A Novel Assurance Procedure for Fair Data Augmentation in Machine Learning

Samira Maghool*, Paolo Ceravolo and Filippo Berto

University of Milan, Department of Computer Science

Abstract

In addressing the limited availability of data for predictive purposes with machine learning, we are concerned with potential biases arising from dataset augmentation. Despite advanced algorithms to generate synthetic data that can preserve the original data distribution, challenges remain, including the risk of perpetuating social biases. Our approach uses a similarity network representation that treats each data point as a node and strategically generates synthetic points near it. A vector label propagation algorithm, complemented by an exponential kernel for adjusting link weights, accurately labels these synthetic points. The primary goal is to reduce the system's dependence on sensitive features without excluding them, thereby avoiding the risk of exacerbating biases or reducing data variation. Implemented in a big data ecosystem, our methodology enables continuous evaluation in an evolving domain, effectively addressing the challenges of data scarcity with a fairness-aware approach.

Keywords

Machine Learning, Fairness, Similarity Network, Data Augmentation

1. Introduction

The widespread adoption of Machine Learning (ML) technologies across industries has ushered in a new era of data-driven decision-making [1]. While ML promises to increase efficiency and productivity, its application in decision-making processes presents a number of challenges, ranging from performance to regulatory compliance [2]. Regulatory frameworks, such as the European Union's proposed Artificial Intelligence Act, emphasize the importance of fairness and accountability [3]. Overcoming these challenges requires industries to establish comprehensive testing frameworks that evaluate ML models' performance, reliability, and generalization across multiple scenarios [4]. But developing frameworks for regulatory compliance is a complex task. As regulations evolve and new data becomes available, industries must establish mechanisms for continuously monitoring and updating ML models [5]. The different techniques used by designers to achieve specific properties in ML systems can conflict with each other. Fairness may come at the expense of accuracy, accuracy at the expense of transparency, and privacy compliance at the expense of explainability [6, 7].

In this paper, our contribution is to *explore the delicate balance between data augmentation and fairness* in tabular data, with the goal of developing a solution suitable for integration into

AIEB 2024: Workshop on Implementing AI Ethics through a Behavioural Lens | co-located with ECAI 2024, Santiago de Compostela, Spain

*Corresponding author.

samira.maghool@unimi.it (S. Maghool); paolo.ceravolo@unimi.it (P. Ceravolo); filippo.berto@unimi.it (F. Berto)

0000-0001-8310-2050 (S. Maghool); 0000-0002-4519-0173 (P. Ceravolo); 0000-0002-2720-608X (F. Berto)

a *continuous assurance framework*. ML model training is inherently data-hungry, requiring a significant amount of data for accurate results, often necessitating the expansion of the dataset through augmentation [8]. Traditional data augmentation techniques applied to tabular datasets focus on creating duplicates and ensuring their consistency with the data distribution. This is achieved by assigning values through random perturbations or by adhering to central tendency [9]. However, many established augmentation techniques do not explicitly incorporate fairness principles. If the training dataset contains social biases, these biases can persist in the augmented data. For instance, applying central tendency functions to features already biased due to social or demographic factors may inadvertently mirror and reinforce these biases in the augmented data, rather than mitigating them.

Addressing fairness in the training of AI models requires careful consideration of the complex issue of discrimination. Discrimination in the context of AI is not a universally applicable concept but is intricately tied to socially salient categories that have historically been subject to unfair and systematic disadvantage [10]. These categories are closely linked to specific subgroups within the population. We use *subgroup* to denote a group of individuals defined by a shared value of a demographic variable such as race, gender, sexual orientation, religion, disability status, etc [11]. In the legal frameworks of various states, certain categories are officially recognized as protected. The recognition of these socially salient categories underscores the need to consider sensitive demographic variables when training AI models. Removing or ignoring these variables doesn't automatically ensure fairness. In fact, it can lead to unintended consequences, as the model may still learn and replicate biases present in the training data even when explicit features are excluded [12]. Moreover, the impact of bias is not limited to the training process but extends to the broader deployment and use of AI systems [13]. Biases in training data can manifest themselves in biased predictions and decisions, perpetuating systemic inequalities. To address these challenges, a holistic approach to fairness is critical. This includes using fairness-aware algorithms, conducting thorough bias audits of training data, and continuously assessing the impact on different subgroups.

Therefore, the methodology we have adopted to create data augmentation techniques that respect the principles of fairness includes the following facets.

- A data augmentation method that can mimic the distribution of data redressing the balance between subgroups in a dataset. This brings to an increased representativeness of underrepresented subgroups.
- Guide the process utilizing the entire feature space, not excluding any variable of the dataset. This comprehensive exploration ensures a holistic consideration of the data without penalizing or favoring any subgroups.
- Use incremental procedures to align with continuous assessment frameworks updating as new data is collected. This incremental approach facilitates ongoing validation and adjustment of the augmentation process in response to evolving data sets that may contain different distributions.

One of the key points of our methodology is to represent the data set using a similarity network. Within this network, we consider each data point as a node and strategically generate synthetic data points in close proximity to the original data points. These synthetic points form connections and inherit features from their neighboring nodes. In previous studies [14] ,

we observed that this network generates a latent feature space capable of capturing grouping patterns and latent associations within the dataset. The latent feature space is crucial for preventing classifiers from being biased by a restricted set of features, reducing the risk of segmenting the data based on sensitive features. In this work, we extend the capabilities of our methodology by using the density of the network to identify less represented groups within the dataset. Generating synthetic data near these groups we naturally balance the representation of different groups in the dataset. To ensure the accurate labeling of these synthetic points, i.e. their association with the target labels to be predicted by a trained ML model, we use a vector label propagation algorithm [15] , complemented by an exponential kernel for fine-tuning the link weights. Importantly, we avoid excluding any features, as such exclusions can worsen biases and compromise data diversity. By incorporating all features into the network, our approach fosters a comprehensive understanding of the dataset's complexity, resulting in a more robust and unbiased representation.

