y} \perp Y \qquad (9)$$

(3) Equalized Odds versus Predictive Rate Parity

Assuming that all events in the joint distribution of $(A, \hat{Y}, Y)$ have a positive probability, if A is dependent on Y, either Equalized Odds is satisfied or Predictive Rate Parity is satisfied, but both cannot be achieved simultaneously.

$$A \perp \hat{Y}|Y \text{ and } A \perp Y|\hat{Y} \text{ implies } A \perp (\hat{Y}, Y), \text{which implies or } A \perp Y \qquad (10)$$

The key takeaway from the theorem is that there are inherent trade-offs and complexities associated in applying all of the different fairness criteria, and it highlights that the urgency of considering the specific context of application and the objectives of the decision-making process when choosing which criteria to prioritize. Many different papers enunciate the importance of carefully considering the context of the objective, models and data at hand before selecting a fairness metric. There are inherent trade-offs between all the different metrics and they all have their own sets of limitations (Pleiss et al., 2017; Vasudevan & Kenthapadi, 2020). Therefore, selecting and interpreting different metrics must be done with great care.

2.2.2 Selecting fairness measures

Due to the vast amount of different definitions of fairness and their corresponding metrics, several researchers have attempted to create guidelines for which circumstances call for which fairness metrics. Saleiro et al. (2018) created a seminal framework, widely recognized in the fairness research community, called AEQUITAS. The framework uses questions such as '*Do you want to be fair based on disparate representation, or, based on disparate errors of your system?*', to guide the user towards the most suitable fairness metric. The framework is visualized in the AEQUITAS Fairness Tree, illustrated in Appendix A.

The first split in the tree is an important one and is made by the question mentioned above. The question splits the tree into on the one hand metrics that can only be measured in the presence of a ground truth label (when measuring for fairness based on disparate errors), and metrics measured without the presence of a ground truth label (measuring for disparate representation). The metrics requiring a ground truth label are metrics much more widely used in fairness research as the large majority of research focuses on proving how and why certain static scenarios are biased or not. Alternatively, prevailing research focusses on creating models on such static scenarios that can mitigate the bias. Metrics that include ground truth labels can quantify fairness much more accurately, making them a better fit for such research objectives.

Metrics requiring a ground truth label include metrics as the False Positive Rate Parity (FPR Parity) or False Omission Rate Parity (FOR Parity). Related to monitoring deployed ML models, ground truth labels are often not readily available as many models don't have direct, or short term, access to ground truth labels. As outlined by Buyl and De Bie (2022) in their paper 'Inherent Limitations of AI Fairness', one of the greatest limitations to current fairness literature is that the overwhelming majority focusses on fairness measurement with ground truth labels, making much of the methods and insights challenging to implement in real life contexts. As there are many scenarios where ground truth labels are available within a reasonable time-frame, both monitoring scenarios will be investigated in this paper.

Zooming in on the non-ground truth based fairness metrics there are two different metrics: Equal Selection Parity and Demographic Parity (Saleiro, et al., 2018). The split between the two metrics is defined

by the need of selecting an equal number of people from each group (Equal Selection Parity) or selecting proportional to their percentage in the overall population (Demographic Parity). The former can be measured through simply counting the number of individuals selected from each group and correspondingly comparing them, whereas the latter can be measured by comparing ratios or means and can be tested through for example a t-test.

Apart from these two measures outlined in the *AEQUITAS* framework, there are also more sophisticated measures that can be employed for measuring discriminatory bias in the absence of ground truth labels. A researched conducted by Aka et al. (2021) investigated eight different metrics ranging from Mutual Information to Demographic Parity, in order to determine which metric is the best representation of discriminatory bias in the absence of a ground truth label. The authors found in their research the normalized Pointwise Mutual Information (nPMI) to be the best in identifying discriminatory bias in the absence of ground truth labels. The nPMI determines whether a value of one category or group co-occurs with an element of a target category more frequently than would be expected by chance. This is a crucial difference to make since even in the absence of bias, components from various categories may co-occur purely by chance. The researchers do however state that their findings have only been verified on the use-case discussed in their paper and that it needs more research for validation of their findings. Disappointingly, this research is the only research that has been conducted focusing on the comparison of different fairness metrics in the absence of ground truth labels, thereby illuminating the existing research gap in the field of fairness literature.

2.2.3 Other considerations for measuring discriminatory bias

2.2.3.1 The fairness and model-performance trade-off

The complex trade-off between model performance and fairness is a central challenge within ML fairness research, with no one-size-fits all solution (Fontana et al., 2022). As outlined by Berk et al. (2018) this balance is ultimately determined by ‘political’ decision making, stakeholders’ values, regulations, and the specific context of the problem at hand. Many research has been done on this topic, where the main focus is to create solutions where both performance and fairness are maximized. For instance, research performed by Reich and Vijaykumar (2021) and Corbett-Davies et al. (2017) suggest algorithms fine-tuned for accuracy whilst still adhering to certain fairness metrics. These studies show that there is no clear cut solution to this challenge and that the appropriate trade-off between model fairness and performance is very context dependent and requires domain knowledge. This underscores the importance of being able to monitor models with respect to fairness and implementing safeguards that can detect and flag discriminatory outputs as they arise. By employing such measures, unless flagged as necessary, models can

be optimized for their performance, without explicitly optimizing for fairness. Hereby, acknowledging the reality that in practice, developers and businesses often prioritize model performance over fairness.

2.2.3.2 Data- and Concept Drift

Deployed models cannot guarantee consistent performance over time. One of the reasons for this is that the underlying data and their generating processes can change stochastically. In this respect, models can be employed on different sub-populations than they were initially trained for, or trained on, often having detrimental effects on the models' output. This phenomenon, called drift, has been well studied in literature and can arise in a sudden or gradual form (Stanley, 2003). Data drift occurs when new data significantly differs from the training data due to real-world data constantly changing. Concept drift occurs when the relationship between the model output and feature variables changes over time, and this can have a large impact on the fairness of a model (Ghosh et al., 2022). They stress the importance of tracking drift in combination with fairness metrics, though they also highlight the limitations of combining fairness metrics with drift monitoring techniques. Being aware of data and concept drift is crucial when monitoring model fairness in ML models, as these drifts can not only substantially influence model performance but also have a profound impact on model fairness.

2.2.4 Auditing and monitoring discriminatory bias

With the growing emphasis on algorithmic fairness in the ML domain, the amount of research conducted related to evaluating and scrutinizing models for fairness is also gradually increasing. However, research on the monitoring of ML models with respect to fairness is still falling behind. At this point it's critical to distinguish between ML model monitoring and ML model auditing, as the large majority of the related fairness research has been done on the latter topic. Model monitoring involves the continuous tracking of a model's performance and fairness metrics in real time, after the model has been deployed. Model auditing, in contrast, is a one-time evaluation procedure that normally takes place prior to deployment in order to scrutinize a model's fairness or performance. This section aims to illustrate the research conducted on ML auditing, which can be valuable with respect to ML model monitoring. Moreover, it aims to introduce the limited research that has been conducted on the monitoring of fairness in ML models.

Koyishama et al. (2021) provide an excellent starting point with their comprehensive conceptual approach to fairness auditing and monitoring, offering valuable insights relevant to this study. First, a 'stop-light dashboard' is recommended, using stop-light interfaces (green, orange and red) for monitoring performance over time in terms of high-, satisfactory- and poor-performance, which helps assess

discriminatory bias. Second, the researchers suggest explicit certification of models, entailing that the model must satisfy a specific standard related to fairness before being deployed. The researchers however don't suggest how this standard should be set, or tested for.

Building on the concept of model auditing, there is research that is explicitly focused on auditing fairness, as shown by the AEQUITAS toolkit paper and other examples. As previously touched upon, AEQUITAS is a toolkit made specifically for auditing bias, for developers and policymakers to track different bias metrics and easily visualize different variable distributions. Similar to AEQUITAS there are many other toolkits which can be used for measuring such metrics such as FairML (Adebayo, 2016), Fairness Measures (Zehlike et al., 2017) or Themis (Bantilan, 2018). Although these toolkits are user-friendly and can be convenient for the calculation of different fairness metrics, they lack the monitoring systems that flag the presence of discriminatory bias as they focus only on stationary settings. There are more advanced auditing methodologies that have been proposed as illustrated by Xue et al. (2020) who have developed a fairness monitoring technique based on optimization and introduced the FaiTH test statistic to assess model fairness. The same researchers also published another paper where they proposed an approach that employs statistical inference for individual fairness by examining models through gradient analysis (Maity et al., 2021). Although the researchers contend in both situations to have found efficient solutions for auditing ML fairness, they can only test on one specific form of discriminatory bias, are quite difficult to implement and are tested in a stationary setting, over a stationary dataset.

A study that does assume a dynamic setting and hereby provides insights into the concept of monitoring in practice is performed by Wilson et al. (2021). The study uses a case study of pymetrics, a company that, among others, employs ML models for job candidate recommendations. The researchers highlight in their study that 'fairness is a performance criterion', thereby emphasizing the need to create guidelines and thresholds on which fairness must adhere to. In the use case of the study the researchers use the 4/5ths rule set by the Equal Employment Opportunity Commission (EEOC), which states that the lowest passing group has to be within 4/5ths of the pass rate of the highest passing group. Accordingly the researchers create and select a model and monitoring test (chi-squared test) to ensure that the model does not cross this threshold. What Wilson et al. (2021) clearly underscore in their publication is that ML models should have transparent thresholds that should be monitored, preferably by third-party organizations. However, the study concentrates on a domain application that already possesses well-defined and transparent rules regarding fairness (as nationally set in America by the EEOC). In this context, monitoring fairness is simpler and more straightforward compared to typical scenarios where discriminatory bias emerges. Furthermore, the study primarily emphasizes the conceptual aspects of establishing a monitoring framework, disregarding challenges that may arise during practical implementation, such as potentially incorrectly assuming discrimination to be present.

In conclusion, it is clear that the literature is significantly biased towards auditing and lacks workable monitoring solutions for real-world, dynamic settings. Additionally, a lot of current research focusses on particular use cases or particular types of discriminatory bias, which restricts their practical applicability across diverse, real-world scenarios. This points to the imperative need for further research in the development of monitoring frameworks that can effectively recognize and flag discriminatory bias in deployed ML models.

2.2.5 Statistical Process Control

Statistical Process Control (SPC) is a general methodology of quality control that leverages statistical methods to monitor and control a process and to ensure the process to operate efficiently. As first introduced by Dr. Shewhart in the 1920s the methodology was intended to monitor and quality control production lines and products. By plotting sample measurements over time, charts enabled operators to identify when a process was operating within acceptable limits or showing unusual variation. This allowed for timely intervention and adjustment to maintain consistent product quality. In other words, control charts, which are the main elements of SPC, show process data over time and include control limits to differentiate between common-cause variation (natural variation) and special-cause variation (Qiu, 2013). By doing this SPC can distinguish systematic changes from chance variation in processes.

There are a variety of different control charts used in SPC, each designed for different types of data, different applications and different aspects of process variation. Shewhart charts include the p-chart (and np-chart) that tracks the proportion with an event for consecutive periods, and the g-chart, that displays the number of cases between events and is specifically designed to detect reductions in event rates. More intricate charts that accumulate information over time include the exponentially weighted moving average (EWMA) chart and the cumulative sum (CUSUM) chart. The EWMA is a weighted moving average of current and past outcomes and is updated after each observation with exponential weight, meaning that the contribution of past observations decreases going back in time. The CUSUM is a metric summarizing the evidence of a shift from the baseline, where higher values indicate stronger evidence. The CUSUM and EWMA can generally detect small increases in event rates more quickly than the p-chart (Qiu, 2013). An ideal chart would take a short time to signal a genuine change in performance and a long time to falsely signal a change.

Although the majority of the SPC literature focusses on industrial applications, there are various other SPC applications. Like the research conducted by Neuburger et al. (2017) which uses clinical data as a use-case to compare the different control charts and the variations in their respective performance for binary data. The research monitored the performance of the different charts using the average run length (ARL), a commonly used metric in SPC, which translates to the average number of observations until a

signal. The out-of-control (OC) ARL measures the average number of observations needed for a chart to detect that the process has actually shifted from its stable state, where a lower OC ARL indicates a better performance. The in-control (IC) ARL on the other hand measures the average number of observations until a false alarm is signaled, where a higher IC ARL is preferred. The researchers found the CUSUM to be the quickest chart in signaling small absolute event rate changes, as well as larger event rate changes. The EWMA was shown to be the runner up in terms of performance, whereas both the Shewhart charts (p-chart and g-chart) were only effective in specific settings. The researchers note that the Shewhart charts are the most accessible, particularly with respect to setting it up. The CUSUM chart on the other hand was found to be the most difficult to construct.

SPC is known for its efficacy in supervising the performance of various processes and in identifying deviations from historical trends and observations. Although there are no documentations of SPC being utilized for ML model monitoring, or for its use in tandem with ML fairness, it holds promise as a valuable tool.

2.2.6 Sequential Analysis

SPC plays an instrumental role in monitoring alert frequencies and can therefore prove to be a useful tool for indicating bias. However, the question remains: Can one ascertain with statistical confidence that a model is exhibiting bias at a given time? To answer this, Sequential Analysis (SA) will be further explored, a method that offers promise of detecting and quantifying bias with statistical certainty.

Often, the volume of data evaluating a model's performance increases over time, making it crucial to continually update the comprehension of the model's fairness as new data emerges. However, if naïve statistical tests were to be applied, also known as Null Hypothesis Significance Tests (NHST), for testing discriminatory bias in a sequential setting, problems arise. NHST, by design, is intended for one-time comparisons. When the same hypothesis is continually tested with growing data, the probability of wrongfully rejecting the null hypothesis at least once increases over time. In other words, NHST does not account for the accumulation of the risk of Type I errors when applied multiple times to iteratively updated data. To put this into context, statistical testing for discrimination can often involve examining mean similarities between population groups, or setting a pre-specified threshold for a fairness metric or some other value. However, with an increasing number of tests over time, the chances of falsely flagging discrimination increases. SA, also known as sequential testing, is a statistical method that allows for the continuous assessment and testing of data, presenting a resolution to this issue.

Group Sequential (GS) design is one of the most commonly known sequential analysis techniques and also serves as a good proxy for explaining the fundamentals of SA (Schönbrodt et al., 2017; Lakens, 2022). With a GS design, the specific intervals and sample sizes for interim tests and a final test are

established before the study begins. The spacing of the sample sizes for the interim tests, along with the key values for the test statistic at each phase, are calculated to ensure that the overall Type I error rate (failing to reject a false null hypothesis) remains within acceptable limits. If the test statistic crosses an upper boundary during an interim test, the data collection process is halted prematurely and a conclusion is drawn. This is because the effect has proven strong enough to be reliably detected in the smaller sample, a strategy often termed as 'stopping for efficacy' (Schönbrodt et al., 2017; Lakens, 2022). If the test statistic doesn't reach this boundary, data collection continues until the next interim test or the final test is scheduled. Some GS designs also include a 'stopping for futility' feature, triggered when the test statistic drops below a lower boundary. In this scenario, it's unlikely that a significant effect could be detected, even if the sample size were increased to the maximum, resulting in accepting H_0. This is a valuable feature in applications such as clinical trials or quality monitoring, where larger sample sizes can significantly increase costs. In the context of testing for fairness this feature is nevertheless redundant, as one cannot conclude over a deployed model its output is fair and will therefore remain fair in the future. The more frequent interim tests that are conducted, the larger the sample size must become to maintain the same level of power as a fixed-n design without interim tests. However, if a significant effect does in fact exist, there's a notable chance that the data collection will stop before reaching the maximum sample size. This early detection in GS, and SA, often reduces the need for a larger sample size that is typically required in NHST, where testing is done at the end of the sample (Lakens, 2022; Schönbrodt et al., 2015).

Many alternative procedures to SA exist. As first introduced by Wald in 1945 and later elaborated upon by Schnuerch and Erdfelder (2020), Sequential Ratio Probability Tests (SPRT) are optimally efficient if researchers are able to analyze the data after every additional observation is collected. SPRT and GS analysis are similar to each other in terms of functioning. The only difference lies in that GS design data is collected in groups, whereas SPRT is designed to be able to analyze data on an ongoing basis as individual datapoints are collected. As opposed to GS analysis, SPRT can be more efficient as it enables real-time decision making, and provides more flexibility with respect to the number of observations and group sizes (Schnuerch & Erdfelder, 2020). An alternative to the SPRT is the General Sequential Ratio Probability Test (GSPRT), which extends the approach to more general situations involving composite hypotheses (Li et al., 2016).

As opposed to GS testing, where the interim tests are planned a priori, Sequential Bayes Factors (SBF), allow for unlimited multiple testing. As introduced by Schönbrodt et al. (2017) SBF offers an alternative to group sequential testing and traditional null hypothesis significance testing, using Bayesian statistics to incorporate prior information into the analysis. Similar to GS testing SBF incorporates stopping rules, however based on different statistics (i.e. the Bayes factors). In the context of sequential testing for

discrimination over deployed ML models, one of SBFs greatest advantages lies in the fact that Bayesian statistics are not affected by the number of times tests are performed on the data (Schönbrodt et al., 2017).

