technische universität
dortmund

Fairness enhancement methods in Anomaly Detection for DEAN

Benning, Marcelo

Bachelorarbeit

Studiengang: Bachelor Informatik
Matrikelnummer: 220846
Erstgutachter: Prof. Dr. Emmanuel Müller
Zweitgutachter: Dr. Simon Klüttermann

Bearbeitungszeit: 21.07.2025 – 17.11.2025

Lehrstuhl für Data Science and Data Engineering
Fakultät für Informatik

Zusammenfassung

Diese Arbeit befasst sich mit der Bewertung und Verbesserung der Fairness im DEAN-Modell (Deep Ensemble Anomaly Detection). Das Ensemble aus mehreren kleineren neuronalen Netzen soll zunächst Anomalien erkennen und anschließend auf Fairness getestet werden. Diese wird in der Arbeit durch ein einzelnes geschütztes Merkmal abgebildet. Der Fokus der Arbeit liegt darin, DEAN durch verschiedene Methoden zu verbessern. Das Modell arbeitet, indem es die Anomalie-Wahrscheinlichkeiten der vielen Submodelle aggregiert und dann als Ensemble den Durchschnittswert bildet. Das ermöglicht ein schnelleres Training und eine höhere Robustheit, da einzelne Submodelle herausgefiltert werden können. Fairness ist im Bereich des maschinellen Lernens noch wenig untersuchtes Thema. Unter Fairness wird verstanden, dass verschiedene Gruppen bei gleichen Bedingungen ähnliche Ergebnisse erhalten. Von Diskriminierung spricht man, wenn ein Modell bei gleichen Voraussetzungen unterschiedliche Fehlerquoten für verschiedene Gruppen aufweist. Ziel dieser Arbeit ist es, die Fairness zu messen und zu verbessern, ohne die Leistungsfähigkeit des Modells (AUC-Score) zu verringern. Es gibt viele verschiedene Methoden, Fairness zu messen – beispielsweise den Unterschied der positiven Raten zwischen Gruppen oder den Unterschied der True-Positive-Raten. In dieser Arbeit wird die Fairness dadurch angegeben, wie stark das Modell mit dem *protected feature* korreliert. Ein Wert nahe 0,5 spricht für ein faires Modell, da keine statistische Abhängigkeit zwischen dem Feature und der Vorhersage besteht.

Die in dieser Arbeit verwendeten Methoden sind:

- eine Fairness-Filterung auf Submodellebene – das heißt, Submodelle oberhalb eines bestimmten Schwellenwerts werden nicht in das Ensemble aufgenommen,
- eine Anpassung der Loss-Funktion der einzelnen Submodelle, sodass fairere Modelle eine höhere Gewichtung im Ensemble erhalten als unfaire.

Das Ziel besteht darin, das Modell fairer zu gestalten, ohne einen Verlust an Leistungsfähigkeit (AUC) zu verursachen. Durch diese Anpassungen wird die Laufzeit des Modells nicht erhöht.

Abstract

This thesis focuses on the evaluation and improvement of fairness in the DEAN model (Deep Ensemble Anomaly Detection). The ensemble, consisting of several smaller neural networks, is first designed to detect anomalies and is then tested for fairness. Fairness in this work is represented through a *single protected feature*. The main focus of the thesis is to improve DEAN by means of different fairness-enhancing methods.

The model operates by aggregating the anomaly probabilities of the many submodels and then computing their average as the ensemble output. This approach allows for faster training and increased robustness, since individual submodels can be filtered out.

Fairness in the field of machine learning is still a relatively underexplored topic. In general terms, fairness means that different groups receive similar outcomes under the same conditions. Discrimination occurs when a model predicts different error rates for groups under identical circumstances. The goal of this work is to measure and improve fairness without decreasing the model performance, expressed through the AUC score.

There are several approaches to measuring fairness, such as the differences in positive rates between groups or the differences in true positive rates. In this thesis, fairness is defined by the degree to which the model output correlates with the *protected feature*. A value close to 0.5 indicates a fair model, since there is no statistical dependency between the protected feature and the model prediction.

The methods used in this work are:

- fairness filtering at the submodel level – meaning that submodels above a specific threshold are not included in the ensemble,
- adjustment of the loss function of individual submodels, so that fairer models receive higher weights in the ensemble compared to unfair ones.

The objective is to improve the model's fairness without causing a loss in performance (measured by ROC-AUC). These modifications also do not increase the runtime of the model.

Contents

List of Figures

List of Tables

1 Introduction

Anomaly detection plays a relevant part in today's ever-growing data-driven world. It is used to find conspicuous patterns in large datasets. These patterns can indicate failure, fraud, or other critical findings. The use cases for anomaly detection are diverse. For example, anomaly detection is used in financial fraud (Niu et al., 2019), technical system failures (Sanami and Aghdam, 2024), or cybersecurity breaches (Zhang et al., 2018). Recent advances in artificial neural networks have improved the ability of the models to detect more complex anomaly types (Huang et al., 2025).
In spite of these advances, one problem is largely overlooked. Many models discriminate against certain groups. This can take place when the training set is imbalanced, so that the model predominantly learns the patterns of the larger, more represented group. As a result, the less represented or protected group may be incorrectly flagged, even though they should not be (Wu et al., 2024). This is why fairness in anomaly detection is becoming increasingly important research topic. The aim is to secure, that the model treats all groups in the same way, without bias towards any one group.
This thesis addresses this issue by modifying the DEAN framework. It proposes variants of DEAN, that reduce its biases towards a protected group. The thesis begins with background knowledge on anomaly detection and related work. Then, it introduces the evaluation metrics used, followed by the datasets applied in this work. After that, the two modification methods for DEAN are explained. Finally, the evaluation results are presented, followed by the conclusion and an outlook for future research.