To ensure this continuous verification, we deployed the training and verification pipelines of the ML model using a platform in provided by the MUSA project [16]. This platform provides an edge-cloud continuum service infrastructure that delivers high-performance computing resources while assuring advanced non-functional properties such as privacy and application security. The solution proposed by [16] utilizes contract-based continuous verification using evidence gathered through transparent monitoring. The data pipelines described in this paper are executed on this platform, monitoring both their behavior and artifacts, and verifying that the model's performance and fairness properties are maintained. In this way, we have evaluated whether the dataset augmentation method we propose is able to enlarge the data points while guaranteeing fairness in the ML models trained after the augmentation.

By addressing the challenges associated with data scarcity, our method makes a significant contribution to the ongoing quest to develop unbiased ML systems that can effectively generalize across diverse and dynamic datasets. Through this research, we aim to foster the creation of ML models that not only exhibit fairness, but also maintain high levels of data variation and representation, ensuring their applicability to a wide range of real-world scenarios. Specifically, the paper is organized as follows. In Section 2, we discuss related work. In Section 3 we introduce the basic notions necessary to understand this work. Section 4 presents the proposed approach and explains how we incorporated it into an assurance procedure. Section 5 discusses in detail the experimental setup, the steps to be taken from data preparation to model description. In Section 6, we demonstrate the results we obtained by implementing our proposed method on a public dataset and evaluation metrics comparing the original dataset with the augmented version. Finally, after Discussion (Section 7), we conclude this paper.

2. Related Works

Three primary approaches to generating synthetic data for augmentation purposes have been identified in the literature [17]:

Generating synthetic data according to a given distribution. This approach involves generating synthetic data points that match the statistical properties and patterns expected in the target distribution [18, 19]. It uses knowledge of distribution properties, such as normal,

exponential, and chi-square, to generate synthetic samples without relying on actual data points. Many techniques are built by combining and balancing bootstrapping and perturbation steps. Bootstrapping involves generating synthetic data by duplicating existing data. Conversely, perturbation introduces controlled noise or randomization into real data. This step is also exploited to create synthetic data while maintaining privacy. By changing sensitive variables or details in the data, synthetic data can be generated that retains the statistical properties of the original dataset while making re-identification very difficult. A recognized limitation of duplication-based approaches is their tendency to lead to overfilling [20]. To mitigate this problem, state-of-the-art methods employ density measures to regulate the duplication ratio in different regions of the dataset [21, 22]. Our approach aligns with this strategy. However, to the best of our knowledge, none of the existing methods evaluate the duplication process using fairness-aware metrics, as we do.

Agent-based modeling (ABM). Addressing the challenge of simulating systems with many interacting parts that evolve over time, ABM is a robust method for generating synthetic data that effectively augments historical data [23]. Conceptually aligned with complex systems, ABM allows events to emerge from interactions among fully autonomous agents that follow rules, behaviors, and decision processes. This approach is particularly useful for understanding the dynamics of complex systems. Its strength lies in providing a high degree of realism and granularity, capturing emergent phenomena that result from the interactions of individual agents. Indeed, it has historically been used in the natural sciences [24, 25]. It is also very effective as an incremental learning technique, where iteration after iteration refines the global knowledge of the system based on information acquired at the local level [26]. However, ABM also has some limitations. Developing accurate ABMs can be complex, requiring careful calibration of parameters, and the computational intensity of simulations can be challenging in terms of computing power and time [27]. Validation and verification of ABM results against real-world data can be difficult due to the dynamic and emergent nature of the results [28]. In addition, ABM often requires detailed data on individual agents, and its sensitivity to parameter changes highlights the importance of robust sensitivity analysis [29]. Our method uses an ABM algorithm but fully exploits the feature space of the target detaset without relying on external or global data, and it is calibrated on fairness metrics.

Generative Adversarial Networks (GANs). Generative Adversarial Networks (GANs) play an important role in the synthetic data generation landscape, enhancing the ability to generate data that not only has statistical similarities, but also has visual and contextual similarities to real-world data [30]. At the core of GANs is the paradigm introduced by Goodfellow et al.[31], which encapsulates the concept of adversarial training. It involves a dynamic interplay between two key components: the generator and the discriminator. The generator is tasked with creating synthetic data, attempting to emulate real-world data distributions. At the same time, the discriminator acts as a classifier, distinguishing between real and synthetic data. Through iterative training, the generator refines its ability to generate increasingly realistic data, while the discriminator refines its ability to distinguish between real and synthetic instances. The adversarial nature of this training process creates a feedback loop that forces both the generator and the discriminator to continually improve. Despite the effectiveness of GANs in generating realistic data, they are susceptible to mode collapse, a limitation in which the generator produces a limited variety of samples. This can prevent the generation of extreme or

diverse data points, limiting the overall effectiveness of the synthetic data generation process. To address the challenge of mode collapse and increase the diversity of the generated data, various strategies have been explored. One common approach is to modify the dataset or learn a fair distribution to mitigate bias [32]. In the context of mitigating unfairness using adversarial learning, it is important to note that the selection and treatment of sensitive features play a critical role. Typically, studies in this area require the predefinition of sensitive features, and the algorithm systematically addresses the mitigation of one sensitive feature at a time. This is a significant limitation that does not apply to our approach, since multiple sensitive features are considered when generating new examples.