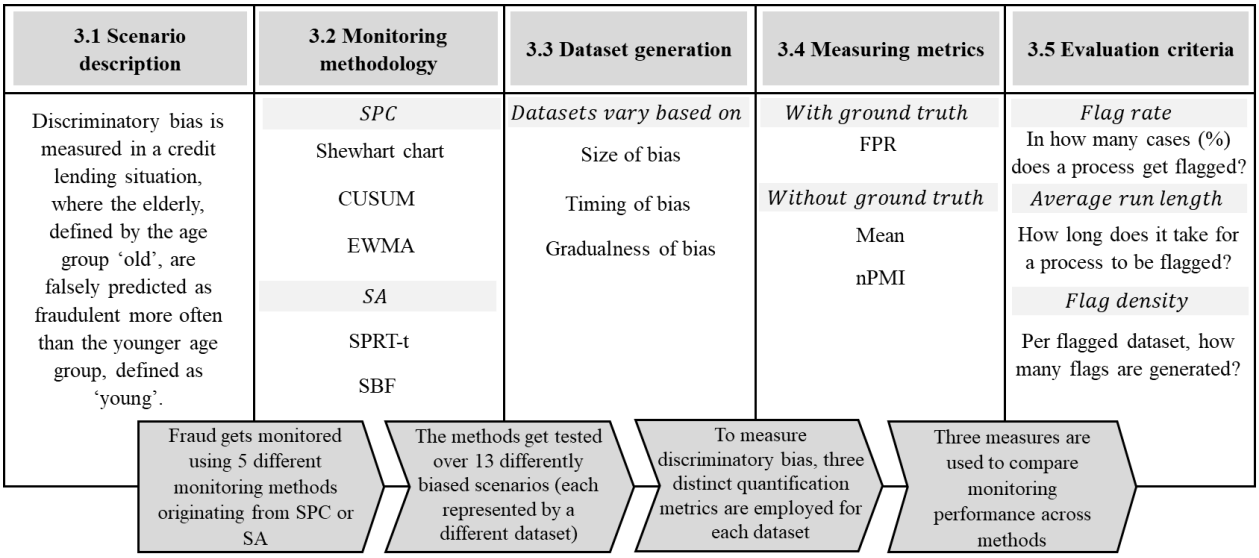
Another alternative to sequential testing, Safe Testing, has been proposed by Grünwald et al. (2019). The research introduces a new alternative to the p-value, the e-value. The e-value is a notion of evidence that, contrary to the p-value, allows for effortlessly combining results from several tests, where the decision to perform a new test may depend on the outcomes of the previous test, which is referred to as optimal continuation. Moreover, the paper introduces another new concept called Growth-Rate-Optimality (GRO), which is similar to power in traditional hypothesis testing, but is adapted for optional continuation scenarios. These GRO e-values are designed to be optimal in the sense for general testing problems with composite null and alternative hypotheses.

These techniques provide strategies that can be adapted to scrutinize the fairness of outputs of ML models in a dynamic context. Their designs provide potential to detect and alert emergence or presence of discriminatory bias as soon as it reaches a certain level of statistical significance. Although not recognized for their use in tandem with ML model monitoring or ML fairness, this research sets out to combine the two.

3 Methodology

The research design used in this paper consists out of five different stages, which will be explained in sections 3.1 to 3.5. Figure 1, illustrated below, explains the different sections of the experimental design, moreover it presents a chronological overview of the methodology chapter. Here is a detailed breakdown of what each step in the figure represents. The study starts by defining the discriminatory biased scenario, which is within a credit lending context (3.1). To analyze the bias a suite of monitoring methods is used, originating from SPC and SA, each presenting different techniques for monitoring bias (3.2). To gauge how these methods behave under different circumstances and how robust they are, they need to be tested across a variety of scenarios. These scenarios are represented in different synthetically generated datasets each simulating different variations of bias (3.3). Once the various biased scenarios have been defined and created, several metrics are employed to quantify, or act as proxy for, the discriminatory bias present. These different representations of discriminatory bias can then be used by the monitoring methodologies to monitor and potentially flag bias (3.4). Having set the scenario, selected monitoring methods, simulated varying biased scenarios and quantified discriminatory bias using specific metrics, the final step is to evaluate the different monitoring methods. Therefore they are assessed based on three specific performance criteria. This allows for evaluating and comparing how effective and efficient the methods are at monitoring discriminatory bias (3.5).

Figure 1. Experimental setup



Before diving into the specific sections, along with Figure 2, a few explanatory notes will be made regarding the experimental setup and underlying assumptions of the monitoring scenario.

- In the described scenario of monitoring discriminatory bias, it is assumed that there is prior knowledge as to which group is being discriminated against by the model, namely the old age group. This assumption may not always hold in practical monitoring settings, but slight adjustments could be made to the methods and monitoring design to account for this uncertainty. This also signifies that when statistical tests are being performed to test for bias (using SA techniques), these are one-sided hypothesis tests.
- Both SPC and SA feature a lower bound threshold to identify when observed values dip below the expected range, indicating potential anomalies or lack of substantial evidence against a hypothesis. In SPC, this lower bound functions as an alert system within a control chart, flagging when a process may be deviating negatively from standard operation. In SA, it serves as a stopping rule, halting the test and accepting a hypothesis if results consistently fall below this limit, indicating the absence of discrimination based on the current data. However, this research won't consider these lower bounds because instances where test statistics, or measurement values are too low typically suggest minimal discrimination, making it unnecessary to flag such instances. Note that this does not necessarily imply an inversion of discrimination towards another group.
- To distinguish between biased and non-biased data, the portions of the datasets are classified as either OC (out of control) or IC (in control), respectively. As such, all generated alerts are labeled and categorized as either OC or IC. This classification facilitates the comparison of different methods, for accurately generating OC alerts for biased data or inadvertently generating IC alerts for non-biased data.

3.1 Scenario description

To monitor discriminatory bias, a scenario is simulated of a sensitive real-life application of ML, where instances of discrimination frequently occur. The scenario, represented by the generated datasets, involves the detection of fraudulent online bank account opening applications at a bank. In this setting, fraudsters will try to either impersonate someone via identity theft, or create a fictional individual to gain access to the banking services. The fraudster promptly maxes out the credit line when given access to a new bank account, or uses the account to receive illicit payments. The bank is responsible for all costs incurred as there is no way of tracing the fraudsters true identity.

This case is regarded as a high-risk area for the adoption of ML. A negative prediction (of an applicant being fraudulent) results in providing access to a new bank account and its credit, whereas a positive prediction results in rejecting the customer’s application for a bank account and flagging them as fraudulent. Since having a bank account is a fundamental right in the European Union (Jesus, et al., 2022),

fraud detection is a very relevant application from a societal standpoint. Moreover, this scenario effectively illustrates the true nature of fairness in ML. Even a slight decline in predictive performance often represents millions in fraud losses. Consequently, in such scenarios, ML models are optimized with respect to performance rather than fairness, emphasizing the crucial role of monitoring practices in preventing discriminatory bias.

3.2 Monitoring methodology

Statistical Process Control

Now that a discrimination-sensitive scenario is defined, the monitoring methods must be selected and defined. To monitor the credit fraud scenario, monitoring methodologies originating from the fields of SPC and SA will be employed, of which the first techniques pertain to SPC. As previously stated, SPC is a methodology used in quality management to monitor and quality control processes, often in industrial settings. It involves the application of statistical techniques to measure and analyze the variability in a process over time, allowing for the identification of patterns, trends and deviations from the desired performance. The primary objective of SPC is to ensure that a process operates within predetermined statistical control limits. By continuously monitoring the process and collecting data at regular intervals, SPC enables organizations to effectively monitor process stability and allows them to improve overall quality. Although SPC methods are most widely known for their application in industrial quality control, they are also applied in other domains such as health care (Neuburger et al., 2017) or software engineering (Card, 1994). The SPC charts that will be used and discussed in this research are the Shewhart chart, the EWMA (exponentially weighted moving average chart) and the CUSUM (cumulative sum chart). All symbols used in the following methodology section are displayed in Appendix B. Moreover, as these monitoring methods are often visualized in charts, a visual representation of all three charts can be found in Appendix C.

3.2.1 Shewhart $\bar{X}$-chart for individual measurements

Control charts for individual measurements, where the sample size equals one, use the moving range between two subsequent observations to measure process variability. The moving range is defined as

$$MR_i = |x_i - x_{i-1}| \tag{11}$$

where x is an individual measurement of a process and i is the measurement's index. The moving range is hence the absolute difference between two consecutive datapoints and is plotted on the control chart. Recall from section 2.2.5 that control charts include control limits to distinguish systematic changes from chance

variation in processes (Qiu, 2013). These control limits are defined by a Lower Control Limit (LCL), an Upper Control Limit (UCL) and a center line in between the two. Even though the LCL will not be considered as an alert threshold in this study due to its representation of bias absence, its construction will still be explained as it helps in gaining a better understanding of the principles behind the control chart. For the $\bar{X}$-chart these plotted lines are defined by

$$LCL = \bar{x} - 3\frac{\overline{MR}}{1.128}, \quad \text{Center line} = \bar{x} \quad \text{and} \quad UCL = \bar{x} + 3\frac{\overline{MR}}{1.128}. \tag{12}$$

Where $\bar{x}$ is the average of all the individual measurements and $\overline{MR}$ is the average of all the moving ranges of two observations. The denominator contains the $d2$ value, which is a statistical constant used to adjust the average moving range to more closely align with the actual process standard deviation, providing a more accurate representation of process variability in control charts. The specific value is determined not by the data itself, but by the subgroup size of the moving range. As the subgroup size is always 2 for this method the denominator becomes a constant of 1.128, the $d2$ value for subgroup size 2 (Montgomery, 2012). The 3 multiple deviation from the center line is defined as the sigma limit and it is common practice in SPC to set this value to 3. In general this ensures that the combined total risk of a Type I and a Type II error is minimized (Lloyd, 2019). However, this value is not set in stone, and depending on the data and the specific use case these limits can be altered, usually widened, in order to avoid frequent false alarms.

In the intended application of SPC, the average, $\bar{x}$, is supposed to be specified beforehand by taking the average value of a process that is IC. In SPC this is referred to as Phase I and Phase II monitoring, where in phase I historical data is analyzed to determine the performance of an IC process and in Phase II new data is analyzed in an ongoing monitoring process. In the context of monitoring discriminatory bias, Phase I SPC would be performed on the test set (or possibly other historical data), assuming that model producers would not deploy a model that is already unfair when being tested on. The test data would in this case represent a process that is IC.

3.2.2 Shewhart p-chart

The Shewhart p-chart is often used for the tracking of the proportion of defective items in a population and is applicable to categorical data. Two variations of the Shewhart chart are used to account for the different data inputs. Suppose that the model is operating in a stable state (i.e. not discriminating), such that the probability that any 'old' individual will be fraudulent is p. If a random sample of n individuals is selected, and D is the number of fraudulent elderly, then D has a binomial distribution with parameters n and p. It follows that the mean and variance of the random variable D are np and $np(1-p)$, respectively.

The sample's proportion, $\hat{p}$, is then defined as the ratio of the number of fraudulent elderly in the sample D to the sample size n; that is (Qiu, 2013),

$$\hat{p} = \frac{D}{n}. \tag{13}$$

From this the mean and variance of $\hat{p}$ follow

$$\mu_{\hat{p}} = p, \qquad \sigma_{\hat{p}}^2 = \frac{p(1-p)}{n}. \tag{14}$$

Suppose that the true fraction of fraudulent elderly p is known, the center line and the control limits can be constructed:

$$LCL = p - 3\sqrt{\frac{p(1-p)}{n}}, \tag{15}$$

$$\text{center line} = p, \tag{16}$$

$$UCL = p + 3\sqrt{\frac{p(1-p)}{p}}. \tag{17}$$

Oftentimes the true value of p is unknown. This means either a standard given value must be chosen for p or an estimate $\bar{p}$ must be made (Qiu, 2013). This can be done in the following manner. If m preliminary samples were to be selected, each with n observations, and there were D_i fraudulent elderly in a sample, the fraction elderly in sample i is

$$\hat{p}_i = \frac{D_i}{n}, \quad i = 1, 2, \ldots, m \tag{18}$$

and the average of these individual sample fractions is

$$\bar{p} = \frac{\sum_{i=1}^{m} D_i}{mn} = \frac{\sum_{i=1}^{m} \hat{p}_i}{m}. \tag{19}$$

The control chart would then be constructed by replacing p by $\bar{p}$.

3.2.3 CUSUM

CUSUM charts, while not as intuitive to operate as Shewhart charts, have been shown to be more effective at detecting small shifts in the mean of a process. It does this by graphing the cumulative sums of the observations' deviations from a target value, by this means directly including all the information in a sequence of observations. This makes the method more sensitive to detecting small process changes than the Shewhart chart (Montgomery, 2012). Assume that j values have been collected and x_i is the value of the ith measurement. Then the CUSUM can be plotted as follows

$$C_i = \sum_{i=1}^{j} (x_i - \mu_0) \tag{20}$$

where μ_0 is the value of the in-control mean of feature x (if no target is available an estimate $\hat{\mu}_0$ is made), and C_i is the cumulative sum up to and including the ith measurement. If the process remains in control centered at μ_0, the CUSUM plot will show variation in a random pattern centered around zero. However, if the mean shifts upwards or downwards to a different value the CUSUM will display a drift in C_i, leading to the identification of a control chart signal (Montgomery 2009).

If the process deviates from the target value μ_0, the CUSUM will generate an alert. The one-sided upper and lower CUSUMs, which are plotted on the chart, work by accumulating derivations from μ_0 that are above target with one statistic C^+ and accumulating derivations from the μ_0 that are below target with another statistic C^-. The statistics C^+ and C^- are computed as follows (Montgomery 2009)

$$C_i^+ = max\left[0, x_i - \left(\mu_0 + k\right) + C_{i-1}^+\right], \tag{21}$$

$$C_i^- = max\left[0, x_i - \left(\mu_0 - k\right) + C_{i-1}^-\right]. \tag{22}$$

Where the starting values C_0^+ and C_0^- are equal to zero and k is a reference value which determines the sensitivity to shifts in the process. It is important to recognize that C_i^+ and C_i^- accumulate deviations from the target value μ_0 that are greater than k, with both amounts being reset to zero when they become negative (Montgomery, 2009). The process is considered OC if C_i^+ or C_i^- surpasses the decision interval h. As a result, the values of h represent the CUSUM chart's control limits. The correct selection of k and h is hence critical as it directly affects the performance of the CUSUM chart. The h and k rely on the following

$$k = \frac{\delta \sigma_x}{2}, \qquad h = dk \qquad \text{and} \qquad d = \frac{2}{\delta^2} \ln\left(\frac{1-\beta}{\alpha}\right) \tag{23}$$

where

- α is the probability of a Type I error;
- β is the probability of a Type II error;
- δ is the amount of shift in the process mean we wish to detect, expressed as a multiple of the standard deviation of the data.

A general rule of thumb to obtain optimal performance, as proposed by Montgomery (2009), is $h = 4$ or 5, $k = 0.5\delta$ and δ set to 3 likewise to the Shewhart $\bar{X}$ chart. Finally, the CUSUM chart is created by charting C^+ and C^- versus the sample number.

3.2.4 EWMA

A permanent change in the process may not instantly result in individual violations of the control limits on a Shewhart control chart as it takes time for patterns in data to emerge. The EWMA control chart is better suited to this purpose (Devore, 2012). The EWMA is a statistic for monitoring the process that averages the data in a way that gives increasingly less weight to data as they are further from the current measurement. The data $x_1, x_2, \dots, x_i$ are the standard measurements ordered by index i, similar to previous methods. The EWMA statistic at the ith measurement is calculated recursively from individual data points, with the first EWMA value, $EWMA_1$, being the arithmetic average of historical data

$$EWMA_{i+1} = \lambda x_i + (1 - \lambda) EWMA_i. \tag{24}$$

Where λ is the weight being applied to past observations and $0 < \lambda \leq 1$. The sensitivity to small changes or gradual drifts is tuned by this value (Hunter, 1986). As λ approaches 1, more weight is assigned to recent data and less weight to older data, while approaching 0 results in the opposite effect (Roberts, 1959). By plotting $EWMA_i$ versus time, the EWMA control chart is created. The center line for the control chart is the average of historical data, unless specified otherwise. The UCL and LCL are defined by:

$$UCL = EWMA_1 + 3\sigma\sqrt{\frac{\lambda}{(2-\lambda)}}, \qquad LCL = EWMA_1 - 3\sigma\sqrt{\frac{\lambda}{(2-\lambda)}} \tag{25}$$

where σ represents the standard deviation of the EWMA statistic and the width of the control limits is set to 3.

Sequential Analysis

The second set of methods falls under the Sequential Analysis procedures. As previously touched upon in section 2.2.6, sequential approaches collect data and constantly analyze and test the data after each data point, or each batch of data points (Wald, 1945). This approach, in the traditional applications of sequential testing, leads to three different outcomes:

1. The testing is *terminated* as there is enough evidence to accept the null hypothesis H_0;
2. The testing is *terminated* because there is enough evidence to accept the alternative hypothesis H_1;
3. The testing *will continue* as there is not yet enough evidence for either of the two hypotheses.

As previously touched upon, this research disregards situation 1, as accepting that there is no discriminatory bias up to date, does not result in a termination of the monitoring process.

3.2.5 Sequential Probability Ratio t-Test

The sequential probability ratio t-test (SPRT-t) is based on Abraham Wald's (1945) SPRT, which is a highly efficient sequential hypothesis test. The SPRT was further refined by Rushton (1950) and Hajnal (1961) utilizing the t-test. The main idea is to convert the observed sequence (which relies on the variance) into a sequence of the related t-statistic (which is independent of variance). H_0 and H_1 are defined as follows, with θ representing the model parameter in the underlying population

$$H_0 : \theta = \theta_0 \quad \text{and} \quad H_1 : \theta = \theta_1. \tag{26}$$

On this test the following requirements are held

$$P(\text{accept } H_i|\theta_i) = \begin{cases} 1-\alpha, & (i=0) \\ 1-\beta, & (i=1) \end{cases} \tag{27}$$

where $P(\text{accept } H_i|\theta_i)$ denotes the probability to correctly accept H_i when θ_i is true. A sequential test is of strength (α, β) when it satisfies these requirements.