2 Background

This section provides background information required for this thesis. It introduces the basic concepts of anomaly detection, neural networks, the DEAN model, pruning, fairness in machine learning, with particular focus on fairness in anomaly detection.

2.1 Anomaly Detection

In order to explain what anomaly detection is and how it works, we first define anomalies. Anomalies are data points or events that stand out from the baseline pattern (StrongDM, 2024). An anomaly does not mean that a failure has occurred in a given system. Rather, it means that something out of the ordinary happened and should, for that reason, be further investigated.
Anomaly detection models are trained to indicate anomalies. They are trained by various methods, but fundamentally they work like most machine learning algorithms. They first learn the pattern of normal data, without any anomalies. The way the model learns these patterns depends on the algorithm chosen for the problem. In the final step, it can be deployed in a setting where anomalies can occur.

2.2 Neural Networks

Neural networks are a subfield of machine learning. As the name suggests, neural networks are inspired by the brain's structure with its connected neurons. Each neuron has one or more inputs and a single output. These outputs can then be used as inputs for other neurons, or directly as the output of the network. Inputs for a neuron differ, since one input can be more important than another. To replicate that, the connections between neurons have a weight. Once a neuron hits some kind of threshold (bias), it fires or changes its output based on the weighted inputs and the activation function. A neural network does not need to have only one hidden layer. Once it has more than two-three hidden layers, it is called a deep learning neural network, like the one used in this thesis.

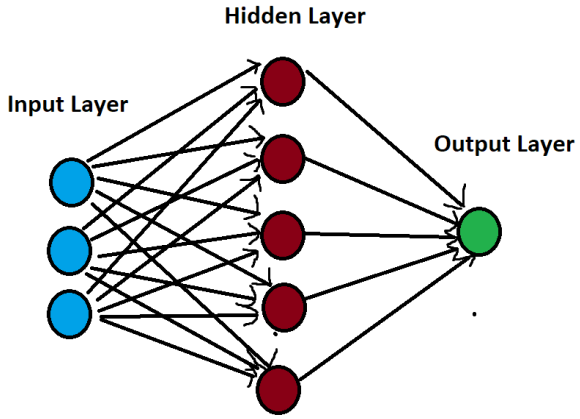

Figure 1: A simple depiction of a neural network with an input layer, one hidden layer and an output layer. The dots are neurons and the arrows are the weighted connection.

2.3 Deep Ensemble Anomaly Detection (DEAN)

DEAN is an unsupervised, deep learning neural network ensemble used for anomaly detection. Ensembles consist of multiple, often simpler, neural networks , which are then combined later into a single model. In terms of anomaly detection, DEAN does not reconstruct the input to predict what the data should be. It learns a target behavior and gives scores on the deviations from it. Once training is completed, samples receive an anomaly score depending on how far they deviate from the target.
Feature bagging is another feature that distinguishes DEAN from other concepts. Since DEAN is an ensemble model, it enables the possibility for feature bagging. Each submodel sees a *random subset of features*, which increases robustness and accuracy by reducing its variance. This is distinctly beneficial for high-dimensional datasets.
These two key properties characterize DEAN. It is simple through constant surrogate anomaly scores. It provides diversity and robustness, through feature bagging, because each submodel specializes on different parts of the data's feature sets. And the aggregation reduces variance and overfitting.
DEAN is also easily scalable, considering each submodel is small and independent, which enables parallelization and predictable runtime even for high-dimensional datasets. Moreover, the modularity of DEAN makes it easy to extend, which is essential for the variants proposed in this thesis.

2.4 Pruning

Pruning in machine learning is an important technique to improve model performance. There are multiple ways to prune a neural network. The commonly used approaches are structured pruning, unstructured pruning and ensemble pruning.

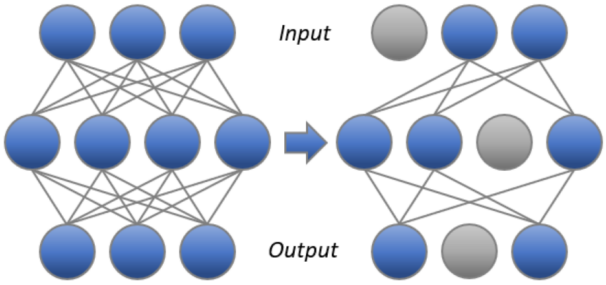

Figure 2: Structured pruning, where complete neurons or sometimes even layers are deleted from the neural network in favor for performance of the model

Structured pruning (Figure 2) sharpens a model by remove whole interpretable components of the neural network. Unlike unstructured pruning, where individual weights are zeroed, it eliminates complete structures such as neurons or even entire layers. Some components contribute little to the predictive performance, but still consume computation time and might even introduce noise. By removing these redundant parts, structured pruning reduces latency and memory consumption, while maintaining and sometimes improving accuracy of the model. After pruning, the neural network is more efficient, faster and easier to deploy.