While the potential of augmentation algorithms using synthetic data generation is promising, it is important to acknowledge and address criticisms that have emerged in the literature. Specifically, concerns have been raised about the potential for these algorithms to inadvertently promote polarization and biased information [33]. Such unintended consequences have the potential to undermine trust in ML systems [34]. This critical finding underscores the need to develop augmentation strategies that go beyond the mere consideration of statistical properties. This involves designing augmentation algorithms that are not only effective in enhancing the diversity of the dataset but also sensitive to and mitigate potential biases.

3. Background notions

The basic concepts necessary to understand the methodology presented in this paper are explained in this section.

3.1. Data Generation using Agent-Based Modeling

In Agent-based modeling (ABM), each agent has distinct characteristics and behaviors, allowing the exploration of emergent phenomena from their collective interactions. ABM emphasizes capturing heterogeneity among agents and local interactions that drive overall system dynamics. ABMs study the interactions among many independent decision-making agents in a discrete spatiotemporal environment. The model operates through discrete time steps, during which agents update their states based on predefined rules and responses to their local environment. Agents in ABM can exhibit a wide range of behaviors, from random actions to adaptive decisions based on prevailing conditions. This flexibility allows the modeling of diverse scenarios.

While agent-based modeling (ABM) often involves specific rules tailored to the characteristics of the modeled system, a general framework can be outlined to illustrate the workings of ABM. Let S_t be the state of the system at time t, A_t represent the actions taken by individual agents at time t, and E_t denote the environment or context at time t. The dynamics of the system can be expressed through the following formula:

$$S_{t+1} = F(S_t, A_t, E_t) \tag{1}$$

In this formula: S_{t+1} is the state of the system at the next time step $(t + 1)$. F is a function that describes how the state of the system at time $t + 1$ is determined by its state at time t, the actions of agents (A_t), and the environment (E_t).

This formula shows the future state of the system is affected by the current state, the actions of the agents, and the environment. The specific form of the function F will vary depending on the characteristics and rules governing the agents and their interactions in the modeled system.

3.2. Agent-Based Vector-Label Propagation algorithm (AVPRA)

Graph-based semi-supervised learning hinges on the idea that adjacent nodes are likely to share similar labels, an assumption becomes more valid referring to the known homophily concept in social network analysis. Label Propagation (LP) algorithms embody this concept by assigning labels to unlabeled nodes based on the similarities among their neighboring nodes. Drawing inspiration from epidemic spread research, these algorithms compile information spread through node contacts to define an individual's features, rather than just assigning a single label. This approach not only uncovers the structural details of the graph but also ensures a balanced representation of the existing features. Expanding on LP algorithms, Label-Vector Propagation algorithms assign a vector of labels to each node instead of one. Our earlier study [35] introduced the application of an agent-based algorithm in the Vector-label Propagation technique. This algorithm employs Vector Labels (VLs) to disseminate weighted labels across the network via edges. A notable aspect of AVPRA is its unique VL size, enhancing its utility in subsequent ML applications. Additionally, it features normalized coefficients within the vector, summing to 1, thereby eliminating biased learning due to uneven feature distribution in the graph. The rationale behind leveraging Agent-Based Model (ABM) is its potential to admit a large number of parameters in modeling the propagation phenomena. ABM adopts simply understandable rules in propagating features through links, tunable to the desired phenomena and output [23]. In this current work, we exploit the aforementioned [`doubleblind`] technique to the correct assignment of labels to the synthetic data points.

3.3. System Fairness

Work on training ML systems that result in fair decisions have defined several useful measurements for fairness: *Demographic Parity*, *Equality of Opportunity*, and *Equality of Mis-Opportunity*. These can be imposed as constraints or incorporated into a loss function to mitigate disproportionate outcomes in the system's output predictions with respect to a protected demographic, such as gender. Prior work in the same scope can be classified into three groups depending on the approach applied to remove bias: **Pre-processing algorithms** [36, 37], **In-processing algorithms** [38, 32] and, **Post-processing algorithms** [39, 40, 41].

3.4. Dataset Complexity

Given the fact that performance metrics do not provide a full insight into the level of complexity of a classification problem that an algorithm has to deal with, we also need to investigate whether the data set itself allows a clear separation between classes. In other words, to investigate how complex a classification problem is in its and whether some features play a more effective role than others, we need to consider other criteria. In this context, we emphasize that if a class is clearly separated by a subset of sensitive features, this could be a sign of biased or discriminatory behavior imposed on the system by the data set. According to [42], these measures are

categorized as “feature-based", “Linearity", “Neighborhood", “Network", “Dimensionality", and, “Class imbalance" measures, and introduce several metrics for this purpose. In this paper, we evaluate some of them by comparing the complexity of the original dataset with the augmented version.

4. The proposed Approach

In this section, we explain our methodology and discuss how it aligns with an infrastructure designed for continuous assurance verification.

4.1. Assuring Fair Data Augmentation

The starting point of our methodology is to redefine the data set through the lens of a similarity network, following the methods described in Section 5.2. This provides us with a view of the structural organization of the dataset that we can exploit in generating synthetic data and interpreting our results. The synthetically generated data points form connections and inherit features from their neighboring nodes. By favoring the generation of synthetic data in close proximity to groups with lower density, we inherently balance the representation of different subgroups in the dataset. The AVPRA algorithm ensures accurate labeling of these synthetic points, as explained in Section 5.3. In addition, the edges of the similarity network provide us with a latent feature space that protects the classifiers from bias due to a limited set of features. This, in turn, mitigates the risk of data segmentation based on sensitive features, as demonstrated in Section 6. Sections 5.4 provide a detailed explanation of the procedure we followed. Another key point of our proposal is to exploit the incremental nature of our ABM algorithm to insert our method into a continuous assurance infrastructure. The metrics evaluated by our algorithm can be inserted into a library of assurance tests periodically executed by the system to monitor the reliability of ML models, as explained in Section 4.3.