The test statistic of the SPRT is based on a likelihood ratio, which is a measure of the data's relative evidence for the given hypotheses. Let $f(X|\theta_i)$ represent the probability function for the observed data X given the population parameter specified in $H_i, i = 0, 1$ (Schnuerch & Erdfelder, 2020). At the *m*-th step of the sampling process, compute a test statistic that conforms to the likelihood ratio, that is, the ratio of probability densities of the observed data $X = x_i, \dots, x_m$ under H_1 versus H_0, which is referred to as LR_m

$$LR_m = \frac{f(data_m|H_1)}{f(data_m|H_0)} = \frac{f(x_1,\dots,x_m|\,\theta_1)}{f(x_1,\dots,x_m|\,\theta_0)}. \quad (28)$$

The model parameter contains the parameters of a normal distribution before being transformed into the t-statistic: the mean, μ, and standard deviation σ. This is why the SPRT test requires prior knowledge about the standard deviation. After transforming the observed values into the associated t-statistic, the model parameter contains the non-central t-distribution parameters: degrees of freedom, df, and the non-centrality parameter Δ

$$f(x_1,\dots,x_m|\mu,\sigma) \rightarrow f(t_2,\dots,t_m|\,df,\Delta). \quad (29)$$

Only the sample size of the group(s) is required for the calculation of the degrees of freedom. The non-centrality parameter additionally requires the estimated effect size in the form of the cohen's d i.e. the true standardized difference of mean(s) of the populations underlying the group(s) (Schnuerch & Erdfelder, 2020). Only the current t_m-statistic is required to compute the LR of the sequential t-test. Rushton (1950) established that an SPRT can be performed at every m-th stage simply by comparing the probability densities for the most recent t_m statistic under the null hypothesis. Hence, the test statistic for a one-sided sequential t-test can be computed as follows

$$LR_{m,\ one-sided\ sequential\ t-test} = \frac{f(t_m|\theta_1)}{f(t_m|\theta_0)}. \quad (30)$$

Following the calculation of the test statistic, the decision will be made to either continue sampling or to stop sampling and accept one of the hypotheses. The rules for this, as introduced by Wald (1945), are as follows

Table 1. Decision rules SPRT-t (Wald, 1945)[1]

Condition	*Decision*
$LR_m < A$	Continue sampling
$LR_m \geq A$	Accept H_1 and reject H_0

[1] Adjusted to discriminatory bias monitoring

The boundary is calculated with previously defined error rates α (Type I error) and β (Type II error) as follows

$$A = \left(\frac{1 - \beta}{\alpha}\right). \tag{31}$$

It is important to note that chosen values for α and β in the SPRT-t test don't guarantee the exact error rates in practice. They represent target values that guide the test's design and decision boundaries. However, the actual error rates may vary depending on many different factors such as for example the sample size or the underlying data distribution.

3.2.6 Sequential Bayes Factor

The last Sequential Testing technique that will be used in this research is SBF. SBF combines Bayesian hypothesis testing with sequential analysis properties. In short, Bayes Factors (BFs) are computed until an a priori set level of evidence is attained. In traditional null hypothesis testing the focus lies on how incompatible the actual data is with the H_0. In contrast, Bayesian hypothesis testing, via BFs, determines if the data are more compatible with the H_0 or an alternative hypothesis H_1 (Schönbrodt et al., 2015).

Priors and posteriors are two fundamental notions in Bayesian statistics. A prior is a probability distribution that expresses beliefs about an uncertain variable before any evidence is considered. For example, the H_0 and H_1 hypotheses are originally viewed as having a priori (prior) probabilities. The posterior, on the other hand, is the updated probability of an event occurring after additional data is included. The prior belief (the prior) gets updated regarding the hypothesis to a new belief (posterior) as more data is included in the experiment. The Bayes' theorem facilitates this updating mechanism.

BFs provide a numerical value that indicates how well a hypothesis predicts the empirical data relative to the alternative hypothesis. Using subjective prior odds, one can create a ratio expressing the probability of the data given the hypothesis. Given the data, the computation of marginal likelihoods yields a factor that allows one to accept or reject the null hypothesis. This leads BFs to be formally defined as:

$$BF = \frac{p(E|H_1)}{p(E|H_0)} \tag{32}$$

If for example $BF = 4$, this signifies that the empirical data E are 4 times more likely if H_1 were true than if H_0 were true. On the other hand, if BF were in between 0 and 1 this would show support for H_0. In NHST as the p-value is disrupted because the framework expects a fixed experimental size, ultimately raising the false positive risk. Bayesian statistics, on the other hand, consider probability as an evolving degree of belief that is updated with each new piece of data, making this less troublesome because

the belief about the hypotheses adjusts continually until either one is accepted. The relationship between the prior odds and the posterior odds is as follows

$$\underbrace{\frac{P(\mathrm{H}_1)}{P(\mathrm{H}_0)}}_{prior} \times BF = \underbrace{\frac{P(\mathrm{H}_1|E)}{P(\mathrm{H}_0|\mathrm{E})}}_{posterior}. \quad (33)$$

When choosing priors, they should not give too much weight to illogical parameter values. For H_0 this can be easier to determine beforehand, in this paper the null hypothesis takes on $H_0: \mu_{Old} = \mu_{Young}$. However, applying a specific value to the mean of prior H_1 can be more problematic, as priors that diverge from the data produce BFs that favor H_0 more. To solve this issue, instead of using a single value for H_1, a distribution is used, for which researchers have found the Cauchy distribution to be the most effective. A Cauchy distribution is similar to a normal distribution but has heavier tails, giving more probability to extreme values. The Cauchy distribution can be adjusted using a scale parameter r, with $0 \leq r \leq 1$, where larger values signify a wider distribution indicating a prior belief that large effect sizes are more likely, and smaller values express the contrary. Hence, as proposed by Rouder et al. (2009) the SBF tests

$$H_0: \mu_1 = \mu_2 \quad \text{against} \quad H_1: \mu_1 \sim Cauchy(r) \quad (34)$$

where r is a scale parameter that controls the width of the Cauchy distribution. This prior distribution defines the plausibility of possible effect sizes under H_1.

The following is an outline of the SBF procedure, as proposed by Schönbrodt et al. (2017), which has been slightly modified for this particular research:

1. Define a priori thresholds that show the requested decisiveness of evidence. These are the thresholds that will define the 'H_1 boundary'.
2. Choose a prior distribution for the effect sizes under H_1. This distribution describes the likelihood that effects of various magnitudes exist.
3. Run the first test over a minimum number of observations, increase the sample size as needed and compute a BF at each stage.
4. Stop sampling and report the final BF if H_1 gets accepted. Plot the entire posterior distribution to observe where potential bias trickled into the model.

The hypotheses get accepted if certain threshold values for the BF are reached. For selecting the thresholds guidelines have been made for accepting hypotheses (Schönbrodt et al., 2017). If 1 < BF < 3, the BF indicates *anecdotal evidence* (in favor of the alternative hypothesis), 3 < BF < 10 *moderate evidence*, 10 <

BF < 30 *strong evidence*, and BF > 30 *very strong evidence*. For this research a threshold of 3 is chosen to accept H_1, however this value is subject to change for different monitoring scenarios.

3.3 Dataset generation

Having established a scenario that is sensitive to discrimination and chosen monitoring methods, the next step is to generate a variety of datasets with varying degrees of bias. This will facilitate the comparison of how these different monitoring methods perform under different bias conditions.

While there are datasets available in the fairness domain, such as the COMPAS dataset (Wadsworth, Vera, & Piech, 2018) or the Bank Account Fraud data suites (Jesus, et al., 2022), which are commonly used for fairness audits or developing fairness-aware algorithms, there is a lack of datasets specifically designed for monitoring discriminatory bias. This further emphasizes the existing research gap in the fairness ML realm regarding the monitoring of discriminatory bias. Therefore this research aims to address this gap by generating synthetic datasets that simulate a variety of situations featuring varying levels and alternations of discriminatory bias. In doing this, this research sets out to assess the effectiveness of different monitoring techniques in detecting and flagging discriminatory biases across a wide range of circumstances.

The datasets generated in this research each contain 10,000 observations and three columns: `customer_age`, `fraudulent` and `ground_truth`. `Customer_age`, the sensitive attribute in the datasets, can be either `young` or `old`. The `old` age group is being discriminated against by being falsely predicted as fraudulent more often compared to the `young` group. The `fraudulent` variable represents the predictions made by a model if an individual is fraudulent yes (`1`) or no (`0`). The `ground_truth` variable represents the ground truth value if an individual was in reality fraudulent yes (`1`) or no (`0`). Take note that the datasets assume hypothetical predictions made by a model, these predictions are not actually made by a model in this research. This approach ensures precise control over the level of bias, enabling a more controlled environment for testing the various monitoring methods.

To assess the robustness of the monitoring procedures different scenarios of discriminatory bias are tested. This research chooses to create different scenarios based on the level of bias, the pace at which bias is introduced and the timing of the introduction of the bias. Bias is introduced, following the guidelines from Saleiro et al. (2018), by an increasing false positive rate (FPR) for the elderly age group. Table 2 illustrates the different datasets used in this research, where the listed FPRs represent the changing FPR for the elderly age group. The younger age group has a stable FPR at the 5% level. As there are no formal guidelines to what ‘large’, ‘medium’ or ‘small’ bias is, these values are chosen based on the bias values of real life discriminatory instances, such as the examples explained in the introduction of this paper.

Table 2. Datasets used in research

			500	5000	8500
	Explanation	No bias	Bias introduced from 500th observation	Bias introduced from 5000th observation	Bias introduced from 8500th observation
dataset_1	No bias: 0.05 FPR	`dataset_1`	-	-	-
dataset_2	Little bias: 0.075 FPR	-	`dataset_2_500`	`dataset_2_5000`	`dataset_2_8500`
dataset_3	Medium bias: 0.1 FPR	-	`dataset_3_500`	`dataset_3_5000`	`dataset_3_8500`
dataset_4	Large bias: 0.125 FPR	-	`dataset_4_500`	`dataset_4_5000`	`dataset_4_8500`
dataset_5	Sudden medium bias: 0.1 FPR	-	`dataset_5_500`	`dataset_5_5000`	`dataset_5_8500`

In addition, to ensure robustness of findings, each dataset was generated 100 times with different seeds, from seed 1 to seed 100, running each method over each different seed of the dataset. The datasets were named: `dataset_2_500_s1`, … , `dataset_2_500_s100`, and were all stored locally.

Two custom functions, were created in R to generate the datasets: `generate_ground_truth_dataset` and `add_fraudulent_predictions`. The first function produces a dataset with ground truth labels and customer age category, using random sampling. This assumes an equal number of fraudulent and non-fraudulent credit applicants, a decision made for maintaining smaller datasets, thereby enhancing computational efficiency without significantly affecting the methods' outcomes or performance. The second function introduces predictions and alters the FPR for the age groups. It sets the young group's FPR as constant (although there is random variation), while the old group's FPR varies linearly within a specified range, allowing a systematic change in the FPR and the degree, gradualness, and timing of bias.

The different methods employed in this research make different assumptions about the underlying properties of the data. All methods assume that data points are independent of one another, which holds for this research. The biggest difference in the assumptions between the methods lies assuming normality (CUSUM, SPRT-t, SBF) and the methods not making any assumptions about the distribution of the data (EWMA, Shewhart $\bar{X}$). The datasets used per method are tested for normality using the Shapiro-Wilk test and show normality at the 1% significance level, meaning the null hypothesis of normality cannot be rejected at this level. Moreover, according to the central limit theorem binomial distributions also take on normality when $\min(np, n(1-p)) \geq 5$, which holds in all cases (Kwak & Kim, 2017). Inherent to these

methods, is that they perform better or worse under specific conditions. Although not deeply explored within the scope of this study, the potential for improvement in different data scenarios will be touched upon in the discussion of the research limitations.

Take note that the datasets created in this research are not continuous or dynamic in nature as the research question to this study may imply. The methods applied to the datasets handle the data in a continuous manner, mimicking the treatment of continuous data in real-life monitoring scenarios.

3.3.1 Dataset transformations

The thirteen datasets previously mentioned don't fully meet the input requirements for all monitoring methods. Additionally, some of the metrics used as input to these methods, like the FPR, still require computation. Therefore, for each previously mentioned dataset three variations are created.

- SPC: These variations are transformed to count data, compatible with the SPC techniques. Four count variables (`count_old_fraudulent`, `count_old`, `count_young_fraudulent` and `count_young`) are created, aggregated per 50 total observations, resulting in a 200-row dataset.
- nPMI: Also the distinct bias quantification metrics still need to be computed, such as the nPMI (normalized Pointwise Mutual Information) , which is calculated between `customer_age` and `fraudulent`. The nPMI was calculated over 50-observation chunks, providing 200 total rows. The nPMI (as well as the FPR) will be further elaborated upon in the subsequent section.
- FPR: These variations record the FPR difference ($FPR_{Old} - FPR_{Young}$). Again, the FPR was calculated over 50-observation chunks, yielding 200 rows per dataset.

Given thirteen different datasets, four variations and 100 different seeds, a total of 5200 datasets were generated. Appendix E, illustrates the development of the averages of the mean of fraudulent predictions, the FPR and the nPMI across all datasets.

3.4 Measuring metrics

With the discriminatory setting now established, appropriate methods selected, and biased datasets generated, it becomes crucial to select and compute metrics that can quantify, or serve as proxy for, the discriminatory bias. These metrics will then be utilized as input for the monitoring methods.

Different fairness monitoring settings, ask for different fairness monitoring measures. An important split when it comes to selecting measures, as highlighted in the AEQUITAS fairness tree (Saleiro, et al., 2018), is if there is a ground truth label available or not. Having a ground truth label available, within a short degree of time from when the prediction is made, allows for a much more accurate representation of fairness through a single metric. However, very often, such ground truth labels are not available (Buyl &

Bie, 2022). In the case of fraud prediction for credit approval, the ground truth labels are not available for the individuals predicted as 'fraudulent'. The distinction between a true positive and a false positive, can hence never be made. Because there are many fairness-sensitive domains where ground truth values do become available within a brief period of time (eg. online advertising), this research will also investigate monitoring discriminatory bias with the presence of ground truth labels. This research will use three different metrics and input data for testing the different monitoring methods: the mean, the nPMI and the FPR.

3.4.1 Mean

The first metric that is employed for measuring fairness in the absence of a ground truth label is the mean. Changes in the mean(s) of the predictions of fraudulency are used to test for demographic parity. Recall demographic parity implies $P(\hat{Y} = 1|A = 0) = P(\hat{Y} = 1|A = 1)$, where $\hat{Y}$ is the prediction and A is the sensitive attribute. Demographic parity will therefore, in this research, be tested by identifying if $P(fraudulent = 1|customer\_age = old) =$ if $P(fraudulent = 1|customer\_age = young)$. The way the mean is being measured is adapted for the different methods, as each method relies on distinct inputs.

Take note that accepting demographic parity as a proxy for fairness assumes that the means of two different sensitive groups are equal, even though this assumption may often not hold. As this research does not aim to investigate the validity and applicability of different fairness definitions this assumption will not be scrutinized. Nevertheless, these are considerations that should be paid attention to when monitoring in real life scenarios.

3.4.2 Normalized Pointwise Mutual Information

The second metric used to quantify discriminatory bias in the absence of a ground truth label is the nPMI (normalized Pointwise Mutual Information), as suggested by (Aka et al., 2021). The nPMI is a measure of association that compares the likelihood of two events co-occuring together with the likelihood of the same events occurring independently. Hereby encapsulating various aspects of demographic parity and equalized odds.

The Pointwise Mutual Information (PMI) of a pair of outcomes x and y corresponding to discrete random variables X and Y quantifies the difference between the likelihood of their coincidence given their joint distribution and their respective individual distributions, assuming independence. Mathematically the PMI is defined as follows

$$PMI(x;y) \equiv log_2 \frac{p(x,y)}{p(x)p(y)} = log_2 \frac{p(y|x)}{p(y)} \quad (35)$$

and the nPMI then takes on the value

$$nPMI(x;y) \equiv \frac{PMI}{-\log p(x,y)} = \frac{\log[p(x)p(y)]}{\log p(x,y)} - 1. \quad (36)$$

The corresponding outputs of the nPMI are:

- no co-occurences, $\log p(x,y) \to -\infty$, so the nPMI is -1;
- co-occurences at random, $\log p(x,y) = \log [p(x)p(y)]$, so nPMI is 0;
- complete co-occurences, $\log p(x,y) = \log p(x) = \log p(y)$, so nPMI is 1.

In this research y is the prediction of fraudulency and x is the sensitive attribute customer age. A larger nPMI will therefore suggest a stronger association between a customer's age group and the prediction of being fraudulent, signifying discriminatory bias.

Take note that the nPMI has been selected as an additional metric to the mean for measuring discriminatory bias in the absence of a ground truth label to potentially uncover aspects of discriminatory bias that the mean may not capture, and to show how bias can be measured when the assumptions underlying demographic parity do not hold. Following the findings of Aka et al. (2021), the assumption is made that the nPMI serves as an adequate representation of fairness. Nevertheless, it is important to acknowledge that the validity of this assumption is not extensively studied and may not hold true in all cases. This will not be researched in this paper as an in-depth investigation into the most suitable metric(s) is beyond its scope. Additionally, it is important to take note that the SA techniques cannot utilize the nPMI as input as they require data distributions as opposed to singular metrics.