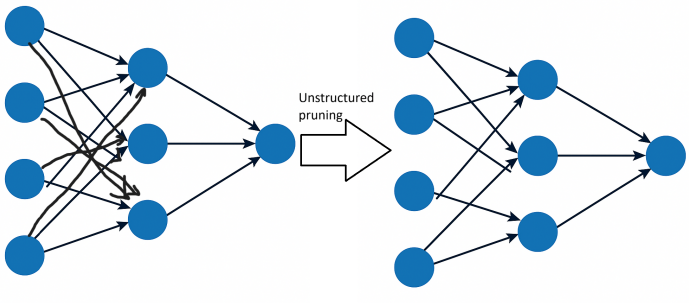

Figure 3: Unstructured pruning, where weights of the neural networks are zeroed and the weight matrices simplified

Unstructured pruning (Figure 3) also makes a model more efficient, but it works by setting individual weights to zero instead of completely removing neurons or layers from the neural network. In doing that, the overall architecture of the model remains the same, while the weight matrices become simpler with the zeroing. This can lower memory usage and sometimes speed improvements. Unstructured pruning keeps the shape of the model and simplifies its parameters, by trading dense computation for sparser ones.
Another pruning method is ensemble pruning. It removes a chosen subset of the models from the ensemble (submodels). The tricky part is figuring out which subset to drop.

A common strategy is to eliminate models that add little to no value or that harm the performance of the ensemble. After pruning, the overall structure of the ensemble is simpler, uses fewer resources, since there are less models to evaluate, and can even perform better, because bad models are dropped.

2.5 Fairness

Fairness in machine learning is critical, because more and more areas such as criminal justice, healthcare and finance include the usage of algorithmic calculations (Caton and Haas, 2024). When the model is biased towards a minority, it will amplify already existing social inequalities.
A well-known example illustrating that problem is the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) system used in the U.S. judicial system. COMPAS is a risk assignment tool, designed to predict how likely a defendant is likely to recidivate. It uses historical data and compares it to the defendant. As a result a risk score will be generated, which the judge uses to assess an appropriate judgment about giving bail or parole and other possibilities. The algorithm sees factors including criminal history, age and social background, but not explicitly the race.
Even though race is not explicitly included as a feature ProPublica (Angwin et al., 2016) showed, that COMPAS exhibits racial bias. Black defendants were twice as likely to be falsely classified as high-risk, when compared to white defendants. On top of that, white defendants on the other hand were more often misclassified as low-risk.
These predictions stem from correlations in the data. Features like prior arrests or socially depending variables are indirectly linked to race, due to preexisting inequalities. The algorithm learned, reflected and reinforced these disparities. This is only one of many other examples why fairness in machine learning and anomaly detection is becoming increasingly crucial. The impact these systems have on people's lives should not be overlooked.
Fortunately there are multiple ways to ensure a fairer model.

1. Data preprocessing to mitigate bias in data
2. Inprocessing at training time, by adding constraints
3. Postprocessing after training time, correcting the results of the classifier

At the end it should be one of the main goals to have the system fair, because otherwise there will be a lack of trustworthiness.

2.5.1 Fairness Metrics

Now that the term fairness has been defined, a metric is needed to evaluate it. For that a confusion matrix is used (Gursoy and Kakadiaris, 2022, Table 1).

	Predicted Positive	**Predicted Negative**
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Table 1: Confusion Matrix

From this matrix, key rates for fairness metrics are defined:

1. **True Positive Rate (TPR)**

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{1}$$

   This measures how well the model identifies actual positives.

2. **False Positive Rate (FPR)**

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{2}$$

   Indicates the rate, in which how often negatives are incorrectly classified as positives.

3. **Positive Predictive Value (PPV)**

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

   This metric is also known as precision. It measures how correct the positive predictions are.

There are many more different metrics that are developed from the confusion matrix, but the three above (Equation 1, Equation 2, Equation 3) are the most valuable ones. Several fairness criteria are based on these.
To work with them, a system will be defined. $\boldsymbol{R}$ represents the predicted classification by the algorithm. $\boldsymbol{Y}$ is the actual outcome for that individual data and $\boldsymbol{A}$ corresponds to the sensitive feature (race, gender, ethnicity). There are various fairness criteria that can be characterized using the metrics and system that was specified. The fundamental ones are *demographic parity* (independence), *equalized odds* (separation) and *predictive parity* (sufficiency).

1. **Demographic Parity** means, that the model's predictions are independent of sensitive attributes, such that protected and unprotected groups have the same probability of being assigned as to a positive class.

$$P(R = 1 \mid A = a) = P(R = 1 \mid A = b) \quad \forall a, b \in A \tag{4}$$

2. **Equalized Odds**, is given when the classifier predicts the same rate of TPR (Equation 1) and FPR (Equation 2)

$$P(R = + \mid Y = y, A = a) = P(R = + \mid Y = y, A = b) \quad y \in \{+, -\} \quad \forall a, b \in A \tag{5}$$

3. **Predictive Parity**, also known as predictive equality, is satisfied if the model has an equal PPV metric between protected and unprotected groups.

$$P(Y = + \mid R = +, A = a) = P(Y = + \mid R = +, A = b) \quad \forall a, b \in A \tag{6}$$