4.2. Infrastructure

Our method has been included in the MUSA platform. This platform provides a 5G-enabled edge-cloud continuum infrastructure, supporting processes with continuous assurance techniques for advanced non-functional requirements. The platform provides multiple cloud services, including storage, computing, and data pipeline management, guaranteeing strong non-functional properties, such as data privacy and locality, security, and isolation, while providing high availability and performance [16, 43].

In this paper, we integrate our synthetic data generation methodology with some of the services provided, in particular the storage, computing, pipeline and assurance platforms, as described in Figure 1. We used services from the Apache stack because of their ease of interoperability and open-source nature.

Storage. The platform’s storage solution is based on Apache Hadoop and can be accessed through a RESTful API using the WebHDFS protocol. The system also features access control and logging, transparent replication and snapshotting, data lineage metadata, and high availability.

Computing. The computing platform utilized is based on Kubernetes and offers containerized

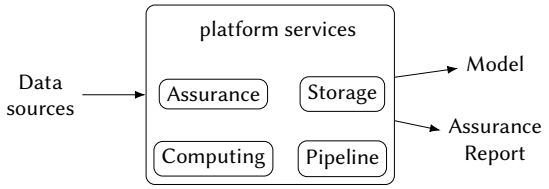


Figure 1: The major components of platform used in this work.

environments for Python-based data analysis. Additionally, the platform provides highly parallelized execution solutions through Apache Spark and Trino clusters, in combination with GPU-accelerated workflows.

Pipeline. This research utilized the Apache Airflow pipeline platform. Apache Airflow is an extensible DAG-based execution scheduler that enables the coordination of analysis pipelines and integrates with numerous services and components through Python bindings and APIs.

Assurance. The platform's assurance component monitors the services in use and continuously verifies a set of required properties. To identify which checks need to be executed, the assurance service uses annotations on the pipeline DAG and a set of templates for common checks. The solution is easily extendable, allowing us to implement custom verification tasks within the pipeline itself.

4.3. Assurance and Certification

The verification and certification of the models' properties is implemented by a continuous assurance process that manages the data analysis and model training tasks in the edge-cloud continuum. This process checks evidence collected from the tasks against a set of formal contracts, inferring whether the associated properties are holding. We first define a set of desired non-functional properties that we want to verify. These are abstract concepts that indicate an inherent behavior or quality of the ML model, in this case, data anonymity, accuracy, and fairness. In order to ensure the validity of a property, it is necessary to gather evidence on the task. This can be achieved by collecting metrics that provide an objective view of the system's state and behavior. Some of these metrics are already available through the execution platform, such as data access logging, execution tracing, resource usage, and source code security and quality analysis. Others, such as the model's scoring metrics (accuracy, fairness), are task/model specific and need to be implemented by the data pipeline. The platform continuously collects measurements of the pipeline's aspects and saves them in a time-series optimized storage solution. Monitoring can be continuous (at a certain frequency or event-based) or on-demand, only running when required by the assurance process. The next step is to define the contracts for the process, each verifying a particular property. These contracts interact with the metrics interface, querying for previously collected or on-demand evidence. An *Accredited Lab*, a trusted entity with full access to the evidence, executes the contracts and provides a report on the inferred information and property status. . In this case, the model is an artifact of the data pipeline and its properties are not expected to change after its release. Therefore, the model and its certificate have tightly linked lifetimes, and certificate revocation is unnecessary.

5. Experimental Setup

5.1. Dataset

In order to evaluate the proposed methodology, we used a public data set of the Curriculum Vitae (CV) of 301 employees[1] which contains both numerical and categorical features.
Payment Categorization: Given that the payment rate is a continuous value in the dataset, but our objective is to address a classification problem, we categorize the `Payment` using automated bin selection. This method determines the number of categories based on maximizing instances within a bin. Consequently, the number of bins varies according to the data distribution, potentially resulting in a multi-class classification problem.

5.2. Similarity measurements

Graph construction from datasets based on instances' similarities can be implemented using various mechanisms to measure similarity values. In this paper, to create the weighted network $\mathcal{N}$ = **(V, E)**, where $\mathbf{V}$ corresponds to the network nodes (vertices), and $\mathbf{E}$ to the links (edges) among them, we use *Gower distance*($\mathcal{GD}$) that is often used for dealing with heterogeneous data [44]. For a set of k variables $\mathbf{G}$=$\{g_1, g_2, ..g_k\}$, the similarity $\mathbf{S}'$ between two patients $\mathbf{a}$ and $\mathbf{b}$ is then defined as the average of the *Gower Similarity* ($\mathcal{GS} = 1 - \mathcal{GD}$) for each of the variables:

$$S'_{a,b} = \frac{\sum_{i=1}^{k} S(a, b, g_i)}{k} \tag{2}$$

The creation of a similarity network requires choosing the threshold value of pairwise similarities, ϵ, that defines the existence (non-existence) of links between points. To make this issue less crucial, kernel matrices that are able to capture the topological characteristics of the underlying graph may provide a solution. Edge weights are represented by an $n \times n$ similarity matrix $\mathbf{W}$ where $\mathbf{W}(\mathbf{a}, \mathbf{b})$ indicates the similarity between data points $\mathbf{a}$ and $\mathbf{b}$. We use a scaled Exponential kernel (Ek) to determine the weight of the edges:

$$W(a, b) = exp(-\frac{\mathcal{GD}^2(a, b)}{\mu\epsilon_{a,b}}), \tag{3}$$

where according to the [45], μ is a hyperparameter that can be empirically set and $\epsilon_{a,b}$ is used to eliminate the scaling problem by defining:

$$\epsilon_{a,b} = \frac{mean(\rho(a, N_a)) + mean(\rho(b, N_b)) + \rho(a, b)}{3}, \tag{4}$$

where $mean(\rho(a, N_a))$ is the average value of the distances between a and each of its k NNs. The range of $\mu = [0.3, 0.8]$ is recommended by [45].