3.4.3 FPR

In the case of measuring fairness in the presence of a ground truth label, the False Positive Rate is chosen, as suggested by the AEQUITAS fairness tree (Saleiro, et al., 2018). In this case, a false positive would have an unfavorable effect on the individual. In order to ensure fairness, we therefore want to ensure that the likelihood of an application being incorrectly labeled as fraudulent is unrelated to the person's sensitive attribute. Accordingly, we measure the ratio between FPRs. The false positive rate represents the proportion of negative instances that were incorrectly classified as positive and is defined as

$$FPR = \frac{FP}{FP+TP}. \quad (37)$$

The FPR threshold (for which both sensitive groups adhere to), is 5%. This metric is typically mandated by clients in the fraud detection domain because it strikes a balance between preventing customer attrition and detecting fraud (Jesus, et al., 2022). Each false positive results in a dissatisfied customer who may want to switch banks after being mistakenly flagged as fraudulent.

Take note that the FPR is employed in two distinct ways throughout this research. First, the FPR is used sequentially, incrementing from the first up until the last observation. This is similar to how the other metrics are being employed. Second, the FPR is employed with a moving window, where only the most recent 1000 observations are used sequentially as input for the methods. This is done to test if the methods perform better when there is more focus on the recent observations and older data points (which may not be representative of the current process) are disregarded.

3.5 Evaluation criteria

After defining the methods, creating scenarios and datasets, and establishing metrics for quantifying discriminatory bias, the subsequent task is to determine which of the monitoring methods performs best. Therefore, three evaluation criteria are defined to measure and compare the performance of the various monitoring methods. For this several criteria will be measured, the ARL (average run length), the flag rate and the flag density.

The ARL is a commonly used measure in SPC theory and it represents the amount of observations, or sets of observations, needed to generate an alert in a control chart. There are two different ARLs that will be used in this research:

$$\text{ARL IC} = \frac{Index\ first\ IC\ flag}{total\ IC\ observations}, \tag{38}$$

$$\text{ARL OC} = \frac{Index\ first\ OC\ flag}{total\ OC\ observations}. \tag{39}$$

The ARL thus represents the average amount of relative time needed before a control chart signals a process shift. Although it is common practice to express ARLs as absolute figures, this research adopts a relative approach to account for the introduction of bias at different timepoints within the datasets. The ARL IC should be maximized, indicating false positive alerts are generated as late as possible, and the ARL OC should be minimized, indicating true positive alerts are generated as soon as possible. The ARL values are averaged separately over IC-flagged datasets and OC-flagged datasets.

Apart from the ARL values, two other metrics are measured to quantify the performance of the methods, of which both are measured for the IC and OC process

$$Flag\ rate = \frac{\#\ of\ cases\ where\ the\ process\ is\ flagged}{\#\ of\ total\ cases}, \tag{40}$$

$$Flag\ density = \frac{Number\ of\ flags\ per\ flagged\ case}{\#\ of\ total\ flagged\ cases}. \tag{41}$$

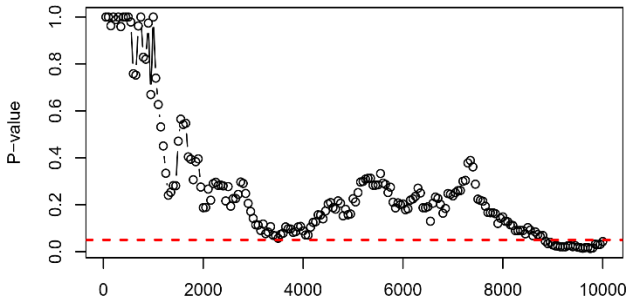
The flag rate represents the amount of times that an IC or OC process has been flagged expressed as a percentage (over the 100 different seeds of the same dataset). The flag density represents the average number of flags or alerts generated by the method per case where alerts have been made. The flag density will only be measured for the SPC methods. This is because an alert in SPC signifies that a process may be out-control and requires investigation, however does not necessarily mean that the process should be stopped. When an alert is made in sequential analysis the test has reached significant evidence to stop the testing process, because H_1 has been accepted. This means that there can only be one alert at maximum in a sequential testing procedure. Rationally, we want to minimize the flag density and flag rate for IC processes ensuring minimal false alerts, and maximize the flag density and flag rate for OC processes ensuring maximal true alerts.

4 Results

The different methods employed throughout this research are designed to gather information to identify and potentially flag discriminatory bias. This evidence can be obtained through methods like SPC, which utilizes alerts to determine if a process is in or out of control, or through sequential hypothesis testing, where acceptance of the alternative hypothesis indicates the presence of the bias. However, why would a traditional statistical test from NHST not work in such a scenario?

To verify the theory, and establish the necessity of the approaches employed in this research, a NHST approach is initially applied to dataset 1, which contains no bias. The test investigates if the false positive predictions hold relation with the customer age category, which is not the case for this dataset. However, when sequentially performing the NHST test, as depicted in figure 2 the p-values progressively decrease until H_1 is accepted. This phenomenon occurs as NHST assumes fixed sample sizes and a singular test. In NHST random variability or noise can lead to small p-values by chance alone. These Type I errors accumulate when successive tests are performed, resulting in the wrongful accepting of H_1. When NHST is sequentially performed over the 100 different seeds of the datasets without bias, almost half of all the datasets are flagged as being discriminatory, indicating that the NHST indeed does not lend itself for monitoring solutions.

Figure 2. NHST for monitoring dataset 1



The rest of this results section will be structured in the following way. First, all methods will be discussed individually, including an explanation of their setup and the rationale behind parameter selection, as well as all individual-level results. Second, comparative plots will be used to compare the results from the different methods. Third, the most effective methods will be selected and combined to demonstrate how efficient these monitoring methods are and how they could be best (collectively) utilized.

4.1 Method-specific results

4.1.1 Shewhart chart

In terms of setting up, the Shewhart chart is very user-friendly as it requires little parameters that need to be specified. Only the center argument needs specification which represents the IC center value of the measurement metrics. This is therefore set to the mean of the datasets containing no bias.

Table 3 presents the results of the Shewhart chart, demonstrating the performance criteria across various biased datasets and input values. Since there are numerous results per method and many show similar trends, only the Shewhart chart's results will be presented in the text. It will serve as an illustrative example of interpreting the study's findings, moreover it will highlight the general trends found in the results across the different methods. For the remaining methods, only the key findings will be presented. Their extensive results can be found in Appendix D.

Table 3 illustrates the three performance criteria, all grouped for IC (in control) and OC (out control) processes. For processes that have no bias, being IC, it is desired to have the least amount of alerts in the least amount of cases, minimal flag density and flag rate respectively. When there is an alert, this alert should be as late in the process as possible, a maximal ARL. For processes that are biased, being OC, the opposite is desired (maximal flag rate & flag density, and minimal ARL). It is important to recognize the tradeoff between IC and OC alerts among methods. Methods with high OC values also tend to have high IC values. As a method becomes more sensitive to OC bias detection, it may over-respond to unrelated noise, leading to poorer IC performance.

The datasets in the table are categorized based on timing of bias introduction: 500, 5000 and 8500, which represent early, middle and late bias introduction respectively. Early bias introduction results in larger OC flag rates because most observations are OC. In the case of late bias, the reverse happens; higher IC alerts are observed due to the majority of datapoints being IC. Further, the datasets, numbered from 1 to 5, differ in terms of bias size and gradualness of introduction. Dataset 1 contains no bias, 2, 3 and 4 contain small, medium and large biases introduced gradually, and dataset 5 contains medium bias introduced abruptly. The detection difficulty increases with smaller bias sizes. As can be seen in the case of the Shewhart chart, and which also holds for the other methods, is that in most cases medium sized abrupt bias (dataset 5) is easier to detect than large gradual bias (dataset 4).

As illustrated in Shewhart charts' results, the flag rates across all metrics are performing quite good in the case of early and middle bias. However, when it comes to late bias the amount of false alarms is seen to significantly increase. Over most cases two to five times as many non-biased cases are flagged than biased cases, thereby compromising the method's reliability. Moreover, the flag rate is 0.6, when using the FPR in the case of no bias, which is very large. This distinguishes the well performing from the poor

performing methods, as a good performing monitoring method should show minimal number of alerts when there is no bias, even when there is no bias for a prolonged period of time.

When comparing the distinct measurement metrics (the mean, nPMI, FPR and moving window FPR) the general trend across all methods is that the FPR stands out as the most reliable, consistently delivering superior results across all tested scenarios. This is also to be expected as it is the most direct representation of bias. Its variant, the FPR moving window, falls slightly short, indicating lower flag densities for the SPC methods and generally providing slightly more inconsistent and unreliable alerts. The mean, despite its simplicity, performs better than the nPMI. The nPMI has higher flag rates and higher degrees of randomness to its results compared to the mean, jeopardizing its overall reliability.

Specifically for the Shewhart chart, this trend holds up to a large extent. The FPR is performing quite good for early and middle timed bias, although for late timed bias in most cases, the IC flag rate surpasses the OC flag rate. The moving window shows worse performance due to a decrease in the OC flag rate and an increase in the IC flag rate.

Upon examining the results of the mean and the nPMI as input measures of the Shewhart chart, low flag rates are found both for IC and OC processes. The flag rate indicates that the mean has a much lower flag rate both IC and OC than the nPMI. In cases of late bias, the IC flag rates are three to fifteen times higher than the OC flag rates when using the mean as measurement.

The ARLs across all input metrics, for the Shewhart chart, are generally poor, although they appear to be the best when solely relying on the last 1000 observations. Furthermore, flag densities remain consistently low across all metrics, aligning with the theory that the Shewhart chart tends to have limited effectiveness in signaling smaller process shifts.

Table 3. Shewhart chart results

	Mean						nPMI					
	Flag rate		*ARL*		*Flag density*		*Flag rate*		*ARL*		*Flag Density*	
	IC	**OC**	**IC**	**OC**	**IC**	**OC**	**IC**	**OC**	**IC**	**OC**	**IC**	**OC**
ds_2_500	-	0.23	-	0.49	-	0.01	0.02	0.41	0.65	0.44	0.10	0.01
ds_3_500	-	0.34	-	0.55	-	0.01	0.02	0.47	0.65	0.46	0.10	0.01
ds_4_500	-	0.41	-	0.56	-	0.01	0.02	0.58	0.65	0.49	0.10	0.01
ds_5_500	0.01	0.32	0.90	0.54	0.10	0.01	0.04	0.52	0.73	0.45	0.10	0.01
ds_2_5000	0.10	0.17	0.45	0.61	0.01	0.01	0.29	0.25	0.56	0.48	0.01	0.01
ds_3_5000	0.10	0.18	0.45	0.63	0.01	0.01	0.28	0.27	0.56	0.51	0.01	0.01
ds_4_5000	0.10	0.23	0.45	0.65	0.01	0.01	0.30	0.31	0.54	0.51	0.01	0.01
ds_5_5000	0.12	0.21	0.52	0.47	0.01	0.01	0.28	0.28	0.49	0.57	0.01	0.01
ds_2_8500	0.15	0.01	0.40	0.80	0.01	0.03	0.38	0.07	0.43	0.53	0.01	0.03
ds_3_8500	0.15	0.02	0.40	0.65	0.01	0.03	0.40	0.07	0.42	0.53	0.01	0.03
ds_4_8500	0.15	0.05	0.40	0.71	0.01	0.03	0.38	0.07	0.42	0.53	0.01	0.03
ds_5_8500	0.16	0.04	0.49	0.53	0.01	0.03	0.42	0.10	0.41	0.57	0.01	0.03
ds_1	0.13	-	0.60	-	0.01	-	0.45	-	0.43	-	0.01	-

	FPR						FPR – Last 1000 obs.					
	Flag rate		*ARL*		*Flag density*		*Flag rate*		*ARL*		*Flag Density*	
	IC	**OC**	**IC**	**OC**	**IC**	**OC**	**IC**	**OC**	**IC**	**OC**	**IC**	**OC**
ds_2_500	0.04	0.8	0.45	0.52	0.1	0.01	0.04	0.81	0.32	0.39	0.12	0.01
ds_3_500	0.04	0.84	0.45	0.51	0.1	0.01	0.04	0.9	0.32	0.38	0.12	0.01
ds_4_500	0.03	0.93	0.33	0.48	0.1	0.02	0.04	0.95	0.32	0.37	0.12	0.02
ds_5_500	0.04	0.93	0.6	0.41	0.1	0.01	0.05	0.94	0.76	0.27	0.1	0.02
ds_2_5000	0.38	0.57	0.48	0.59	0.01	0.01	0.51	0.53	0.42	0.43	0.01	0.01
ds_3_5000	0.36	0.72	0.48	0.59	0.01	0.02	0.51	0.62	0.42	0.43	0.01	0.02
ds_4_5000	0.31	0.85	0.5	0.62	0.01	0.02	0.51	0.68	0.42	0.45	0.01	0.02
ds_5_5000	0.29	0.79	0.51	0.5	0.01	0.02	0.5	0.68	0.48	0.39	0.01	0.02
ds_2_8500	0.6	0.21	0.38	0.58	0.01	0.04	0.68	0.1	0.41	0.32	0.01	0.03
ds_3_8500	0.59	0.32	0.37	0.57	0.01	0.04	0.68	0.13	0.41	0.31	0.01	0.04
ds_4_8500	0.58	0.48	0.36	0.62	0.01	0.05	0.68	0.18	0.41	0.34	0.01	0.04
ds_5_8500	0.45	0.52	0.39	0.56	0.01	0.04	0.56	0.25	0.36	0.41	0.01	0.03
ds_1	0.6	-	0.45	-	0.01	-	0.6	-	0.38	-	0.01	-

4.1.2 CUSUM

When setting up the CUSUM chart there are two parameters that hold tuning relevance. The first parameter is the decision interval, representing h. Recall that the CUSUM chart is found to function optimally when h takes on the value of 4 or 5. Generally, increasing h reduces the sensitivity of the control chart to process changes, thereby increasing flag rates and flag densities and decreasing ARLs, for both the IC and OC processes. For this research h is set to 4, striking a suitable balance between bias detection and avoiding an excessive number of false alarms. The second parameter that can be adjusted is the `center` argument, which represents the mean or center line used in the CUSUM computation. Likewise to the Shewhart chart this value can be set to the mean value of an IC process.

Although, all general trends described in the Shewhart results also hold for the CUSUM there are few other notable findings. The CUSUM performs better than the Shewhart chart across almost all scenarios, with generally slightly lower flag rates both IC and OC, however the method is more reliable in terms of alerts. The FPR shows quite good performance for the CUSUM, although there still is still an IC flag rate of 0.44 in the case of no bias, which is very large. The most notable finding however is the large difference in flag densities, where the OC flag densities are generally multiple times larger than the IC flag densities, especially in the case of the FPR. This presents an opportunity to possibly distinguish IC from OC alerts based on the number of, or closeness of these alerts. The full results of the CUSUM can be observed in Appendix D.

4.1.3 EWMA

When setting up the EWMA, the first parameter that needs adjustment is the `center` argument, which serves the same purpose as and is set likewise to the CUSUM chart. The second tuning parameter is λ, where $0 \leq \lambda \leq 1$, determining the depth of memory of the EWMA. A larger λ becomes more effective when the available data serves as a more precise indicator of bias. This makes intuitive sense, as when the available data is a poor representation of discriminatory bias, there will be a higher degree of unsystematic error or random noise in the data. Making it less effective to heavily rely on current data alone as it may not accurately reflect the actual bias. The λ was set to 0.5 for the mean and nPMI EWMA chart, and set to 0.9 for the FPR EWMA chart. In the case of the FPR moving window the λ was decreased to 0.5 again. This was done because focusing solely on the last 1000 observations inherently disregards old data.

The EWMA shows a poor performance over all metrics, with generally very high IC and OC flag rates and low flag densities, making the method unreliable. The ARLs are slightly lower compared to other methods, however this goes for both IC as OC alerts, rendering it of little value. Using a moving window significantly decreases performance compared to the regular FPR. This is because the EWMA starts reflagging observations which were not flagged in previous iterations, thereby amplifying the

oversensitivity of the chart in a negative way. Since the EWMA already places more emphasis on new data by adjusting λ, it proves ineffective to mimic this effect manually through a moving window.

4.1.4 SPRT-t

To tune the SPRT-t, there are three parameters that need adjustment, the power, α and Cohen's d. Cohen's d represents the true standardized difference of means of the populations underlying two groups (Schnuerch & Erdfelder, 2020). Because the difference in means will be minor in the case of detecting discriminatory bias (in the case of this research), a d of 0.01 is selected for the mean and the FPR SPRT-t test. In the moving window scenario, d is increased to 0.05, suggesting a greater expected mean difference and increasing the method's sensitivity to larger shifts. On the one hand the method is able to detect substantial changes with less data, which is beneficial as there is less data available. On the other hand, the method becomes less likely to pick up on smaller changes, which is beneficial, as the method will less likely falsely flag variance and noise which make up a larger relative portion of the dataset.

The power parameter, $1 - \beta$, is the probability of correctly rejecting the null hypothesis when it is false. It indicates the test's ability to accurately reject the null hypothesis, thereby reducing the likelihood of committing a Type II error. When the true effect size is smaller, the power argument needs to be reduced. In this research, the power argument is set to 0.1 for the mean and FPR, and slightly higher for the moving window, at 0.3. The higher value for the moving window accommodates its increased relative noise and variance, requiring a stronger effect size to accept the alternative hypothesis.