2.5.2 Fairness-Accuracy Trade-off

Even with all the metrics and definition there remains a key challenge, when designing fairness-aware machine learning models. It is the inherent trade-off between fairness and predictive performance of the model. Fairness tries to ensure equal treatment across demographic groups, while accuracy's goal is it to maximize the predictive quality of the model. These two goals often conflict, because the aspect of fairness often restricts the optimization of the learning algorithm of the model.
As shown in Kleinberg et al., 2016, certain fairness definitions, as the one explained above (Equation 4, Equation 5, Equation 6), cannot be all satisfied simultaneously, unless the model has a perfect accuracy and the rates across all groups are identical. This is based on the fact, that they all rely on different conditional probabilities. These conditions are mutually exclusive in most real world cases, because of the differences in base rates for positive outcomes. For example in the COMPAS dataset, it becomes clear, that it is impossible to satisfy all three fairness criteria. In the dataset, the likelihood to reoffend, differs considerably across racial groups. Looking at African-American defendants, they exhibit a higher base rate then Caucasian defendants. In the analyses made by ProPublica (Angwin et al., 2016), it is shown that 45% of African-American reoffend within two years, while the rate of Caucasian lies only around 28%.
This inequality in base rates makes it impossible for the three fairness criteria to be satisfied at the same time, since when giving in to fairness, the accuracy is lowered. For example when the model enforces demographic parity, by predicting equal positive rates throughout groups. It will then over predict the risk of a Caucasian to reoffend and under predict for a African-American to reoffend. This frequently occurs in real-world scenarios, making it mathematically impossible to satisfy independence, separation and sufficiency at the same time. Not only considering the mathematical trade-off, there is also an ethical part, that plays into making a decision. For example when deploying the application in high stakes domain, such as judicial justice system or healthcare, it might be worth a trade-off worth taking. Increased fairness for more equality, but a loss in accuracy.

2.5.3 Fairness in Anomaly Detection

Fairness in anomaly detection is becoming an increasingly important topic than ever, due to the increasing amount of deployments of anomaly detection systems in socially

sensitive areas, such as the one already mentioned. Unlike in traditional supervised classification tasks, anomaly detection operates most of the time in unsupervised settings, where anomalies are sparse and labels are not a given. This setting brings forward unique challenges for fairness evaluation. A key issue is imbalanced group representation. Minorities or the protected group, may be underrepresented in training data, leading models to learn patterns dominated by the majority groups. Consequently, individuals from the minority groups might be incorrectly flagged as anomalies, simply due the fact their behavior deviates from the learned majority pattern (Wu et al., 2024). Another problem, anomaly detection models typically predict a continuous anomaly score, rather than a binary one as output. Fairness for that matter can then vary depending on the chosen threshold for classification (Xiao et al., 2025). Ensembles, such as DEAN, can amplify bias if individual submodels are biased, since aggregating anomaly scores may reinforce structural inequalities. For example, if the majority of submodels assign higher anomaly scores to a protected feature, because of feature correlations, the ensemble will most likely propagate that bias (Du and Zhang, 2021). Given these conditions, fairness can be interpreted as making sure, that protected and unprotected groups have a similar anomaly score distributions, when given comparable conditions, fulfilling with the idea of independence. This is problematic, since achieving this without comprising the performance of the model is very difficult, because anomaly detection inherently is based on deviations from learned patterns, which may correlate with demographic features.
To manage these problems, there are several mitigating strategies, that also were mentioned before. In form of preprocessing, reweighing or resampling to balance group representation. Inprocessing uses fairness constraints, which are put into the loss function. Postprocessing to adjust the anomaly score threshold per group to equalize error rates. Lastly ensemble filtering, which is removing biased submodels based on a given fairness evaluation. All those strategies try to reduce bias, while also keeping a high performance.

3 Related Works

Research on fairness in machine learning has grown drastically in recent years, because more models are deployed into high stakes domains, such as criminal justice, healthcare and finance. However, there is still considerably less researches on unsupervised classification than on supervised classification.

3.1 Fairness in Machine Learning

Caton and Haas, 2024, already provide a wide survey about fairness in machine learning. They review group fairness metrics under demographic parity (Equation 4), equalized odds (Equation 5) and predictive parity (Equation 6). The authors analyzed three main mitigation techniques to improve the fairness metrics. Preprocessing, reweighting samples and generating more data for a fairer representation. Inprocessing, adding constraints directly to the model's training process. This is done by embedding regularization terms or adversarial debiasing into the loss function. The last main technique done in this work is postprocessing, in form of changing the threshold per group. The results of this work show that, when considering fairness, there often will be a trade-off with accuracy and that there is no single method (in their work), that solves the fairness issue. They emphasize, that the domain in which the model is deployed plays a crucial role in selecting the adequate fairness metric and mitigation strategy.

3.2 Fairness in Anomaly Detection

Wu et al., 2024, engage with one the hardest challenge in anomaly detection. The disproportionate impact of imbalanced group representation in fairness. Their work shows, that minorities are more likely to be flagged as abnormal, merely because their behavior differs from the majority's pattern, even though it is normal. This phenomenon occurs, because anomaly detection models are trained to learn what is "normal". When given a more dominant group in the training data, all other minority groups appear suspicious. To tackle this bias, Wu et al. advocates fairness aware anomaly detection methods, that considers group sensitive adjustments into the scoring process. More precisely, they implemented techniques, that normalize anomaly scores for demographic groups and apply reweighting strategies during training, to make sure minorities are reasonably represented. The results of their work reveal, that these additions significantly reduce disparities in false positive rates without hindering the anomaly detection accuracy. The paper is relevant since it highlights the unique requirements, need for fairness in an unsupervised setting. Working with statistical deviations, rather than with explicit classifications.