[1] https://rpubs.com/rhuebner/hrd_cb_v14

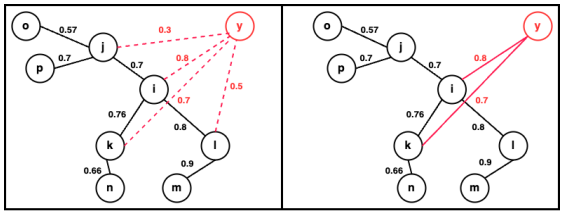


Figure 2: Schematic link creation process between synthetic nodes (red), recently added to the similarity network constructed by original nodes (black). On the left, the similarity of the newly added node (y) with the rest of the network is calculated. Possible links are demonstrated by the dashed red line. On the right, after considering the average link's weight of each original node with its original neighbors ((i,j), (i,l), (i,k)) links (y,i) and (y,k) are remained and the rest ge excluded. This process takes place at the end of each round of adding synthetic nodes, and revising the links.

5.3. Implementation of AVPRA

Leveraging this algorithm, each node(edge) of the graph is mapped into a latent space by a d-dimensional Vector-Labels (VL) containing normalized coefficients of the target labels in the dataset ($\sum_l \mathbf{VL}_i[l](t) = 1$). In this paper, we have altered the aggregation function as Eq. 5, to capture the weight of links, $\mathbf{W}_{ij}$, we calculated before using EK.

$$b(l) = \mathbf{VL}_i[l](t) = w_1 \mathbf{VL}_i[l](t) + w_2 \sum_{j \in \Gamma(i)} \mathbf{W}_{ij} \mathbf{VL}_{j \in \Gamma(i)}[l](t-1) \tag{5}$$

According to this algorithm, at each time step t, $b(l)$, the belonging coefficient of an element of the $\mathbf{VL}_i[l](t)$, can be updated by aggregating the k neighbors' $\mathbf{VL}_{j \in \Gamma(i)}[l](t-1)$. Where w_1 and w_2 are the weight of currently assigned labels l of node i and the weight of the neighbors $\Gamma(i)$, respectively. In a basic scenario, $w_1 = w_2 = \frac{1}{\Gamma(i)+1} = \frac{1}{k+1}$, hence, for all the common elements in $\mathbf{VL}_i$ and $\mathbf{VL}_j$ vectors, the values of the given elements l increase unconditionally and will be normalized to 1 by the inverse of the cardinality of $\Gamma(i)$.

5.4. Model Description

Considering a given dataset, aimed to be augmented, we initiate our approach by creating a similarity graph out of N data points. The quantified similarity values are calculated out of the feature space as previously described in Sec. 5.2 to construct the edges among nodes in the similarity graph. Synthetic data generation will take place repeatedly until the assurance condition is provided. Regarding the calibration of Fairness metrics, *Equal Opportunity* and *Equal mis-opportunity*, as assurance criteria, we need to continuously evaluate them after adding some synthetic data points. At each round, we add N_{syn} nodes with features $\{\mathbf{f_i}\}$ adopted randomly from the same feature distribution as the original dataset without labels assigned. At each round, by measuring the similarity of one synthetic data point with all the original data points, we create links only if the similarity value is greater than the average weights of the original nodes with original neighbors (see Fig. 2). Having the updated network with newly generated synthetic nodes and links, we initiate the AVPRA algorithm, diffusing the labels from the original data points to the other part of the similarity network. For each node

$i \in N$, the labels are vectorized to the maximum dimension of the number of target labels. For instance, if the labels contain three different categories such as L_i where for node i, after uniquely vectorizing the labels, $VL_i = [b_1, b_2, b_3]$ where $b_k = 1$ if the $L_i = k$ and the rest are equal to zero. As the result of the AVPRA algorithm using the aggregation function in Eq. 5 after multiple iterations, the VL of synthetic nodes gets updated. The most weighted label of the VL will be adopted as the label of the synthetic node. At the end of each round of augmentation by $x\%$, the fairness metrics evaluation takes place in order to ensure the augmentation process is not replicating the bias in the synthetic data set.

To assess the equilibrium between accuracy, transparency, and fairness, we employ a range of evaluation metrics in our analysis.

Considering a predictive task based on the dataset, we propose implementing a classification model to predict the payment levels of employees. To facilitate the analysis, we explore various scenarios, differing by dataset representation as **Original data set**; this is the dataset after preprocessing and categorization of payments into *Pay rate* levels. **Balanced data set**; a preprocessed dataset where we have addressed the class imbalance in the *Pay rate* levels. **Data set with balanced groups**; a preprocessed dataset that is balanced considering protected features for discriminated groups. In our experiment, we balanced by the `Gender` and `Hispanic/Latino` features. **Augmented data using our proposed method;** Leveraging the proposed algorithm, we have augmented the data set by 100 percent.

6. Experimental results

After setting up the workflow and infrastructure as extensively described in the Section. 5, the evaluation metrics on the given dataset are presented as follows:
Accuracy: in order to explore the fidelity of our proposed method used for the augmentation purposes, in preserving the dataset still informative for the prediction purposes, we trained two classifiers predicting the *Pay rate* of employees in the given dataset. Table 1, represents the implemented models on different dataset represenations. The original dataset acquires the $F1$ score about 67% for both Classifiers. In the second trial, the data set classes are balanced using the SMOTE algorithm, we witness an increase in the performance of the prediction compared to the original data. Regarding the bias reduction, introduced by the sensitive features into the dataset affecting the output of the predictors, we have tried to balance the dataset (by adding) according to the sensitive features such as gender and ethnicity in our dataset. As a result, in comparison to the original dataset, the balanced datasets are slightly more accurate. The above mentioned models are implemented using 10-fold cross-validated data.