The last parameter, α, affects the likelihood of truthfully accepting H_1. Reducing α can make it more difficult to truthfully accept H_1 by increasing the likelihood of a Type II error, whereas increasing α can make it easier to accept H_1 but also increase the likelihood of a Type I error. α is set to its default value of 0.05.

The SPRT-t shows generally lower flag rates both IC and OC than the SPC methods. The relative difference between OC and IC flag rates increases substantially compared to the SPC methods, rendering it a more reliable method. Especially with the FPR the SPRT-t shows very reliable performance with low IC flag rates. Although there are still cases where IC processes are being flagged (dataset 1 has an IC flag rate of 0.21), over every single tested biased scenario more cases that are OC are being flagged than cases that are IC. The ARLs perform quite good in the case of the moving window, but in the other cases the bias is picked up on later than in the cases of SPC. This is not surprising, as SA methods generate only one alert, whereas SPC methods can generate numerous alerts, making it more likely for the first alert to be earlier in the process (for SPC).

The last notable result is the significantly worse performance of the moving window, a larger decline in performance than compared to the SPC methods. This can be explained by several factors. First,

because there is less data at each step, there is less statistical power, meaning the test will be less likely to detect an effect when it is there, decreasing the OC flag rate. Second, less data increases the likelihood of random variability appearing as a significant effect, leading to increased Type I error rates. Third, the sequential nature of the SPRT-t test might prematurely stop the test in the moving window scenario based on smaller, possibly unrepresentative, data samples. Last, if the bias is not consistently present in each window this can result in decreased true flags and increased false flags due to the varying representations of the underlying data distribution.

4.1.5 Sequential Bayesian Factor

The SBF has two parameters that are important to properly set before performing the tests, `prior.r` and `prior.loc`. Both parameters specify information about the prior beliefs, in particular, with regards to the Cauchy distribution. `prior.r` specifies the scale of the Cauchy distributed function, where larger values signify a prior belief that large effects are more likely and a smaller value indicate that large effects are less likely. For this research this value is set to 0.05 when testing on both the FPR scenarios and to 0.01 when testing on the mean. The r is set to a smaller value for testing on the mean as the discrimination is not as pronounced as in the FPR cases, thus the prior belief of the effect size is smaller.

`prior.loc` specifies the location of the Cauchy distributed prior function, or the center of the peak of the distribution. When there is a prior belief that there is discrimination, the prior location can be set to a positive value, reflecting this initial belief. However, if there is no prior belief about the expected effect size, i.e. the most probable value for the effect size is zero, the prior location should be set to zero. In the case of monitoring discriminatory bias, assuming that the model being deployed shows no discriminatory bias, a prior location of zero should be chosen. If a prior location of zero is chosen and at a given point the data suggests an effect size different from the prior location, the data will 'override' the prior if the evidence is strong enough.

Important to note before discussing the results from the SBF is that five runs were performed per dataset as opposed to 100 as with the other methods. This was due to the computational complexity of the SBF procedure[2]. Therefore keep in mind that the results may be less generalizable and robust than the results discussed before.

What stands out about the results of the SBF is the scarce IC alerts compared to the previously discussed methods. Although the mean shows somewhat inconsistent performance, the FPR (and moving window) are the first instances across all methods to show no IC alerts. Noteworthy is that none of the late

[2] Each run took 20-50 minutes. Consider a total of 13 datasets, 2 measurement metrics & a moving window and several seeds per dataset.

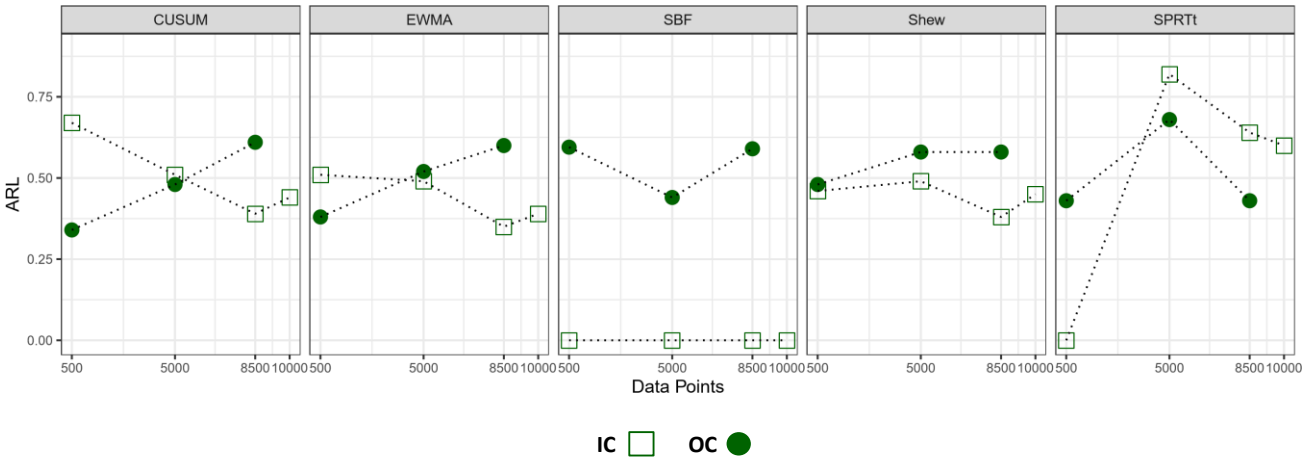
biased setting are being flagged, however when further analyzing the SBF plots, it shows that the BF approaches the threshold of 3, suggesting the method is detecting the bias correctly, though insufficient evidence has been gathered to generate an alert. Overall, the findings suggest that the SBF is a reliable method for detecting discriminatory bias, with a very low rate of false alerts. Moreover, the continuous updating of the BF makes it a useful indicator of existing evidence of discriminatory bias. Users could potentially use the method to assess the risk of discrimination without necessarily needing a definitive decision for there to be discrimination yes, or no. This characteristic makes it a practical tool for real-world applications.

4.2 Comparative analysis

Having individually discussed each method's results, a comparative analysis is performed. The remaining results, including the comparative analysis, focuses solely on scenarios where the FPR is the input without moving window. The reason the FPR is chosen, over the mean or nPMI, is because it is a direct representation of discriminatory bias. If a monitoring method overlooks bias, it implies a deficiency in the method itself, not the measurement metric's inability to capture the bias. Consequently, using the FPR as the input metric provides a more accurate comparison across the different methods.

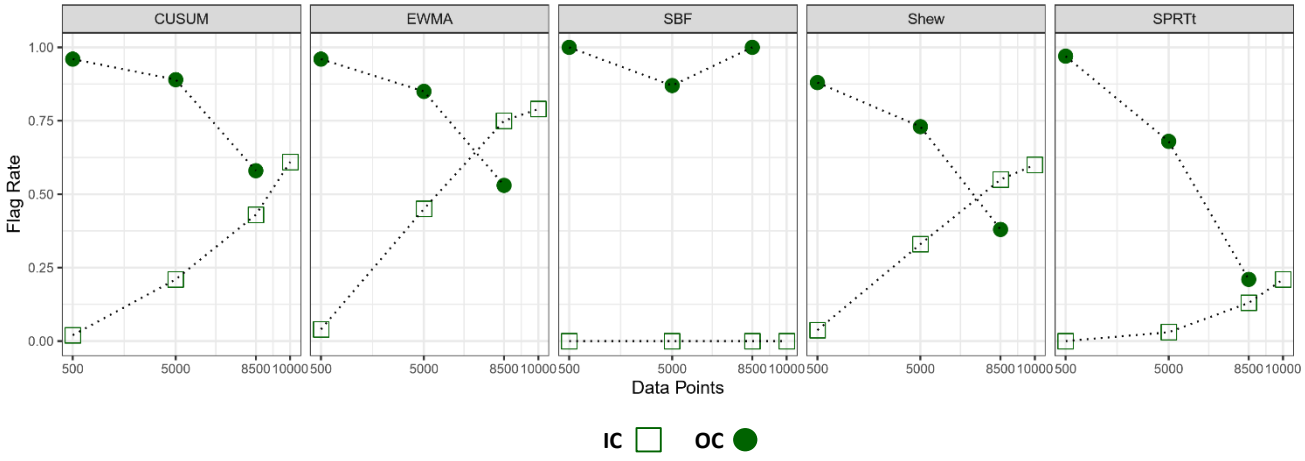
First, the ARLs are compared across the five different methods, as illustrated in figure 3. The figure illustrates the average values over all datasets summarized by early, middle and late bias. The 10.000th data point corresponds to dataset 1, the dataset containing no bias at all. Observing the early bias, most methods correctly identify bias at around the same time, except for SBF, which detects it slightly later. However, when considering the disparity between IC and OC ARLs the CUSUM, SPRT-t and SBF show the best performance showcasing large disparities. For biases introduced in the middle of the dataset, all methods exhibit similar behavior, detecting IC bias approximately at the same time as OC bias, except for the SBF, which does not flag any IC bias. Considering late timed bias, the SPRT-t method proves to be efficient by on average flagging OC bias earlier than IC bias. When examining the datasets without bias SPRT-t and SBF showcase the best performance while all the SPC methods produce comparable results.

Figure 3. Comparative analysis ARL values



The flag rates across all different methods are depicted in figure 4. In the case of early bias, it is clear that all methods exhibit similarly high OC flag rates and low IC flag rates. Moving on to middle timed bias, the flag rates show a considerable decrease, particularly for the SPRT-t method, while the IC flag rate remains relatively unchanged. Among the SPC methods, there is a smaller reduction in OC flag rates, but a substantial increase in IC flag rates, with the CUSUM method demonstrating the best performance in this regard. As for late timed bias, both the EWMA and Shewhart chart display poor performance, with larger IC flag rates compared to OC flag rates. Likewise in the absence of bias, the Shewhart chart and EWMA showcase the worst performance.

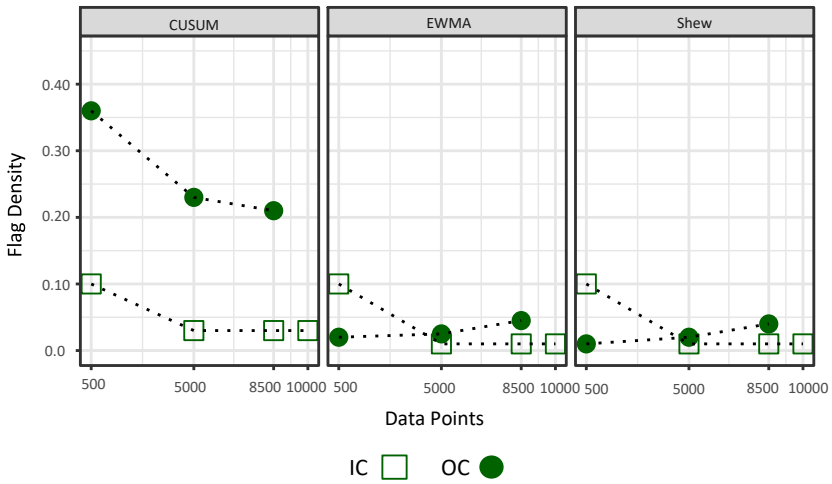
Figure 4. Comparative analysis flag rates



When focusing on flag density, as shown in figure 5, only the SPC methods can be analyzed. It is evident that the CUSUM exhibits the largest flag densities across all scenarios. What's especially

noteworthy is the difference between the IC and OC density, which is very large for the CUSUM. This could allow users to distinguish IC from OC alerts based on the flag density. On the other hand, the EWMA and Shewhart chart do not demonstrate this characteristic, exhibiting small differences between IC and OC flag densities.

Figure 5. Comparative analysis flag densities



4.3 Consolidating best practices

Based on the previous two sections it is evident that there are three methods showcasing the best performance across all different scenarios: the CUSUM, SPRT-t and SBF. For the remainder of this research these three methods will be considered and further investigated. In addition it will be considered how these methods can be combined to create a reliable and effective monitoring system.

4.3.1 Fine-tuning the CUSUM

The reliability of alerts generated by the CUSUM is noticeably lower compared to the other two methods. This may raise concerns about the applicability of the CUSUM method, as it could be consistently overshadowed by the SPRT-t or the SBF. However, it is worth noting the significant divergence between the flag densities of IC and OC alerts. This suggests the potential to establish an alert threshold that effectively distinguishes correct alerts from false alerts. Such a threshold could significantly reduce the occurrence of false alerts, while maintaining the relatively high OC flag rate and maintaining satisfactory OC ARL performance.

To establish this threshold first all cases where alerts had been made are summarized for the number of IC and OC alerts and the distance between them. Using these values, distributions were mapped for the IC and OC processes. As expected, the average number of alerts was more than eight times larger for OC

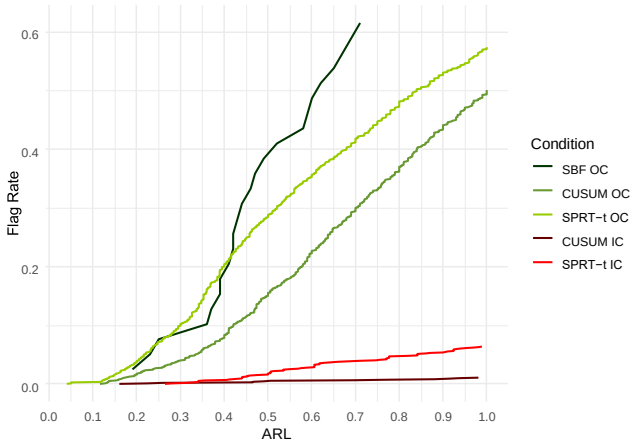
processes than IC. As well as the distances between these alerts were on average 3.3 times smaller for OC processes as IC processes. Using these distributions, as illustrated in Appendix F, a threshold was set of 14 total alerts, indicating that the 14th alert signifies the presence of discriminatory bias. By setting this threshold almost 80% of all false positives have been removed, whilst only removing 6% of true positives. Although the threshold is set in hindsight, in reality it is not as straightforward to choose an optimal value, it does signify the potential of the CUSUM to efficiently detect discriminatory bias with high OC flag rates and low IC flag rates. In spite of the OC ARL declining in performance slightly (as the 14th alert indicates bias as opposed to the first alert the ARL will automatically drop), it does improve the method's usability by a large extent. The remainder of the results will assume the CUSUM to function with this alert threshold.

4.3.2 Contrasting and combining methods

After having established a threshold alert frequency for the CUSUM, the flagging effectiveness of the three best performing methods can be analyzed. Figure 6 illustrates how many cases containing discriminatory bias are being alerted over the run length of the bias, in other words, the flag rate is being plotted against the ARL, for both IC and OC processes.

It appears that the IC rates are negligibly low for most methods, only the SPRT-t has an IC flag rate approaching 0.08. Despite having the least robust results from an experimental design perspective, the SBF proves to be the most efficient method, signaling bias detection very early with high flag rates. Further, the SPRT-t seems to outperform the CUSUM slightly, however this comes with a larger number of false flags.

Figure 6. ARL vs Flag Rate top – 3 methods



The key message this plot communicates is that all three methods can, with adequately low likelihood of false alerts, detect discriminatory bias in a substantial share of the cases. Even though there

are many undetected instances by these methods, their collective use can enable the detection of the majority of biased causes. To put this to the test the alerts generated by both the CUSUM and the SPRT-t are combined, to explore their combined flagging effectiveness. Note that the SBF was not included because it had not been applied on all 100 different seeds. Collectively, they were able to flag about 0.63 of the fraudulent cases. Notably, many of the cases flagged as fraudulent by the CUSUM were also identified by the SPRT-t, however when combined there was still a 10% increase in number of correctly flagged total cases.

Together, these methods were able to detect the large majority of the cases. What's more, of the undetected cases two thirds were datasets containing late bias, meaning that the bias would likely still be picked up on at a later point in time. Furthermore, among the undetected cases, two-thirds were datasets with late bias, indicating a high probability of detecting the bias at a later stage. The remaining non-detected datasets were mostly datasets containing small amounts of bias (dataset 2). What's more, these results disregard the detections made by the SBF, which, according to the findings, should significantly boost the number of correct alerts. An overview of the combined flagged datasets is illustrated in Appendix G.

To summarize the key performance differences among the three methods, the following list outlines their main distinctions:

- The CUSUM, taken without alert threshold, excels in terms of OC ARL, however generates a relatively large number of false alerts.
- Once an alert threshold is derived for the CUSUM, the OC ARL and OC flag rate drop, although given the large decrease in IC alerts, the CUSUM performance increases significantly.
- All three methods have similar flag rates, except in the case of late bias the SBF exhibits significantly better performance.
- As the bias size increases, the SPRT-t method shows significant improvement, the other methods also show improvement, however to a lesser extent.
- Despite being the least robust, the SBF stands out as the most reliable method with minimal false alerts.

4.3.3 Creating a monitoring system

Analyzing the results of this research, it's clear that identifying discriminatory bias isn't a simple procedure. Even within the boundaries of a well-structured study, with controlled settings where the introduction of bias is clear and deliberate, consistent detection proves challenging across various scenarios. When considering real-world monitoring settings, the multitude of influencing factors will only increase

the complexity of the challenge. Due to the wide array of intricate factors at play, it becomes difficult to definitively point out the existence of discriminatory bias. This study facilitated concluding with certainty whether discriminatory bias was present or not, however such black-and-white determinations may not hold up in real-world settings. Especially considering scenarios where ground truth labels are absent, even after flagging, it can become a challenging task to verify the validity of an alert. Consequently, it's unlikely that an effective solution to monitoring discriminatory bias would involve monitoring a system to flag bias with a binary yes-or-no response to discriminatory bias.