3.3 Fairness Constraints in Anomaly Detection

Xiao et al., 2025, deals with fairness in anomaly detection by using fairness-aware anomaly detection projection techniques, combined with inprocessing constraints during the model's training duration. Their approach of this paper is based on the knowledge, that anomaly detection models often learn a representation of the data, that consequently encodes sensitive attributes, resulting in biased anomaly scores. To manage this problem, they introduce a **Fair Projection method**. This projects the data onto a new hidden feature space, where sensitive features have minimal influence on the anomaly scores. This projection is learned together with the normal anomaly detection objective, assuring that the constraints containing fairness are already integrated into the optimization process, rather than after. The authors of the paper encoded this by adding a regularization term to the loss function, that penalizes correlations between sensitive features and anomaly scores. They further push this, by adding an adversary model, which tries to predict sensitive features from the projected representation. After that, the main model's goal additionally becomes to minimize the success of that adversary model. This setup brings the model to produce results, which are informative for anomaly detection, while also being uninformative for sensitive features.
Their results show that fairness aware projection confidentially reduces disparities in false positive rates and anomaly score distributions across demographic groups. Most importantly, they show, that while improving fairness, they only decreased in detection performance only by a little. They also showed, that this approach, fairness constraints embedded in the training process, outperform postprocessing methods.

3.4 Bias in Ensemble Models

Du and Zhang, 2021, published a selective ensemble algorithm for network anomaly detection. Its goal is to increase the detection accuracy and robustness, while reducing redundancy among submodels. In a normal ensemble methods, all submodels are aggregated at a same value, but their approach evaluates each submodel before aggregating based on two factors:

- individual performance
- diversity contribution

They measured performance using normal metrics scales, like detection accuracy and false positive rate, while diversity made sure only models that provide new, complementary views, as opposed of already represented predictions. The procedure involves a parallel learning framework, where multiple base learners are trained at the same time, but on different feature subsets and or sampling variations. One the training is done, the algorithm calculates a selection score for each submodel. This selection score is a combined factor of that submodel's accuracy and diversity contribution. Only submodels with a high enough selection score are included in the final ensemble. This

selective procedure reduces the computational overhead, needed to create an ensemble, and improves the generalization compared to full ensembles. Their results for the selective ensemble method show on benchmark for network anomaly detection show, that their ensemble outperforms traditional ensemble and single models in terms of detection accuracy and robustness.
This paper does not address fairness, but the idea is still relevant since the selection factors can be edited, so that fairness becomes a relevant factor for each submodel.

4 Methodology

The metrics used for this thesis will be presented in this chapter. The one to calculate the accuracy and the fairness. Furthermore, the ideas and how it was done, to tweak DEAN, will be explained in this chapter.

4.1 Evaluation Metrics

Since anomaly detection is not a simple classification task, but rather a scoring problem, where each instance of data points receives a value between [0,1] and a threshold, which determines whether or not the data point is considered anomalous, the ROC-AUC metric was chosen for evaluation. Because unlike accuracy, which depends on a fixed threshold, ROC-AUC evaluates the model's strength to distinguish between normal and anomalous data across all thresholds. The Receiver Operating Characteristic (ROC) curve compares the trade-off between the True Positive Rate (Equation 1) against the False Positive Rate (Equation 2) for different thresholds. This method gives a good overview of the trade-off between the two rates.

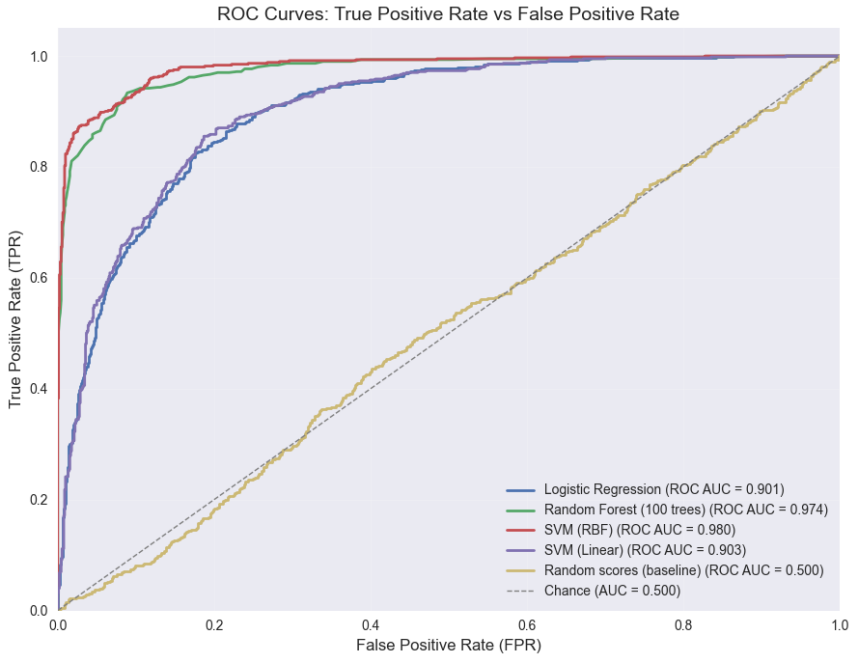

Figure 4: This picture depicts different ROC curves for different classifiers. The x-axis shows the False Positive Rate and the y-axis shows the True positive Rate. The legend reads die different ROC-AUC for each classifier and also a baseline ROC-AUC of 0.5.