Transparency: Investigating the importance of features in the output of a predictive model leads to pointing out those sensitive features playing a role in the model. Fig. 3 is presenting the SHAP values resulted by an RF algorithm on three datasets; original, Balanced classed using SMOTE, and, augmented data using our method. Considering these plots, we witness the depolarised distribution of data point values for sensitive features using in the augmented dataset in contrary to the other datasets.

Dataset representations	Random Forest	XGBOOST
Original data set	0.666 ± 0.056	0.680 ± 0.059
Original data set with balanced classes	0.842 ± 0.038	0.817 ± 0.046
Original data set with balanced protected feature (gender)	0.700 ± 0.071	0.686 ± 0.067
Original data set with balanced protected feature (ethnicity)	0.783 ± 0.049	0.777 ± 0.055
Augmented dataset	0.645 ± 0.043	0.667 ± 0.015

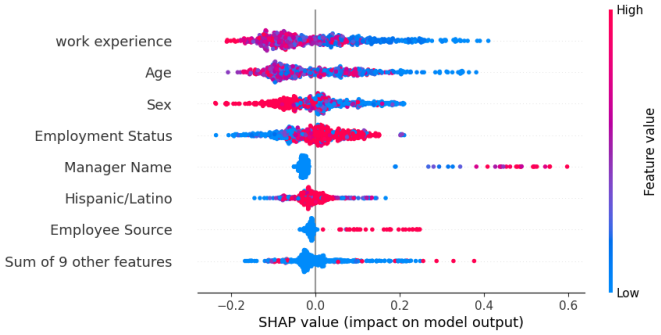
Table 1
The mean $\pm$ standard deviation of F1 weighted score for 10-fold Cross-validated data using XGBOOST and RF classification algorithms on the CV data set considering original and class-balanced data set; on balanced group data set respectively gender and ethnicity; on the augmented dataset by our proposed method.

Data set	Groups	True Positive Rate	False Positive Rate	(approx.)Equal Opportunity	(approx.)Equal mis-Opportunity
original	Privileged group	0.252 ± 0.33	0.077	0.35	0.17
	Unprivileged group	0.27 ± 0.199	0.2435		
augmented	Privileged group	0.42± 0.15	0.0790	0.17	0.08
	Unprivileged group	0.60± 0.06	0.162162		

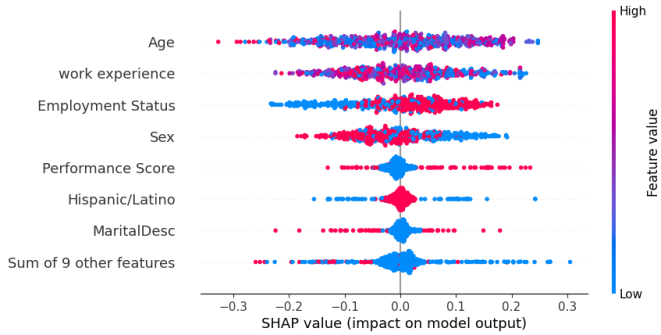
Table 2
Fairness metrics, true positive rate (recall) and False positive rate calculated for privileged and unprivileged groups. The metrics are reported for original and augmented data sets after implementing RF algorithm for predicting Pay Rate.

Classification complexity measurments: In order to have an insight on the dataset distribution specification, for measuring how a classification task could be difficult to implement on a given dataset we have calculated some of the common metrics comparing the original dataset before and after augmentation process. The $C2$ measure that is a well known index computed for measuring class balance demonstrate an improvement by 60% in balancing the classes using our method. *Clustering Coefficient* and *density* of networks before and after augmentation do not tangibly change. The Collective Feature Efficiency ($F4$) measure to get an overview of how the features work together [46], lower values of $F4$ indicate that it is possible to discriminate more examples and, therefore, that the problem is simpler. Our measurements shows $F4$ increased by 34% which means the classification problem is more difficult in augmented data due to debiasing process as we expected. Another crucial metric to consider is Dimensionality Measures which are indicatives of data sparsity. Measuring $T4$ shows that the dataset sparsity stays the same as original dataset with negligible changes.
Fairness: In order to prove that our augmentation method is improving the fairness by creating synthetic data points, we have measured Equal Opportunity and Equal Mis-Opportunity as mentioned in the Supplementary. Table 2 demonstrates the measured metrics on the original data and after augmentation in which the bias has clearly improved in the extended version.

(a)



(b)



(c)

Figure 3: The most important features using SHAP values, extracted from the RF classifier predicting the pay rate on the 3(a)) original data set, 3(b)) Balanced data set using SMOTE, 3(c)) Augmented dataset using our proposed algorithm by 100%.

7. Discussion and Conclusion

In this paper, we focus on the implementation of an innovative assurance procedure for fair data augmentation. To achieve this goal, we present a comprehensive methodology and infrastructure that provides an environment for the seamless implementation of our method. Throughout this process, we continuously evaluate certification criteria, such as the accuracy of ML models or