The solution this research presents relates back to the 'stoplight dashboard', which proposes monitoring bias using red, orange and green colors to indicate the risk for discriminatory bias, as introduced in the research by Koshiyama et al. (2021). Utilizing this notion a multi-faceted approach can be created that integrates various indicators from the CUSUM, SPRT-t and SBF, enabling a nuanced and dynamic system to estimate the potential presence of bias.

In a scenario where a deployed ML model is monitored in real-time, each method should serve as a unique, real time indicator of the potential risk for discriminatory bias. Rather than treating each stream of data independently as a binary indicator of discriminatory bias, these would be combined into a singular, shared discriminatory risk factor. The composite risk factor would incorporate all nuanced insights, providing a more comprehensive view of potential bias. Such an approach would also reduce the risk of making Type I errors, as a Type I error produced by one method can be balanced out by another method not making this error.

To demonstrate how the methods could be combined in a dynamic setting all three methods were combined on one of the middle timed, large biased datasets (dataset_4_5000). The outputs of the different methods are illustrated in figure 7 and a discriminatory risk factor monitor plot is included tracking the composite risk factor. Keep in mind that the risk factor scores illustrated in the figure are presented for demonstrative purposes and are not intended to represent definitive guidelines for setting the 'risk factors'.

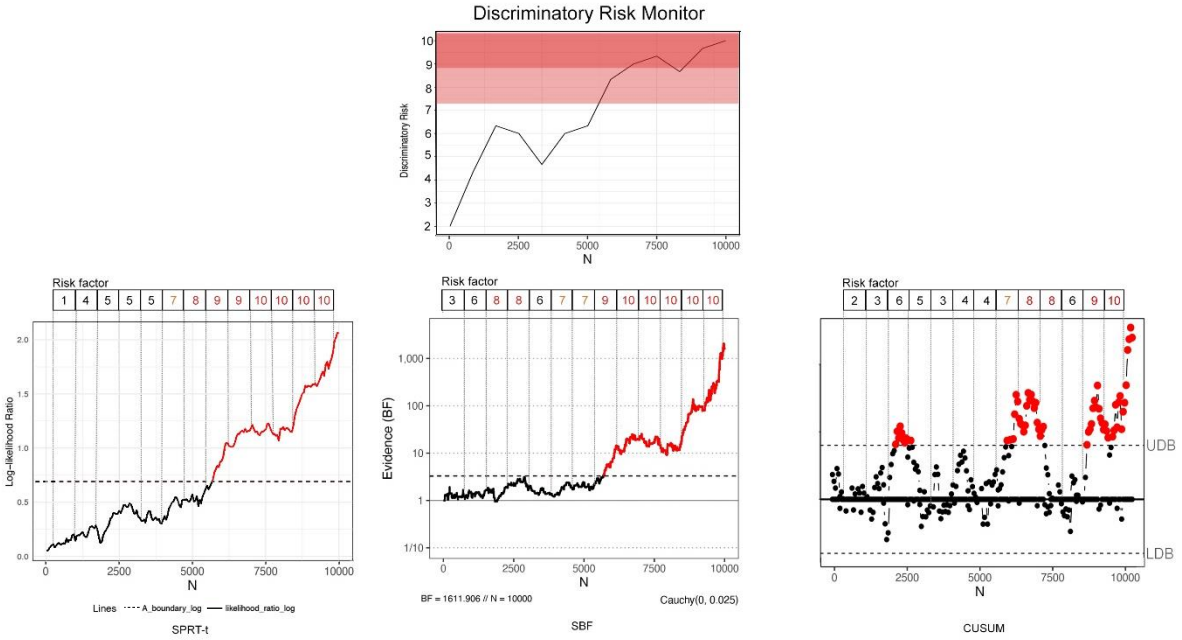
Observing figure 7, the SPRT-t functions quite good with no false positives, picking up on the bias rather soon. The SBF on the other hand, is on the fringe of falsely flagging for discriminatory bias around the 2500th observation, falling just short of the threshold value. The risk factor therefore increases to indicate a higher level of risk according to the SBF. The benefit of this collaborative approach is that as the other methods show no significant signs pointing towards a risk for bias, the overall risk score only increases slightly. The same goes for the CUSUM, which shows a set of OC alerts slightly later on. At the same time, when all three methods start to pick up on the bias, the overall discriminatory risk increases, indicating a high risk of present discriminatory bias.

The advantage of this approach is hence twofold. First, by combining the different methods and aggregating their respective 'evidence' towards discriminatory bias into a risk score prevents the accepting

of false alarms. Second, when bias is indeed picked up on by the methods, there is much stronger evidence to definitively accept that there is indeed discriminatory bias and that it is not merely a false alarm.

This approach is very suitable for implementation in practical settings. Real life monitoring scenarios have much higher levels of noise and variance, meaning the likelihood of making Type I errors will significantly increase. This approach minimizes this risk. Moreover, treating discriminatory bias as a risk factor, rather than a binary label, enables users to proactively identify and address its emergence. This approach allows for early detection and prevention, minimizing its potential negative impact.

Figure 7. Comprehensive discriminatory risk monitoring system



5 Discussion

This research introduced novel applications of monitoring methods by applying them for discriminatory biased settings of ML models. It was found that not every technique is effective at detecting bias in ML models: the Shewhart chart and EWMA were inconsistent across different biased situations, often resulting in numerous false positives and overall unreliability. Conversely, the CUSUM, SPRT-t and SBF did prove to be efficient in monitoring discriminatory bias. The past literature coincides with the finding of the CUSUM being the most effective of the SPC methods. As Neuburger et al. (2017) pointed out, the CUSUM is the best in detecting small absolute changes, as well as large absolute changes in processes. However, the poor performance of the EWMA in this study does not seem to be explained by the covered literature. The Shewhart charts' inefficiency is better understood, with Neuburger et al. (2017) suggesting its efficacy to be limited to specific contexts.

While the study was able to delineate effective bias monitoring methods, it also questioned the reliability of certain measures cited in previous research. Specifically, the nPMI was found to be less reliable than expected in quantifying discriminatory bias when ground truth values are absent, thus challenging the findings made by Aka et al. (2021). This discrepancy, as has also been stated by the researchers, can be attributed to the lack of generalizability of the findings from Aka et al. (2021). Once again, this highlights the high context dependency of ML fairness research, underscoring that fairness ML is a multifaceted field. This complexity is further enunciated by the findings expressed by Wilson et al. (2021), stating that 'fairness is a performance criterion' and Koyishama et al. (2021) whom concede the necessity of actively monitoring and auditing for ML model fairness.

A factor which could have negatively influenced the monitoring performance of the methods is the relatively small dataset size of 10,000 observations. Small datasets make it challenging to achieve statistical significance. This could have also worsened the performance of the mean and nPMI. Because these metrics don't serve as direct representation of what is being monitored for, the larger noise inherent to smaller datasets can lead to misleading results. The worse performing results of the moving window methods can also be tied to the same causes. On the one hand, there is more (relative) statistical noise (random variation) making it difficult to determine if detected changes are due to process changes or random variation. Moreover, the smaller windows possibly lack the statistical power to detect significant effects. What's more, given the scenarios where the methods were performing poorly, either small or late bias, the absolute shifts were rather minimal. Keep in mind that in some instances, the absolute difference in counts of OC false positives between the two groups across the entire dataset was less than 10 observations[3]. This limited number of cases has limited statistical power making it difficult to truthfully pick up on the effect.

[3] For late biased datasets with small bias: 1500 · 0.5 (old ratio) · 0.5 (fraudulent ratio) · 0.025 (bias ratio)

The choice of the chunk size for data accumulation may have also influenced the sub-optimally performing Shewhart chart and EWMA. The detection ability of these charts depends on the relative magnitude of the process shift to the process noise and the frequency of these shifts relative to the chunk size. When the bias had been increased in each dataset to its maximum amount, per chunk there were on average only 0.3, 0.6 and 1.25 more false positive fraudulent elderly individuals for the small, medium and large biased cases respectively[4]. Detecting these shifts can become quite difficult for these control charts, since these charts effectively detect patterns above the noise level. Especially the EWMA can struggle in this respect because it focusses more on recent information, making it more sensitive to recent changes, thus making it more noise sensitive. The CUSUM on the other hand focusses on cumulative trends over time, making it inherently less sensitive to random noise or variation in the data, allowing it to overlook random fluctuation focusing more on persistent changes in the mean (Montgomery, 2012). The same argument can be applied to the nPMI performing below expectation, where the limited amount of information per chunk may have hindered the accurate computation of the metric.

There were also methodological assumptions and constraints negatively impacting the results. The assumptions underlying the methods were met up to an agreeable extent over all scenarios, however, the methods may very well have been monitoring at sub-optimal level due to imperfect input data. Furthermore, there was minimal robustness of the SBFs results due to the limited number of runs, which in turn compromised the validity of these findings.

Lastly, the tuning process of the different methods posed a methodological challenge, as the method's results, especially for the SA methods, were contingent on the pre-determined parameters. This implies that the methods used may have shown performance standards that may be difficult to ascertain in reality. Despite all parameters being based on readily available theoretical knowledge underlying the methods and purposefully not tuning the methods to perfection, there was an opportunity to experiment with parameters to understand how varying values correspond to different outputs and accordingly choose better performing parameters. This may have shown an optimistically painted picture of reality. A similar argument can be made for setting the threshold value of the number of alerts for the CUSUM plot. Although the main driver for setting this threshold was to demonstrate the potential effectivity of the method in practice, it possibly embellished its results.

[4] 25 elderly per chunk · 0.5 (fraudulent ratio) · process change (0.025 ∨ 0.05 ∨ 0.075)

6 Conclusion

In conclusion, this research evaluated distinct monitoring methods for monitoring discriminatory bias in ML models. The five different methods were tested over a variety of thirteen differently biased scenarios, each one replicated 100 times. The Shewhart chart and the EWMA showed poor overall performance with high IC flag rates. The remaining three methods, the CUSUM, SPRT-t and SBF, showed promising outcomes, with large OC flag rates in most scenarios and minimal IC flag rates for the SPRT-t and SBF. Only the CUSUM exhibited high IC flag rates. However, thanks to the large discrepancy between IC and OC flag densities, it is possible in most scenarios to distinguish these two types of alerts from one another. When comparing the three best performing methods on the FPR, the SBF outperformed the CUSUM and the SPRT-t, although with the least robust results. The CUSUM and SPRT-t showed comparable performance.

Upon examining the results of different input metrics, the mean and nPMI performed poorly compared to the FPR and the moving window FPR. This outcome is not surprising as both these metrics measure bias without a ground truth, thereby rendering them less reliable bias representations than when ground truth values are available. Between the two metrics, the mean outperformed the nPMI in most cases, contradicting previous research outcomes in this field. This disparity emphasizes the context dependency of measurement metrics' efficacy. When comparing the moving window FPR to the FPR, the former was found to be less effective in most scenarios, usually reducing both IC and OC flag rates and densities. Although a moving window could potentially be beneficial in real-life monitoring scenarios, the size of the window may not have been optimal in the case of this study.

The aim of this study was to investigate if discriminatory bias can be monitored effectively on deployed ML models, with continuous streams of incoming data. Past studies have indicated that fairness cannot be confined to a single, universally applicable notion; instead, it is deeply reliant on context, drawing from a wide array of sources, definitions and measures. Through testing various monitoring techniques this research suggests that three methods – the CUSUM, SPRT-t and SBF – are particularly effective. Additionally, this research advocates for a comprehensive monitoring approach enabling a nuanced and robust fairness monitoring system. Building on the understanding that fairness is a convoluted concept, such an approach moves beyond treating discriminatory bias as a binary classification task. Instead, it treats bias as a composite risk factor, incorporating results from different monitoring methods to enable early detection and intervention, thus reducing its potential harm. When implemented correctly, the proposed system enables early bias detection, reduces Type I errors, and provides robust evidence of bias, making it a suitable solution for real-world settings.

These findings narrow the research gap of fair ML monitoring, being the first research to investigate and propose solutions for monitoring fairness in practical settings. This monitoring approach acknowledges

the prevailing reality in ML model building where the optimization of model performance overshadows the consideration for fairness. For marketers this study provides tools to enhance fairness and integrity in marketing practices, helping to maintain trust and positive relationships with diverse customer groups, elements crucial to successful marketing strategies in a data-driven world. However, this research also fills a gap for the broader ML community, as it provides guidelines for ethical practices in the rapidly evolving landscape of ML. Given the complexity and challenges of detecting bias, particularly in less controlled real-world scenarios, this research ultimately underlines the pressing need for further research and exploration of fairness monitoring in ML.

6.1 Limitations and future research

The nascent domain of ML fairness research still has many unexplored topics and unanswered questions that need further investigation. Although this study was a first attempt into uncovering insights into the monitoring of fairness in ML models it is important to acknowledge its limitations and the opportunities for future research to further enhance the understanding and effectivity for monitoring in real-life contexts.

First, the study's focus on measuring fairness in the presence of ground truth labels limits the applicability of the monitoring system. Many monitoring scenarios don't have timely access to the ground truth labels and are therefore reliant on other metrics, such as for example the mean or nPMI. While this study investigated efficient monitoring methods and touched upon various metrics, further research is necessary to identify the most effective non-ground truth metrics for different scenarios. This research is crucial to develop a monitoring system applicable in diverse contexts, including those without access to ground truth labels.

The second challenge this research faced is the choosing of parameter values when setting up the methods. As there was access to future data there was a possibility to experiment with different values, which in reality is more difficult. Two insightful opportunities for future research arise. To begin with, research should be conducted on how different fairness monitoring scenarios, considering different data and metrics, call for different parameters. In addition, further research should focus on how these methods can be optimized when only IC (test) data is accessible, aiming to fine-tune the methods within this constraint. This would focus on effectively adapting the methods' parameters given limited (unbiased) data. As new data would become available post-deployment the research could further investigate incremental refinement of the parameters. These insights would ensure robust and effective use of the methods in real-world applications.

The third limitation this study posed is its inability to construct guidelines or rules for the proposed method-combining monitoring approach. The absence of these guiding principles also precluded a robust

evaluation of the system's effectiveness. Hence, future research should investigate how the bias risk score can be computed in the most efficient manner by researching which methods are more reliable under which conditions. Additionally, this could look into other potential factors that can influence model fairness indirectly and factor these into the calculation of the risk score, such as varying socio-economic or demographic factors, model performance metrics or confidence intervals. Understanding and representing the complex interplay between these different factors and the risk for discriminatory bias could create a better representation.

Fourth, this research posed limitations related to the scope of bias testing. Although the study attempted to mimic a wide array of different biased scenarios, many different definitions and forms of bias have not been tested on, thereby limiting the general applicability of the insights of this research. Future research should focus on monitoring with other definitions of fairness, possibly more complex definitions such as individual or counterfactual fairness. Moreover, future research could look into using multivariate control charts, which allow for measuring the process of several variables thereby taking into account the correlation structure among them. Applying this for multiple (non-ground truth) fairness metrics, could better capture the underlying complexities, and monitor their trade-offs more effectively.

Last, this research has limited external validity due to the high degree of experimental control. Future research should verify these findings on real life biased datasets, possibly transforming such datasets in order to test over differently biased real-life scenarios. The robustness of the findings in this paper could then be tested for factors as data structure, noisiness, variance or class imbalance.

6.2 Closing note

In conclusion, monitoring ML fairness is no clear cut task – it is a complex challenge that mirrors the intricacy of the models it seeks to oversee. Given the increasing influence of ML in society, the potential of such systems to amplify biases is a serious risk, underscoring the urgency of robust monitoring practices. It is difficult to state that a model is ‘fair’ or ‘unfair’ in the same sense that it is difficult to state that a model is ‘accurate’. In the words of Wilson et al. (2021): ‘Fairness is a performance criterion’, and therefore, it should be treated, and monitored, as one.

7 References

Adebayo, J. (2016). FairML : ToolBox for diagnosing bias in predictive modeling. *Massachusetts Institute of Technology*.

Aka, O., Burke, K., Bäuerle, A., Greer, C., & Mitchell, M. (2021). Measuring Model Biases in the Absence of Ground Truth. *Association for Computing Machinery*.

Armitage, P., McPherson, C. K., & Rowe, B. C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society: Series A (General)*.

Bantilan, N. (2018). Themis-ml: A fairness-aware machine learning interface for end-to-end discrimination discovery and mitigation. *Journal of Technology in Human Services*.

Barocas, & Selbst. (2016). Big Data's Disparate Impact . *California Law Review*.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2018). Fairness in Criminal Justice Risk Assessments: The State of the Art. *SAGE journals*.

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in. *Proceedings of Machine Learning Research*.

Buyl, M., & Bie, T. D. (2022). Inherent Limitations of AI Fairness. *ARXIV:* 2212.06495.

Card, D. (1994). Statistical process control for software? *IEEE*.

Chen, Z., Zhang, J. M., Hort, M., Sarro, F., & Harman, M. (2023). Fairness Testing: A Comprehensive Survey and Analysis of Trends. *ARXIV:* 10.1145.

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *ScienceAdvances*.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness Through Awareness. *ITCS*.

Fontana, M., Naretto, F., Monreale, A., & Giannotti, F. (2022). Monitoring Fairness in HOLDA. *IOS Press*.

Gajane, P., & Pechenizkiy, M. (2018). On Formalizing Fairness in Prediction with Machine Learning. *ARXIV:* 1710.03184.