The Area Under the Curve (AUC) is the condensed form in a numerical form between 0 and 1. If the AUC is 1, it indicates a perfect classifier, while an AUC of 0.5

corresponds to random guessing. ROC-AUC is the most common ranking metric for anomaly detection.
For fairness evaluation, the same principle can be adopted. Instead of comparing anomaly scores to the true label, the scores are compared to the protected feature. Meaning, given the anomaly score, how much of the sensitive data are leaked. In this scenario, regarding are fair model, it should not allow the protected feature to be predicted from its anomaly scores. For that reason, an ROC-AUC value close to 0.5 indicates a fair model, as it advocates no correlation between anomaly scores and sensitive attributes. On the opposite, a value higher than 0.5 implies that the anomaly scores carry information about the protected feature, meaning the model shows bias.

4.2 DEAN variants

In this section, the two different approaches to improve fairness in a DEAN framework will be presented, after some more detailed explanation of the baseline DEAN. Each builds are simple tweaks to the core DEAN architecture, that bring distinctive methods to enhance the fairness of the model.

4.2.1 Baseline DEAN Model

The concept of DEAN was already explained in subsection 2.3. The following subsection describes the training procedure, the loss function and the scoring function in more detail. Firstly each submodel, trained on normal pattern, tries to output a constant value.

$$g_{\text{DEAN}}(x) = 1 \quad \forall x \in X_{\text{train}} \tag{7}$$

This simple function ensures a smooth training phase, making the model robust and computationally more efficient. The learned function of each submodel *f(x)* is therefor optimized to minimize the deviation from the constant 1. The loss function and scoring function results into:

$$\mathcal{L} = \sum_{x \in X_{\text{train}}} |f(x) - 1|, \qquad score(x) = \|f(x) - q\| \tag{8}$$

with the output of a submodel

$$q = \frac{1}{\|X_{\text{train}}\|} \sum_{x_T \in X_{\text{train}}} f(x_T) \approx 1 \tag{9}$$

This loss function enforces, that normal data points result into an output close to 1, while anomalous data points, which by definition deviate from the learned pattern, lead to larger differences to 1. The anomaly score function then aggregates the deviations across all submodels:

$$score_F(x) = \frac{1}{\|F\|} \sum_{f_i \in F} \|f_i(x) - q_i\|^{power} \tag{10}$$

F is the space of submodels, q_i is the output of the given submodel f_i and *power* denotes the sensitivity parameter, to which degree the submodels are aggregated.
With this simple architecture, DEAN is in the next two points edited to reinsure fairness enhancement.

4.3 Submodel Pruning

This method here introduces a filtering mechanism at submodel level. The core function is to measure each submodel's fairness score individually and discard it, when its fairness score deviates too strongly from the desired value. The fairness score is derived by measuring the correlation between the given score and protected feature using the ROC-AUC metric explained in subsection 4.1. The closer the score to 0.5, the less predictive power regarding the protected feature, which satisfy the desired fairness.
The deviation of each submodel's fairness value is calculated as:

$$D = |fairness\_score - 0.5|, \qquad D \leq fairness\_threshold \tag{11}$$

Submodels whose deviation exceeds a predefined threshold (fairness_threshold) are considered bias and discarded. Only submodels that satisfy the second constraint are kept in the ensemble. This pruning step makes sure, that the final ensemble gets it's aggregated predictions from submodels, that satisfy the desired fairness, thereby reducing the overall bias of the ensemble.
The aggregation of anomaly scores are kept the same from the baseline DEAN, and so is the loss function.

4.4 Fairness Constraint on Loss Function

Another method to enhance the fairness of DEAN is to extend the original loss function Equation 8 by adding a regularization term, which penalizes dependence between the model's prediction and a protected feature $p \in \{0, 1\}$. While the original loss continues to minimize the deviation from the target value 1, the additional term penalizes covariance between the output and the protected feature. The new loss function looks like the following:

$$\mathcal{L} = \sum_{x \in X_{\text{train}}} |f(x) - 1| + \lambda_{\text{fair}} \left|\text{Cov}(y, p)\right| \tag{12}$$

This added covariance based fairness regularization measures the linear dependence between the model output $y = f(x)$ and the protected attribute p. The function behind $\text{Cov}(y, p)$ is further explained following:

$$\text{Cov}(y, p) = \frac{1}{\|X_{\text{train}}\|} \sum_{x \in X_{\text{train}}} (f(x) - \bar{y})(p_x - \bar{p}), \tag{13}$$

where

$$\bar{y} = \frac{1}{\|X_{\text{train}}\|} \sum_{x \in X_{\text{train}}} f(x), \qquad \bar{p} = \frac{1}{\|X_{\text{train}}\|} \sum_{x \in X_{\text{train}}} p_x. \tag{14}$$

A covariance close to zero implies, that being part of the protected group has little influence on the the output of the model . λ_{fair} is a parameter used for this model. Higher values, strengthen the enforcement of fairness for the model. With this addition to the loss function, DEAN now minimizes the original loss function and the newly added, creating less bias against the protected group.

4.5 Dataset

For this thesis two commonly used datasets were used, in addition with some synthetic dataset. The two common used datasets are COMPAS, which was already presented in this thesis and largely analyzed. The second dataset used in this thesis Adult dataset, published by the UCI Machine Learning Repository (Dua and Graff, 2019). It contains socio-economic information, like age, education level, hours worked per week, occupation and other demographic variables. The machine learning task with this dataset is it to determine whether an individual has an annual income exceeding 50.000 USD. Adult dataset is with COMPAS a standard benchmark in fairness research, because several of the attributes, for example gender and race, hold structural imbalances, which can lead to a biased model, if not addressed by constraints. For that reason the dataset was chosen for this thesis, since it represents a real world scenario, in which fairness considerations are essential.