fairness metrics (see Section 5 for experimental setup details). Our methodology is based on computing the similarity, quantified and tuned by a kernel function, to construct a similarity network of instances from a given dataset. The generation of synthetic nodes, each with features randomly drawn from the original feature distribution, is done by creating links to more similar nodes in the initially constructed network. Building on our previous work on the AVPRA algorithm for semi-supervised learning, we assign labels to the unlabeled synthetic nodes after several iterations of the AVPRA method. This involves aggregating and propagating labels across links based on their weights. We provide a detailed discussion of the proposed approach, outlining all the necessary steps, in Section 4. After establishing the workflow (see Section 5), we perform a comprehensive evaluation of several metrics on different data representations and report the results in Section 6. Our results confirm the effectiveness of our approach in augmenting datasets while ensuring fairness metrics. We evaluate the accuracy of predicting employee pay rates in several scenarios, including the original dataset, an original dataset with balanced classes, an original dataset with balanced sensitive characteristics (Sex and ethnicity), and an augmented dataset using our methodology. In particular, we observe an increase in accuracy in the balanced dataset, both in terms of classes and subgroups. However, for the augmented dataset, the accuracy remains stable compared to the original data, possibly due to the debiasing process that decentralizes the focus of the classification tasks. Taking transparency into account, we visualize the crucial, potentially sensitive features by using the SHAP values, as shown in Figure 3. The figures show biases in the output of wage rate predictions with respect to *Sex* and *Age* features. Even with a balanced data set, the sensitivity of the features remains. However, in the augmented data using our methodology, the biases in the feature values are significantly improved. In addition, we examine measures of dataset complexity in the augmented dataset and observe improvements towards a more fair dataset in factors such as class balance and collective feature efficiency. Stability is observed in other factors, including *Clustering Coefficient*, *Density*, and *T4*. In the final stages of our evaluation process, we periodically measure fairness levels after each round of synthetic data point generation. This is done by calculating metrics such as *Equal Opportunity* and *Equal Mis-Opportunity*, which correspond to *True Positive* and *False Positive* rates, respectively. We made a binary classification for two wage levels, considering the privileged and non-privileged groups. We considered `Gender` and `Age` as two discriminating features, defining the privileged group as `Sex`: *male* and `Age` $\leq$ *40*, while considering the remaining instances as an unprivileged group. Table 2 shows that the fairness metrics have improved in the extended dataset (by about 50%).

8. Conclusion

Implementing accurate ML models requires a handful of dataset. Data scarcity may force a data scientist to adopt augmentation techniques for enlarging the original dataset. Concerning the possible biases in the original dataset, for instances imposed by data collector, the augmentation process may amplify the bias in the extended data set. To address this issue we have proposed a methodology based on the similarity network representation of dataset. Links in the similarity network, composed of original and synthetic nodes, are adjusted using exponential kernel functions mapping the similarity values of nodes smoothly to the weight of links. Our previously

proposed algorithm, `doubleblind`, is considered for labelling newly synthetic nodes in a semi-supervised fashion. The main objectives of this work as reducing the dependency of ML models on sensitive features while improving the fairness is achieved while evaluating by different metrics continuously in a big data ecosystem.

References

[1] M. D. Lytras, V. Raghavan, E. Damiani, Big data and data analytics research: From metaphors to value space for collective wisdom in human decision making and smart machines, International Journal on Semantic Web and Information Systems (IJSWIS) 13 (2017) 1–10.

[2] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, et al., An ethical framework for a good ai society: Opportunities, risks, principles, and recommendations, Ethics, governance, and policies in artificial intelligence (2021) 19–39.

[3] V. Almeida, L. S. Mendes, D. Doneda, On the development of ai governance frameworks, IEEE Internet Computing 27 (2023) 70–74.

[4] V. Riccio, G. Jahangirova, A. Stocco, N. Humbatova, M. Weiss, P. Tonella, Testing machine learning based systems: a systematic mapping, Empirical Software Engineering 25 (2020) 5193–5254.

[5] M. Anisetti, C. A. Ardagna, N. Bena, E. Damiani, Rethinking certification for trustworthy machine learning-based applications, IEEE Internet Computing (2023).

[6] H. Chang, R. Shokri, On the privacy risks of algorithmic fairness, in: 2021 IEEE European Symposium on Security and Privacy (EuroS&P), IEEE, 2021, pp. 292–303.

[7] B. Rastegarpanah, M. Crovella, K. P. Gummadi, Fair inputs and fair outputs: The incompatibility of fairness in privacy and accuracy, in: Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization, 2020, pp. 260–267.

[8] J. Ding, X. Li, X. Kang, V. N. Gudivada, A case study of the augmentation and evaluation of training data for deep learning, Journal of Data and Information Quality (JDIQ) 11 (2019) 1–22.

[9] K. Maharana, S. Mondal, B. Nemade, A review: Data pre-processing and data augmentation techniques, Global Transitions Proceedings 3 (2022) 91–99.

[10] B. Hutchinson, M. Mitchell, 50 years of test (un) fairness: Lessons for machine learning, in: Proceedings of the conference on fairness, accountability, and transparency, 2019, pp. 49–58.

[11] S. Barocas, M. Hardt, A. Narayanan, Fairness and Machine Learning: Limitations and Opportunities, MIT Press, 2023.

[12] F. Kamiran, T. Calders, Classifying without discriminating, in: 2009 2nd international conference on computer, control and communication, IEEE, 2009, pp. 1–6.

[13] D. Roselli, J. Matthews, N. Talagala, Managing bias in ai, in: Companion Proceedings of The 2019 World Wide Web Conference, 2019, pp. 539–544.