Ghosh, A., Shanbhag, A., & Wilson, C. (2022). FairCanary: Rapid Continuous Explainable Fairness. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*.

Giffen, B. v., Herhausen, D., & Fahse, T. (2022). Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. *Journal of Business Research*.

Grünwald, P., Heide, R. d., & Koolen, W. (2019). Safe Testing. *IEEE*.

Hajnal, J. (1961). A two-sample sequential t-test. *Biometrika*

Hardt, M., Price, E., & Srebro, N. (2016). Equality of Opportunity in Supervised Learning. *Conference on Neural Information Processing Systems*.

Hu, L., & Chen, Y. (2018). A Short-term Intervention for Long-term. *WWW2018*.

Hunter, J. S. (1986). The Exponentially Weighted Moving. *Journal of Quality Technology* .

Jesus, S., Pombal, J., Alves, D., Cruz, A., Saleiro, P., Ribeiro, R. P., . . . Bizarro, P. (2022). Turning the Tables: Biased, Imbalanced, Dynamic Tabular Datasets for ML Evaluation. *ARXIV:* 2211.13358.

Kim, M., Reingold, O., & Rothblum, G. (2018). Fairness Through Computationallly-Bounded Awareness. *Conference of Neural Information Processing Systems*.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent Trade-Offs in the Fair Determination of Risk Scores. *ITCS*.

Koshiyama, A., E. K., Treleaven, P., Rai, P., Szpruch, L., Pavey, G., . . . Chamorro-Premuzic, T. (2021). Towards Algorithm Auditing. *SSRN* .

Kwak, S. G., & Kim, J. H. (2017). Central limit theorem: the cornerstone of modern statistics . *Korean Journal of Anesthesiology* .

Lakens, D. (2022). Improving Your Statistical Inferences. *10.5281*.

Li, X., Liu, J., & Ying, Z. (2016). Generalized Sequential Probability Ratio Test for Separate. *Sequential Analysis*.

Lloyd, R. (2019). *Quality Health Care: A Guide to Developing and Using Indicators.* Jones & Bartlett Learning.

Loureiro, T. P., Lisboa, F. V., Cruz, G. O., Peixoto, R. M., Guimarães, G. A., Santos, L. L., . . . Mar. (2023). BIAS AND UNFAIRNESS IN MACHINE LEARNING MODELS: A. *Big Data and Cognitive Computing*.

Maity, S., Xue, S., Yurochkin, M., & Sun, Y. (2021). STATISTICAL INFERENCE FOR INDIVIDUAL FAIRNESS. *ICLR* .

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. *ACM Computing Surveys*.

Menon, A. K., & Williamsonv, R. C. (2018). The cost of fairness in binary classification. *PMLR: Proceedings of Machine Learning Research*.

Miller, J., & Steinley, D. (2021). A Simple, General, and Efficient Method for Sequential Hypothesis Testing: The Independent Segments Procedure. *Psychological Methods*.

Montgomery, D. C. (2012). *Introduction to Statistical Quality Control.* John Wiley & Sons.

Neuburger, J., Walker, K., Sherlaw-Johnson, C., Meulen, J. v., & Cromwell, D. A. (2017). Comparison of control charts for monitoring clinical performance using binary data. *BMJ Quality Safety*.

Olteanu, A., Castillo, C., Diaz, F., & Kıcıman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*.

OpenAI. (2023). ChatGPT (version march 2023) [Large language model]. https://chat.openai.com

Ovalle, A., Dev, S., Zhao, J., Sarrafzadeh, M., & Chang, K.-W. (2022). Auditing Algorithmic Fairness in Machine Learning for Health with. *ARXIV*.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On Fairness and Calibration. *ARXIV:* 1709.02012.

Proschan, M. A., Lan, K. K., & Wittes, J. T. (2007). Statistical Monitoring of Clinical Trials: A Unified Approach. *Jounral of the Interantional Biometrics Society*.

Qiu, P. (2013). *Introduction to Statistical Process Control.* CRC Press.

Quy, T. L., Roy, A., Iosifidis, V., Zhang, W., & Ntoutsi, E. (2022). A survey on datasets for fairness-aware machine learning. *WIREs Data mining and knowledge discovery*.

Reich, C. L., & Vijaykumar, S. (2021). A Possibility in Algorithmic Fairness: Can Calibration and Equal Error Rates Be Reconciled? *Schloss Dagstuhl*.

Rouder, J., Speckman, P., Sun, D., Morey, R., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*.

Russel, C., Kusner, M., & Loftus, J. (2017). When Worlds Collide: Integrating Different. *Conference on Neural Information Processing Systems*.

Rushton, S. (1950). On a sequential t-test. *Biometrika*

Rydning, J. (2022, May). *Worldwide Global StorageSphere Forecast, 2022–2026: An Installed Base of 7.9ZB of Storage Capacity in 2021 Came at a Cost of $370 Billion — Is It Enough?* Retrieved from IDC: https://www.idc.com/getdoc.jsp?containerId=US49051122

Ryu, H. J., Adam, H., & Mitchell, M. (2018). InclusiveFaceNet: Improving Face Attribute Detection with Race and Gender Diversity. *Computer Vision and Pattern Recognition*.

Saleiro, P. P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., . . . Ghani, R. (2018). Aequitas: A Bias and Fairness Audit Toolkit. *ARXIV:* 1811.05577.

Sam Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. *ARXIV:* 1701.08230.

Schnuerch, M., & Erdfelder, E. (2020). Controlling Decision Errors With Minimal Costs: The Sequential. *American Psychological Association*.

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential Hypothesis Testing With Bayes Factors: Eciently Testing. *Psychological Method*.

Shankar, V., Roelofs, R., Mania, H., Fang, A., Recht, B., & Schmidt, L. (2020). Evaluating Machine Accuracy on ImageNet. *PMLR*. Retrieved from Evaluating Machine Accuracy on ImageNet.

Shewhart, W. A. (1920). *Economic Control Of Quality Of Manufactured Product.*

Shrestha, S., & Das, S. (2022). Exploring gender biases in ML and AI academic research through systematic literature review. *Frontiers in Artificial Intelligence*.

Stanley, K. (2003). Learning Concept Drift with a Committee of Decision Trees. *Informe técnico*.

Suresh, H., & Guttag, J. V. (2019). A Framework for Understanding Unintended Consequences of Machine Learning. *Equity and Access in Algorithms, Mechanisms, and Optimization*.

Vasudevan, S., & Kenthapadi, K. (2020). LiFT: A Scalable Framework for Measuring Fairness in ML Applications. *ARXIV:* 2008.07433.

VII of the Civil Rights Act (USA 1964).

Wadsworth, C., Vera, F., & Piech, C. (2018). Achieving Fairness through Adversarial Learning: an Application to Recidivism Prediction. *ARXIV:* 1807.00199.

Wald, A. (1945). Sequential Tests of Statistical Hypotheses. *Annals of Mathematical Statistics*.

Wilson, C., Ghosh, A., Jiang, S., Mislove, A., Baker, L., Szary, J., . . . Polli, F. (2021). Building and Auditing Fair Algorithms: A Case Study in Candidate Screening. *FAccT* .

Xue, S., Yurochkin, M., & Sun, Y. (2020). Auditing ML Models for Individual Bias and Unfairness. *Proceedings of the 23rdInternational Conference on Artificial Intelligence and Statistics*.

Yeom, S., & Fredrikson, M. (2020). Individual Fairness Revisited: Transferring Techniques from Adversarial Robustness. *Joint Conference on Artificial Intelligence* .

Zafar, M. B., Valera, I., Rodriguez, M. G., & Gummadi, K. P. (2017). Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. *WWW2017*.

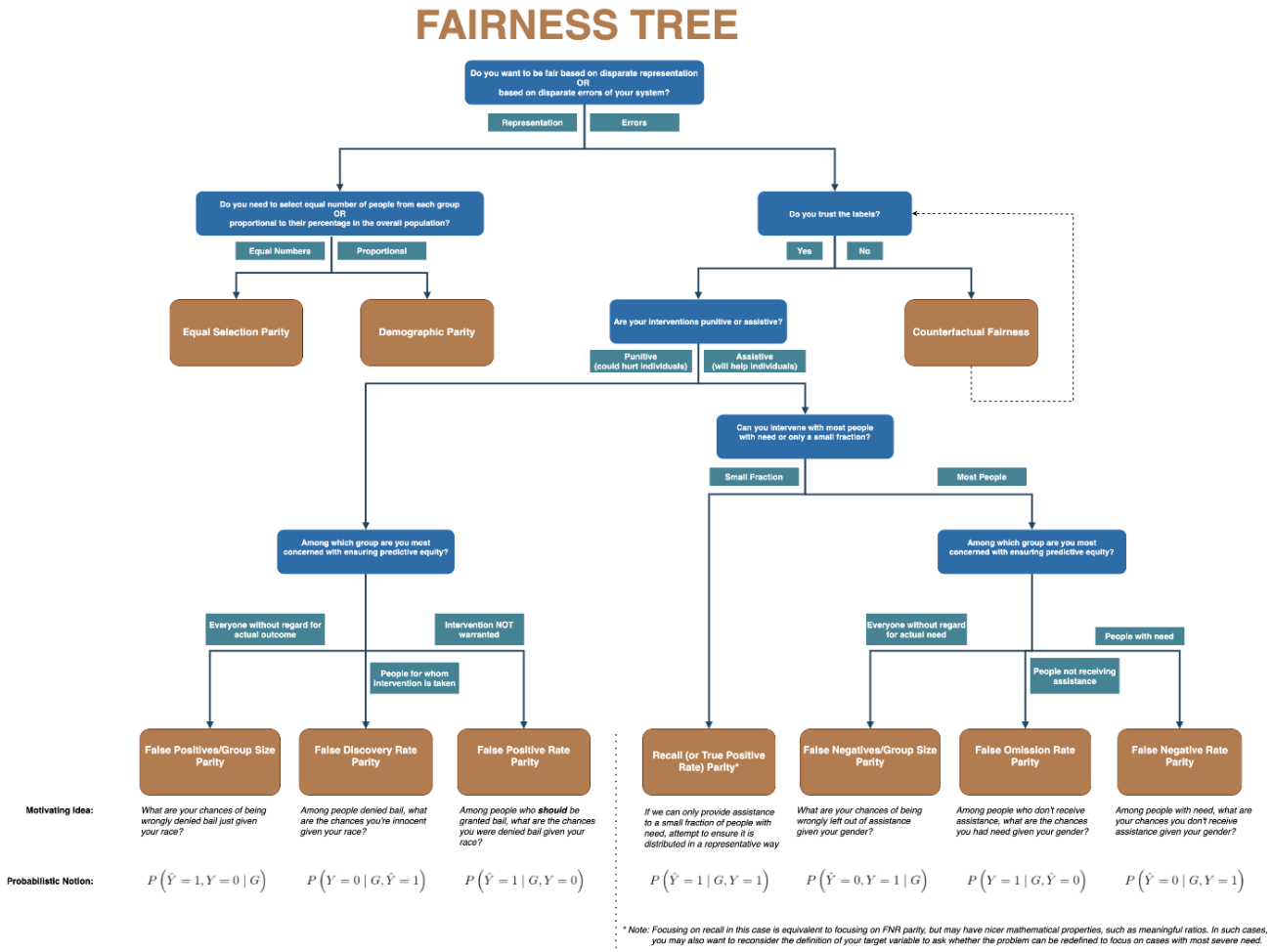
Zarya, V. (2018). *The Share of Female CEOs in the Fortune 500 Dropped by 25% in 2018*. Retrieved from Fortune.

Zehlike, M., Castillo, C., Bonchi, F., Hajian, S., & Megahed, M. (2017). *Fairness Measures: Datasets and software*. Retrieved from http://fairness-measures.org/

Appendices

Appendix A: Fairness measure decision tree

Figure 8. The Aequitas fairness tree as proposed by Saleiro et al. (2018).



Appendix B: Signs used in methodology section

Table 4. Symbol methodology section

Symbol	Defintion
x	An individual measurement of a process
$\bar{x}$	Average of all the individual measurements
i	Index
p	True ratio of 'old' fraudulent individuals
$\hat{p}$	Ratio of 'old' fraudulent individuals in a sample
$\bar{p}$	Estimate of ratio of 'old' fraudulent individuals
D	The number of fraudulent elderly in a sample
m	The number of (preliminary) samples
n	Number of individuals in a sample
j	Number of values collected
h	Decision interval CUSUM
k	Reference value determining sensitivity to process shifts of CUSUM
a	Probability of a Type I error
b	Probability of a Type II error
d	Cohen's d, i.e. the true standardized difference of mean(s) of the populations underlying the group(s)
E	Measured empirical data

Appendix C: Statistical Process Control charts & SBF chart

Figure 9. Shewhart chart

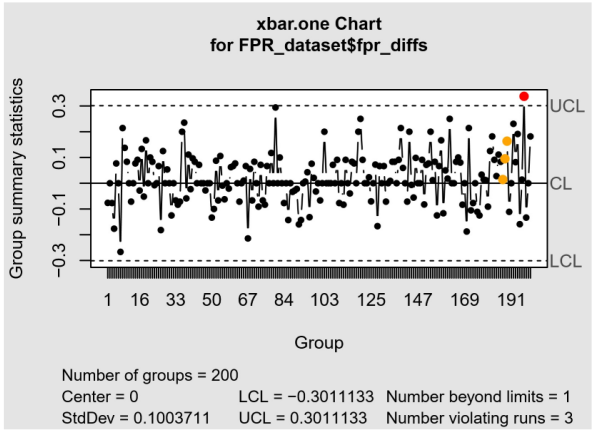


Figure 10. CUSUM chart

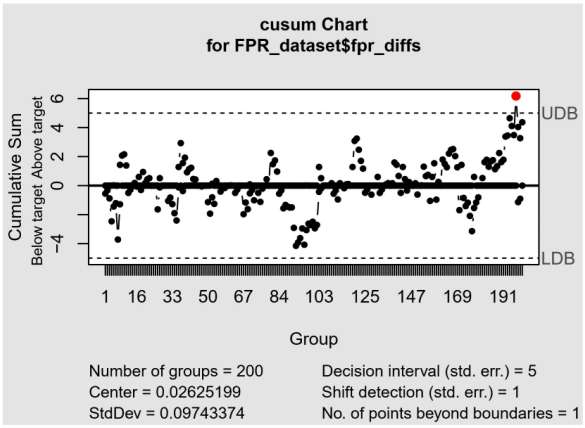
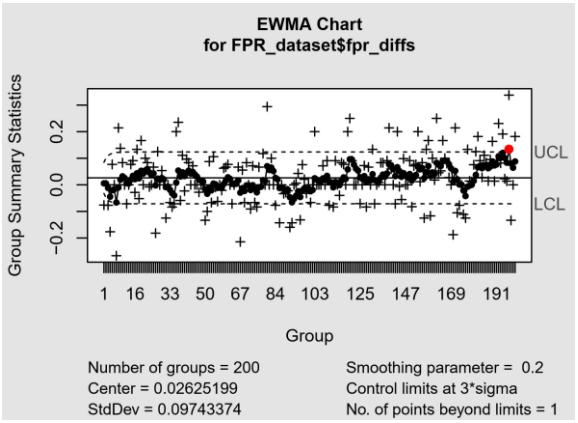


Figure 11. EWMA chart



Appendix D: Results

Table 5. CUSUM results

	Mean						nPMI					
	Flag rate		*ARL*		*Flag density*		*Flag rate*		*ARL*		*Flag Density*	
	IC	**OC**	**IC**	**OC**	**IC**	**OC**	**IC**	**OC**	**IC**	**OC**	**IC**	**OC**
ds_2_500	0.00	0.62	-	0.46	-	0.04	0.02	0.54	0.70	0.43	0.20	0.03
ds_3_500	0.01	0.86	0.90	0.44	0.10	0.07	0.02	0.78	0.70	0.46	0.20	0.05
ds_4_500	0.01	0.99	0.90	0.40	0.10	0.16	0.03	0.87	0.67	0.44	0.13	0.09
ds_5_500	0.02	0.88	0.90	0.36	0.10	0.11	0.02	0.82	0.55	0.36	0.25	0.06
ds_2_5000	0.27	0.33	0.53	0.45	0.02	0.06	0.28	0.37	0.46	0.43	0.04	0.04
ds_3_5000	0.25	0.49	0.53	0.55	0.03	0.07	0.28	0.44	0.46	0.50	0.04	0.05
ds_4_5000	0.24	0.65	0.55	0.57	0.03	0.09	0.27	0.51	0.44	0.52	0.04	0.06
ds_5_5000	0.24	0.59	0.50	0.47	0.03	0.10	0.31	0.53	0.51	0.46	0.04	0.07
ds_2_8500	0.33	0.10	0.45	0.53	0.02	0.06	0.50	0.10	0.47	0.54	0.02	0.06
ds_3_8500	0.33	0.12	0.45	0.60	0.02	0.10	0.51	0.12	0.46	0.52	0.02	0.08
ds_4_8500	0.32	0.19	0.44	0.70	0.02	0.11	0.52	0.15	0.47	0.54	0.02	0.09
ds_5_8500	0.39	0.21	0.49	0.58	0.03	0.14	0.59	0.23	0.47	0.46	0.03	0.12
ds_1	0.44	-	0.45	-	0.02	-	0.52	-	0.41	-	0.02	-