5 Results

COMPAS dataset

Adjustment	AUC-ROC	ΔAUC	Fairness	ΔFairness
DEAN (Baseline)	0.921	0.000	0.577	0.000
DEANpruning ($\tau = 0.0125$)	0.887	-0.034	0.566	-0.011
DEANpruning ($\tau = 0.025$)	0.875	-0.046	0.567	-0.010
DEANpruning ($\tau = 0.05$)	0.911	-0.010	0.574	-0.003
DEANpruning ($\tau = 0.10$)	0.915	-0.006	0.578	+0.001
DEANloss ($\lambda = 0.20$)	0.914	-0.007	0.577	0.000
DEANloss ($\lambda = 0.35$)	0.907	-0.014	0.578	+0.001
DEANloss ($\lambda = 0.50$)	0.911	-0.010	0.582	+0.005
DEANloss ($\lambda = 1.00$)	0.911	-0.010	0.579	+0.002

Table 2: AUC-ROC and Fairness for DEAN (baseline), pruning-based variants and DEANloss on COMPAS dataset, including difference to the DEAN-Baseline.

Adult dataset

Adjustment	AUC-ROC	ΔAUC	Fairness	ΔFairness
DEAN (Baseline)	0.580	0.000	0.469	0.000
DEANpruning ($\tau = 0.0125$)	0.557	-0.023	0.441	-0.028
DEANpruning ($\tau = 0.0250$)	0.575	-0.005	0.454	-0.015
DEANpruning ($\tau = 0.0500$)	0.576	-0.004	0.454	-0.015
DEANpruning ($\tau = 0.1000$)	0.568	-0.012	0.457	-0.012
DEANloss ($\lambda = 0.20$)	0.573	-0.007	0.463	-0.006
DEANloss ($\lambda = 0.35$)	0.573	-0.007	0.474	+0.005
DEANloss ($\lambda = 0.50$)	0.568	-0.012	0.453	-0.016
DEANloss ($\lambda = 1.00$)	0.596	+0.016	0.502	+0.033

Table 3: AUC-ROC and Fairness for DEAN (baseline), pruning-based variants and DEANloss on Adult dataset, including differences relative to the DEAN baseline.

The evaluation of the different DEAN variants on both COMPAS and Adult datasets clearly show a persistent relationship between the fairness adjustments and their changes in the predictive performance of the model. Across all experiments, whether COMPAS, Adult or the synthetic generated datasets, increasing the fairness constraints, either by lowering the thresholds τ to enforce stronger fairness submodels or by increasing the regularization weights λ to increase the penalty, they all lead to a slight improvement in fairness values, often along an equal amount of decrease in ROC-AUC.

5.1 Pruning results

In the pruning method, fairness improves with stricter threshold, because more sub-models, that illustrate biased behavior are discarded from the ensemble. This can be seen in both datasets examples. Looking at a small threshold ($\tau = 0.0125$ or $\tau = 0.025$), the fairness is consistently reduced, closing to 0.5. But at the same time the ROC-AUC tends to lower more. This observation indicates that pruning with strict parameters removes model diversity, that otherwise was still important for detecting anomalies. The more moderate pruning constraints ($\tau = 0.05$ or $\tau = 0.10$) present a better balance between enhancement of the fairness value and the maintaining a competitive ROC-AUC score.

5.2 Regularization results

In regularization based method, the fairness penalty influenced by the factor λ, directly imposes impact to the optimization dynamics. Larger values, brings the model to learn less statistically dependent representation based on the protected feature. As with the pruning method, increasing the strictness of the fairness constraint, the fairness metric improves. Also similar to the pruning method, with overly strong fairness constraints, the model tends to reduce its ROC-AUC. Contrary to the expected results, what can be examined in Adult dataset, at $\lambda = 1.0$ the ROC-AUC is increased when compared to the base model. This might be due to noise, but it may also suggesting, that the base model unintentionally exploited spurious correlations related with the protected feature.

6 Conclusion and Future Work

The results clearly show the fairness problem presented in subsubsection 2.5.2. DEAN is not an exception to the trade-off problem, where stronger fairness constraints result in reduced bias, but at the same time decrease the model's anomaly detection performance. The model has less possibility to use group specific information, that may correlate with the underlying anomaly structure. The results of both techniques show, that it is possible to reduce bias in the base DEAN model. The algorithmic adaptations counteracted dependencies and surpassed results, which would otherwise resulted in disproportionately higher anomaly scores to minorities.
However, these improvements must be seen with a broader view, about the context of fairness in real world data. It is impossible to achieve perfect fairness, because most real datasets are inherently bias, because they are collected in a biased world through structural, historical and institutional constraints. Given a completely unbiased model, it is impossible for the model to ignore the constraints dictated by biased data distribution. Precisely for that reason, it is important to monitor the model behavior in order to prevent harmful patterns from being reproduced.
Future work should investigate in methods to remove bias directly in data-generating process or how to process it. This thesis can also be further improved with other variants for DEAN, since it is easily adjustable. For example, instead of pruning models, give bias models lower weight in the aggregation function than less bias models. It is also possible to combine different variants, which should result into an interesting less bias model.