[14] S. Maghool, E. Casiraghi, P. Ceravolo, Enhancing fairness and accuracy in machine learning

through similarity networks, in: International Conference on Cooperative Information Systems, Springer, 2023, pp. 3–20.
[15] V. Bellandi, E. Damiani, V. Ghirimoldi, S. Maghool, F. Negri, Validating vector-label propagation for graph embedding, in: International Conference on Cooperative Information Systems, Springer, 2022, pp. 259–276.
[16] M. Anisetti, C. A. Ardagna, F. Berto, An assurance process for Big Data trustworthiness, Future Generation Computer Systems 146 (2023) 34–46. URL: https://www.sciencedirect.com/science/article/pii/S0167739X23001371. doi:10.1016/j.future.2023.04.003.
[17] J. Fonseca, F. Bacao, Tabular and latent space synthetic data generation: a literature review, Journal of Big Data 10 (2023) 115.
[18] G. Douzas, F. Bacao, Geometric smote a geometrically enhanced drop-in replacement for smote, Information sciences 501 (2019) 118–135.
[19] G. Menardi, N. Torelli, Training and assessing classification rules with imbalanced data, Data mining and knowledge discovery 28 (2014) 92–122.
[20] B. Krawczyk, Learning from imbalanced data: open challenges and future directions, Progress in Artificial Intelligence 5 (2016) 221–232.
[21] C.-T. Lin, T.-Y. Hsieh, Y.-T. Liu, Y.-Y. Lin, C.-N. Fang, Y.-K. Wang, G. Yen, N. R. Pal, C.-H. Chuang, Minority oversampling in kernel adaptive subspaces for class imbalanced datasets, IEEE Transactions on Knowledge and Data Engineering 30 (2017) 950–962.
[22] G. Douzas, R. Rauch, F. Bacao, G-somo: An oversampling approach based on self-organized maps and geometric smote, Expert Systems with Applications 183 (2021) 115230.
[23] C. M. Macal, M. J. North, Tutorial on agent-based modeling and simulation, in: Proceedings of the Winter Simulation Conference, 2005., IEEE, 2005, pp. 14–pp.
[24] T. J. Cocucci, M. Pulido, J. P. Aparicio, J. Ruíz, M. I. Simoy, S. Rosa, Inference in epidemiological agent-based models using ensemble-based data assimilation, Plos one 17 (2022) e0264892.
[25] M. B. Hooten, D. S. Johnson, E. M. Hanks, J. H. Lowry, Agent-based inference for animal movement and selection, Journal of Agricultural, Biological and Environmental Statistics 15 (2010) 523–538.
[26] D.-o. Kang, J. W. Bae, C.-H. Lee, J. Jung, E. Paik, Self-evolving agent-based simulation platform for predictive analysis on socio-economics by using incremental machine learning, in: Proceedings of the 2018 Winter Simulation Conference, 2018, pp. 4254–4254.
[27] F. Lamperti, A. Roventini, A. Sani, Agent-based model calibration using machine learning surrogates, Journal of Economic Dynamics and Control 90 (2018) 366–389.
[28] C. Bianchi, P. Cirillo, M. Gallegati, P. A. Vagliasindi, Validating and calibrating agent-based models: a case study, Computational Economics 30 (2007) 245–264.
[29] M. Riddle, C. M. Macal, G. Conzelmann, T. E. Combs, D. Bauer, F. Fields, Global critical materials markets: An agent-based modeling approach, Resources Policy 45 (2015) 307–321.
[30] R. Sauber-Cole, T. M. Khoshgoftaar, The use of generative adversarial networks to alleviate class imbalance in tabular data: a survey, Journal of Big Data 9 (2022) 98.
[31] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Advances in neural information processing systems 27 (2014).
[32] B. H. Zhang, B. Lemoine, M. Mitchell, Mitigating unwanted biases with adversarial learning,

in: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 2018, pp. 335–340.

[33] P. N. Howard, S. Bradshaw, B. Kollanyi, G. Bolsover, D. J. Trump, Junk news and bots during the french presidential election: What are french voters sharing over twitter in round two? comprop data memo 2017.4 / 4 may 2017, 2017.

[34] S. B. Naeem, R. Bhatti, A. Khan, An exploration of how fake news is taking over social media and putting public health at risk, Health Information & Libraries Journal 38 (2021) 143–149.

[35] V. Bellandi, P. Ceravolo, E. Damiani, S. Maghool, Agent-based vector-label propagation for explaining social network structures, in: International Conference on Knowledge Management in Organizations, Springer, 2022, pp. 306–317.

[36] F. Kamiran, T. Calders, Data preprocessing techniques for classification without discrimination, Knowledge and information systems 33 (2012) 1–33.

[37] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, S. Venkatasubramanian, Certifying and removing disparate impact, in: proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, 2015, pp. 259–268.

[38] Y. Zhang, J. Sang, Towards accuracy-fairness paradox: Adversarial example-based data augmentation for visual debiasing, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 4346–4354.

[39] F. Kamiran, S. Mansha, A. Karim, X. Zhang, Exploiting reject option in classification for social discrimination control, Information Sciences 425 (2018) 18–33.

[40] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, K. Q. Weinberger, On fairness and calibration, Advances in neural information processing systems 30 (2017).

[41] M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning, Advances in neural information processing systems 29 (2016).

[42] A. C. Lorena, L. P. Garcia, J. Lehmann, M. C. Souto, T. K. Ho, How complex is your classification problem? a survey on measuring classification complexity, ACM Computing Surveys (CSUR) 52 (2019) 1–34.

[43] M. Anisetti, N. Bena, F. Berto, G. Jeon, A DevSecOps-based Assurance Process for Big Data Analytics, in: 2022 IEEE International Conference on Web Services (ICWS), IEEE, Barcelona, Spain, 2022, pp. 1–10. URL: https://ieeexplore.ieee.org/document/9885738/. doi:10.1109/ICWS55610.2022.00017.

[44] J. C. Gower, Some distance properties of latent root and vector methods used in multivariate analysis, Biometrika 53 (1966) 325–338.

[45] B. Wang, A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains, A. Goldenberg, Similarity network fusion for aggregating data types on a genomic scale, Nature methods 11 (2014) 333–337.

[46] A. Orriols-Puig, N. Macia, T. K. Ho, Documentation for the data complexity library in c++, Universitat Ramon Llull, La Salle 196 (2010) 12.