	FPR						FPR – Last 1000 obs.					
	Flag rate		*ARL*		*Flag density*		*Flag rate*		*ARL*		*Flag Density*	
	IC	**OC**	**IC**	**OC**	**IC**	**OC**	**IC**	**OC**	**IC**	**OC**	**IC**	**OC**
ds_2_500	0.03	0.86	0.67	0.44	0.1	0.1	0.01	0.7	0.70	0.58	0.1	0.02
ds_3_500	0.02	0.99	0.8	0.4	0.1	0.33	0.01	0.79	0.70	0.63	0.1	0.03
ds_4_500	0.01	0.99	0.7	0.33	0.1	0.54	-	0.9	-	0.63	-	0.05
ds_5_500	0.02	1	0.5	0.17	0.1	0.46	-	0.63	-	0.48	-	0.03
ds_2_5000	0.23	0.69	0.47	0.55	0.03	0.08	0.19	0.58	0.50	0.61	0.02	0.05
ds_3_5000	0.21	0.92	0.48	0.54	0.03	0.16	0.14	0.74	0.51	0.65	0.01	0.09
ds_4_5000	0.19	1	0.47	0.47	0.03	0.3	0.06	0.97	0.47	0.67	0.01	0.13
ds_5_5000	0.2	0.98	0.62	0.37	0.03	0.39	0.05	0.86	0.70	0.49	0.01	0.16
ds_2_8500	0.46	0.31	0.39	0.63	0.03	0.14	0.41	0.29	0.42	0.63	0.02	0.11
ds_3_8500	0.47	0.56	0.42	0.63	0.03	0.16	0.37	0.47	0.43	0.67	0.02	0.13
ds_4_8500	0.43	0.73	0.4	0.62	0.02	0.23	0.30	0.66	0.41	0.66	0.02	0.20
ds_5_8500	0.37	0.73	0.34	0.54	0.04	0.32	0.31	0.63	0.35	0.58	0.02	0.29
ds_1	0.61	-	0.44	-	0.03	-	0.5	0	0.45	-	0.02	-

Table 6. EWMA results

	Mean						nPMI					
	Flag rate		*ARL*		*Flag density*		*Flag rate*		*ARL*		*Flag Density*	
	IC	**OC**	**IC**	**OC**	**IC**	**OC**	**IC**	**OC**	**IC**	**OC**	**IC**	**OC**
ds_2_500	0.37	1.00	0.40	0.11	0.11	0.05	0.29	1.00	0.46	0.10	0.13	0.05
ds_3_500	0.37	1.00	0.39	0.11	0.11	0.06	0.29	1.00	0.46	0.10	0.12	0.06
ds_4_500	0.37	1.00	0.40	0.11	0.11	0.07	0.29	1.00	0.46	0.09	0.12	0.06
ds_5_500	0.35	1.00	0.49	0.08	0.12	0.07	0.41	1.00	0.55	0.09	0.12	0.07
ds_2_5000	0.98	0.98	0.21	0.19	0.04	0.05	1.00	1.00	0.24	0.18	0.05	0.05
ds_3_5000	0.99	0.98	0.22	0.19	0.04	0.06	1.00	1.00	0.24	0.18	0.05	0.06
ds_4_5000	0.98	1.00	0.22	0.20	0.04	0.06	1.00	1.00	0.24	0.18	0.05	0.06
ds_5_5000	1.00	1.00	0.23	0.19	0.04	0.06	0.99	1.00	0.25	0.20	0.05	0.06
ds_2_8500	1.00	0.77	0.11	0.41	0.04	0.06	1.00	0.81	0.13	0.42	0.05	0.06
ds_3_8500	1.00	0.79	0.11	0.39	0.04	0.07	1.00	0.82	0.13	0.43	0.05	0.06
ds_4_8500	1.00	0.82	0.11	0.40	0.04	0.07	1.00	0.83	0.13	0.43	0.05	0.07
ds_5_8500	1.00	0.83	0.11	0.34	0.05	0.08	1.00	0.83	0.12	0.37	0.05	0.07
ds_1	1	-	0.08	-	0.04	-	1	-	0.09	-	0.05	-

	FPR						FPR – Last 1000 obs.					
	Flag rate		*ARL*		*Flag density*		*Flag rate*		*ARL*		*Flag Density*	
	IC	**OC**	**IC**	**OC**	**IC**	**OC**	**IC**	**OC**	**IC**	**OC**	**IC**	**OC**
ds_2_500	0.04	0.93	0.45	0.43	0.1	0.01	0.41	1	0.48	0.07	0.12	0.08
ds_3_500	0.04	0.98	0.45	0.42	0.1	0.02	0.4	1	0.47	0.07	0.12	0.11
ds_4_500	0.04	0.99	0.45	0.38	0.1	0.03	0.4	1	0.47	0.06	0.12	0.14
ds_5_500	0.07	0.97	0.7	0.29	0.1	0.02	0.47	1	0.52	0.05	0.12	0.13
ds_2_5000	0.48	0.67	0.48	0.53	0.01	0.02	1	1	0.16	0.16	0.06	0.07
ds_3_5000	0.47	0.84	0.49	0.55	0.01	0.02	1	1	0.16	0.15	0.06	0.08
ds_4_5000	0.44	0.96	0.47	0.55	0.01	0.03	1	1	0.16	0.14	0.06	0.1
ds_5_5000	0.43	0.94	0.5	0.43	0.01	0.03	1	1	0.18	0.11	0.06	0.11
ds_2_8500	0.78	0.32	0.34	0.6	0.01	0.04	1	0.76	0.1	0.25	0.07	0.06
ds_3_8500	0.76	0.47	0.34	0.61	0.01	0.04	1	0.82	0.1	0.27	0.07	0.06
ds_4_8500	0.75	0.65	0.35	0.63	0.01	0.05	1	0.87	0.1	0.28	0.07	0.07
ds_5_8500	0.69	0.71	0.36	0.56	0.01	0.05	1	0.95	0.09	0.28	0.06	0.08
ds_1	0.79	-	0.39	-	0.01	-	1	-	0.08	-	0.06	-

Table 7. SPRT-t test results

	Mean				FPR				FPR – last 1000 obs.			
	Flag rate		*ARL*		*Flag rate*		*ARL*		*Flag rate*		*ARL*	
	IC	**OC**	**IC**	**OC**	**IC**	**OC**	**IC**	**OC**	**IC**	**OC**	**IC**	**OC**
ds_2_500	-	0.28	-	0.77	-	0.87	-	0.63	0.00	0.31	-	0.54
ds_3_500	-	0.65	-	0.73	-	1	-	0.46	0.00	0.62	-	0.58
ds_4_500	-	0.84	-	0.66	-	1	-	0.38	0.00	0.84	-	0.57
ds_5_500	-	0.85	-	0.55	-	1	-	0.24	0.00	0.63	-	0.49
ds_2_5000	0.04	0.09	0.73	0.63	-	0.31	-	0.71	0.10	0.12	0.40	0.38
ds_3_5000	0.04	0.16	0.73	0.71	-	0.6	-	0.72	0.10	0.23	0.40	0.56
ds_4_5000	0.04	0.21	0.73	0.73	-	0.87	-	0.68	0.10	0.32	0.40	0.58
ds_5_5000	0.03	0.32	0.65	0.65	0.03	0.93	0.82	0.59	0.09	0.32	0.56	0.48
ds_2_8500	0.13	0.01	0.64	0.17	0.12	0.13	0.62	0.33	0.21	0.03	0.44	0.47
ds_3_8500	0.13	0.02	0.64	0.48	0.12	0.18	0.62	0.44	0.21	0.04	0.44	0.58
ds_4_8500	0.13	0.03	0.64	0.64	0.12	0.25	0.62	0.49	0.21	0.04	0.44	0.58
ds_5_8500	0.08	0.03	0.67	0.49	0.17	0.26	0.69	0.44	0.20	0.06	0.49	0.52
ds_1	0.13	-	0.71	-	0.21	-	0.6	-	0.15	-	0.46	-

Table 8. SBF results

	Mean				FPR				FPR – last 1000 obs.			
	Flag rate		*ARL*		*Flag rate*		*ARL*		*Flag rate*		*ARL*	
	IC	**OC**	**IC**	**OC**	**IC**	**OC**	**IC**	**OC**	**IC**	**OC**	**IC**	**OC**
ds_2_500	0.2	0.4	0.14	0.03	-	1	-	0.67	-	0.2	-	0.59
ds_3_500	0.2	0.6	0.14	0.33	-	1	-	0.52	-	0.8	-	0.55
ds_4_500	0.2	0.6	0.14	0.28	-	1	-	0.39	-	1	-	0.51
ds_5_500	-	1	-	0.74	-	1	-	0.23	-	1	-	0.47
ds_2_5000	-	-	-	-	-	0.8	-	0.62	-	0.6	-	0.62
ds_3_5000	-	-	-	-	-	0.8	-	0.46	-	0.8	-	0.64
ds_4_5000	-	0.4	-	0.95	-	0.8	-	0.39	-	0.8	-	0.56
ds_5_5000	-	0.2	-	0.08	-	1	-	0.59	-	1	-	0.46
ds_2_8500	-	-	-	-	-	-	-	-	-	-	-	-
ds_3_8500	-	-	-	-	-	-	-	-	-	-	-	-
ds_4_8500	-	-	-	-	-	-	-	-	-	0.4	-	0.78
ds_5_8500	0.4	-	-	0.16	-	-	-	-	-	0.6	-	0.68
ds_1	0.2	-	0.06	-	-	-	-	-	-	-	-	-

Appendix E: Dataset plots

Figure 12. Mean fraudulent predictions over datasets[5]

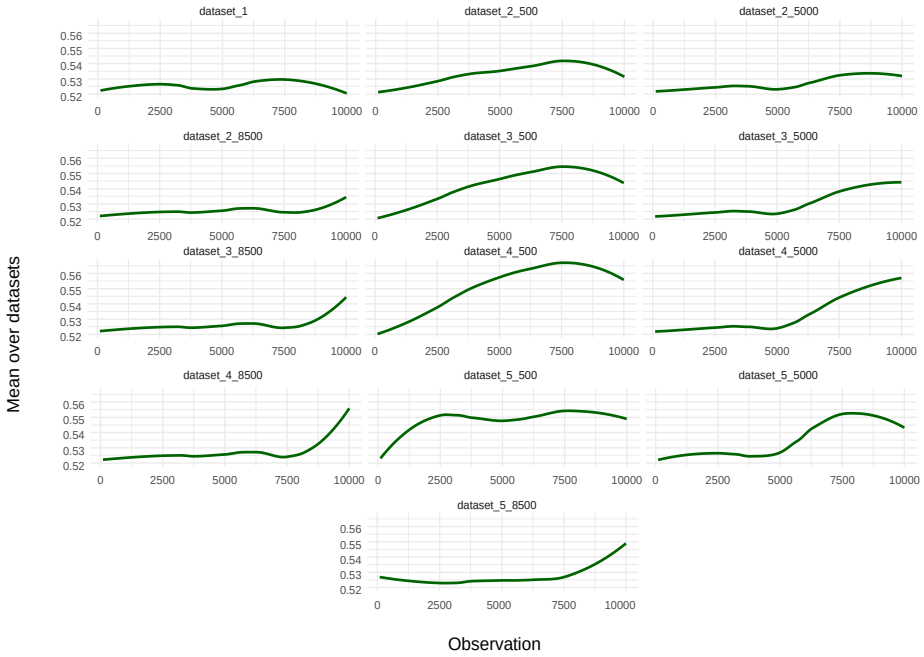
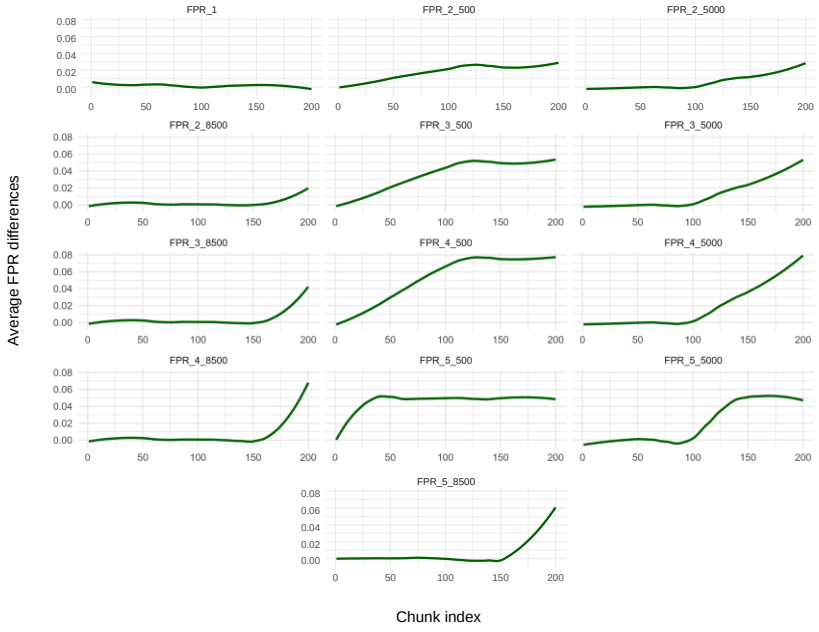
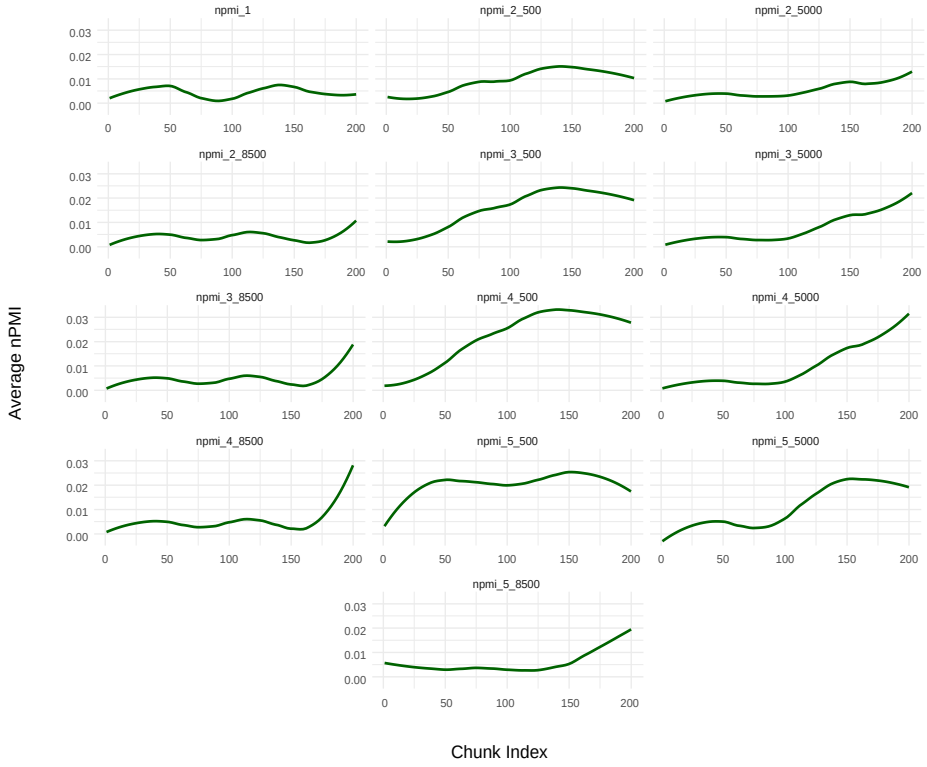


Figure 13. Average FPR differences over datasets



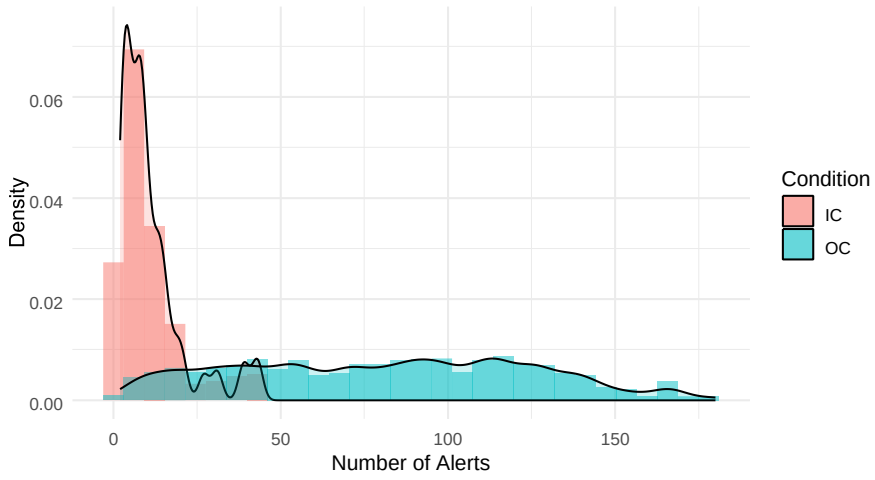
[5] Note that the average fraudulent value is slightly larger than 0.5 because the datasets have FP values, without having FN values.

Figure 14. Average nPMI over datasets



Appendix F: Setting the CUSUM threshold

Figure 15. Distribution alert frequency IC and OC CUSUM



Appendix G: Flagged datasets SPRT-t and CUSUM combined

Table 9. Number of flagged OC cases SPRT-t and CUSUM (with threshold rule) combined

Dataset	Count
dataset_3_500	100
dataset_4_500	100
dataset_5_500	100
dataset_5_5000	95
dataset_4_5000	93
dataset_2_500	88
dataset_3_5000	73
dataset_5_8500	44
dataset_2_5000	37
dataset_4_8500	34
dataset_3_8500	24
dataset_2_8500	18