References

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May). *Machine bias: Risk assessments in criminal sentencing*. ProPublica. Retrieved November 16, 2025, from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Caton, S., & Haas, C. (2024). Fairness in machine learning: A survey. *ACM Computing Surveys*, *56*(7), 1–38. https://doi.org/10.1145/3616865

Du, H., & Zhang, Y. (2021). Network anomaly detection based on selective ensemble algorithm. *Journal of Supercomputing*, *77*, 2875–2896. https://doi.org/10.1007/s11227-020-03374-z

Dua, D., & Graff, C. (2019). Uci machine learning repository: Adult data set. https://archive.ics.uci.edu/dataset/2/adult

Gursoy, F., & Kakadiaris, I. A. (2022). Equal confusion fairness: Measuring group-based disparities in automated decision systems. *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, 137–146. https://doi.org/10.1109/ICDMW58026.2022.00027

Huang, H., Wang, P., Pei, J., Wang, J., Alexanian, S., & Niyato, D. (2025). Deep learning advancements in anomaly detection: A comprehensive survey. *IEEE Internet of Things Journal*.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Niu, X., Wang, L., & Yang, X. (2019). A comparison study of credit card fraud detection: Supervised versus unsupervised. *arXiv preprint arXiv:1904.10604*.

Sanami, S., & Aghdam, A. G. (2024). Aero-engines anomaly detection using an unsupervised fisher autoencoder. *2024 IEEE 63rd Conference on Decision and Control (CDC)*, 553–558.

StrongDM. (2024). *Anomaly detection: What it is and why it matters* [Accessed: 2025-10-25]. https://www.strongdm.com/blog/anomaly-detection

Wu, Z., Zheng, L., Yu, Y., Qiu, R., Birge, J., & He, J. (2024). Fair anomaly detection for imbalanced groups. *arXiv preprint arXiv:2409.10951*.

Xiao, F., Tang, X., & Fan, J. (2025). Fairness-aware anomaly detection via fair projection. *arXiv preprint arXiv:2505.11132*.

Zhang, J., Vukotic, I., & Gardner, R. (2018). Anomaly detection in wide area network mesh using two machine learning anomaly detection algorithms. *arXiv preprint arXiv:1801.10094*.

Eidesstattliche Versicherung

(Affidavit)

Benning, Marcelo

Name, Vorname
(surname, first name)

220846

Matrikelnummer
(student ID number)

☒ Bachelorarbeit
(Bachelor's thesis)

☐ Masterarbeit
(Master's thesis)

Titel
(Title)

Fairness enhancement methods in Anomaly Detection for DEAN

Ich versichere hiermit an Eides statt, dass ich die vorliegende Abschlussarbeit mit dem oben genannten Titel selbstständig und ohne unzulässige fremde Hilfe erbracht habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie wörtliche und sinngemäße Zitate kenntlich gemacht. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

I declare in lieu of oath that I have completed the present thesis with the above-mentioned title independently and without any unauthorized assistance. I have not used any other sources or aids than the ones listed and have documented quotations and paraphrases as such. The thesis in its current or similar version has not been submitted to an auditing institution before.

Dortmund, den 25.11.2025

Ort, Datum
(place, date)

MBenning

Unterschrift
(signature)

Belehrung:
Wer vorsätzlich gegen eine die Täuschung über Prüfungsleistungen betreffende Regelung einer Hochschulprüfungsordnung verstößt, handelt ordnungswidrig. Die Ordnungswidrigkeit kann mit einer Geldbuße von bis zu 50.000,00 € geahndet werden. Zuständige Verwaltungsbehörde für die Verfolgung und Ahndung von Ordnungswidrigkeiten ist der Kanzler/die Kanzlerin der Technischen Universität Dortmund. Im Falle eines mehrfachen oder sonstigen schwerwiegenden Täuschungsversuches kann der Prüfling zudem exmatrikuliert werden. (§ 63 Abs. 5 Hochschulgesetz - HG -).

Die Abgabe einer falschen Versicherung an Eides statt wird mit Freiheitsstrafe bis zu 3 Jahren oder mit Geldstrafe bestraft.

Die Technische Universität Dortmund wird ggf. elektronische Vergleichswerkzeuge (wie z.B. die Software „turnitin") zur Überprüfung von Ordnungswidrigkeiten in Prüfungsverfahren nutzen.

Die oben stehende Belehrung habe ich zur Kenntnis genommen:

Official notification:
Any person who intentionally breaches any regulation of university examination regulations relating to deception in examination performance is acting improperly. This offense can be punished with a fine of up to EUR 50,000.00. The competent administrative authority for the pursuit and prosecution of offenses of this type is the Chancellor of TU Dortmund University. In the case of multiple or other serious attempts at deception, the examinee can also be unenrolled, Section 63 (5) North Rhine-Westphalia Higher Education Act (*Hochschulgesetz, HG*).

The submission of a false affidavit will be punished with a prison sentence of up to three years or a fine.

As may be necessary, TU Dortmund University will make use of electronic plagiarism-prevention tools (e.g. the "turnitin" service) in order to monitor violations during the examination procedures.

I have taken note of the above official notification:*

Dortmund, den 25.11.2025

Ort, Datum
(place, date)

MBenning

Unterschrift
(signature)

***Please be aware that solely the German version of the affidavit ("Eidesstattliche Versicherung") for the Bachelor's/ Master's thesis is the official and legally binding version.**