




UnbiasedNets: a dataset diversification framework for robustness bias alleviation in neural networks

Mahum Naseer¹  · Bharath Srinivas Prabakaran¹ · Osman Hasan² · Muhammad Shafique³

Received: 12 November 2021 / Revised: 23 December 2022 / Accepted: 2 February 2023 /
Published online: 1 March 2023
© The Author(s) 2023

Abstract

Performance of trained neural network (NN) models, in terms of testing accuracy, has improved remarkably over the past several years, especially with the advent of deep learning. However, even the most accurate NNs can be biased toward a specific output classification due to the inherent bias in the available training datasets, which may propagate to the real-world implementations. This paper deals with the *robustness bias*, i.e., the bias exhibited by the trained NN by having a significantly large robustness to noise for a certain output class, as compared to the remaining output classes. The bias is shown to result from imbalanced datasets, i.e., the datasets where all output classes are not *equally represented*. Towards this, we propose the *UnbiasedNets* framework, which leverages K-means clustering and the NN's noise tolerance to diversify the given training dataset, even from relatively smaller datasets. This generates balanced datasets and reduces the bias within the datasets themselves. To the best of our knowledge, this is the first framework catering to the robustness bias problem in NNs. We use real-world datasets to demonstrate the efficacy of the *UnbiasedNets* for data diversification, in case of both binary and multi-label classifiers. The results are compared to well-known tools aimed at generating balanced datasets, and illustrate how existing works have limited success while addressing the robustness bias. In contrast, *UnbiasedNets* provides a notable improvement over existing works, while even reducing the robustness bias significantly in some cases, as observed by comparing the NNs trained on the diversified and original datasets.

Keywords Bias · Data-centric bias alleviation · K-means clustering · Neural networks · Noise tolerance

Editors: Dana Drachler Cohen, Javier Garcia, Mohammad Ghavamzadeh, Marek Petrik, Philip S. Thomas.

✉ Mahum Naseer
mahum.naseer@tuwien.ac.at

Extended author information available on the last page of the article

1 Introduction

Machine learning (ML)-based systems are becoming increasingly ubiquitous in today's world, with their applications ranging from small embedded devices [like health monitoring in smartwatches (Esteva et al., 2019)] to large safety-critical systems [like autonomous driving (Fink et al., 2019)]. Their success is often attributed to the Neural Networks (NNs) deployed in these systems, which have the ability to learn and perform decision-making with a high accuracy, without being explicitly programmed for their designated task. Typically, these NNs are trained on large datasets, with tens to hundreds of thousands of input samples, using various supervised training algorithms. Testing accuracy is often the most commonly (and possibly the only) used metric to analyze the performance of these NNs.

This spotlights two major limitations: (a) there is a notable reliance on large, labeled datasets, obtaining which is a significant challenge for the ML community, especially for new use-cases, and (b) the trained NN may experience problems like robustness bias, i.e., the robustness of NN to noise is not the same across all output classes, which accentuate in the presence of noisy real-world data.

Even when large datasets are available, they may contain a significantly large number of samples from one output/decision class. For instance, the MIT-BIH Arrhythmia dataset (Moody and Mark, 2001) contains a considerably larger number of normal ECG signals as compared to the ECG signals indicating a specific arrhythmia. Likewise, the IMDB-WIKI dataset (Rothe et al., 2018) comprises mostly of Caucasian faces. The NNs trained on such datasets are, therefore, less likely to detect arrhythmia or non-Caucasian faces, with high confidence—the problem aggravates under noisy input setting. However, the number of inputs from each output class is not the only parameter that leads to an *imbalanced dataset*.

1.1 Motivating example

Consider a NN trained on the Leukemia dataset (Golub et al., 1999)—details of the dataset and NN are provided in Sect. 5 along with further experiments. The training dataset contains an unequal number of inputs from the two output classes. Figure 1 (left) shows the classification performance of this network under the application of varying noise. Not surprisingly, the trained NN is more likely to misclassify inputs from the output class with less number of training inputs.

The experiments were then repeated, deleting randomly selected inputs from the class with a larger number of inputs in the training dataset each time, hence ensuring an equal number of inputs from both classes in the dataset. The graphs in Fig. 1 (right) give the classification performance of these networks under the application of varying noise. As shown in the graphs, simply having an equal number of inputs in both classes may still lead to a trained network significantly misclassifying inputs from one class.

It must also be noted that the bias becomes apparent only in the presence of noise, since the trained NNs do not indicate misclassifications in the absence of noise. Hence, the robustness bias in a trained NN may go undetected before the deployment of the NN in a real-world application. This gravitates the need to address robustness bias and calls for the better description and acquisition of *balanced datasets* that may enable training unbiased NNs. However, obtaining such datasets is not a straightforward task.

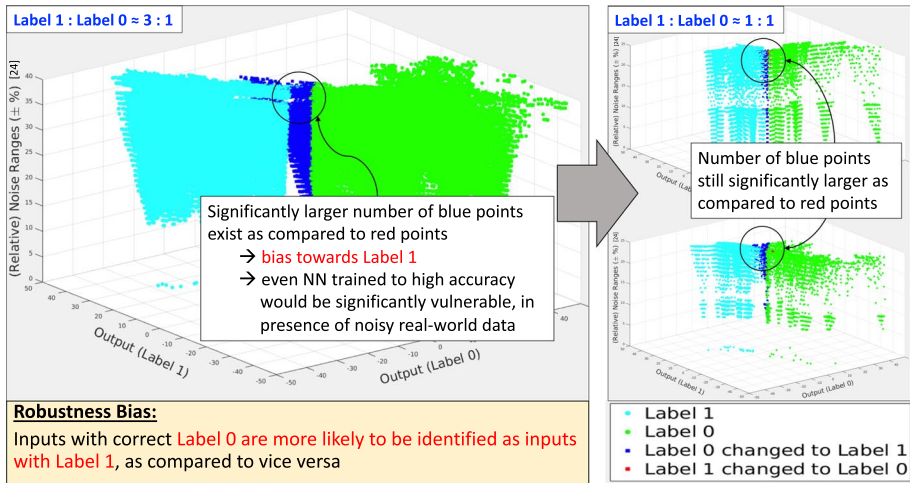


Fig. 1 Networks trained on unequal (left) and equal (right) number of inputs from the classes: *Label 0* and *Label 1*. All networks used the same network architecture and training hyper-parameters, and all indicate a higher likelihood of *Label 0* being misclassified as compared to *Label 1*

The existing works dealing with bias alleviation either aim to improve the training algorithms to ensure unbiased training, or manipulate training data to obtain datasets that favor minimal NN bias. Yet, most of these works (Gat et al., 2020; Le Bras et al., 2020; Nam et al., 2020) encounter the following limitations, making robustness bias alleviation a challenging task:

1. Most works (Li & Vasconcelos, 2019; Li et al., 2018; Zhao et al., 2017) focus on either the *dataset bias*, i.e., the lack of generalization of the available dataset to real-world data, or *representation bias*, i.e., flaws in the dataset acquired during its collection process. However, they rarely focus on biases like *robustness bias*, which generally becomes evident only during NN deployment, since noisy inputs are common in practical real-world systems.
2. A *limited notion of balanced dataset* is often used in literature (Bagui & Li, 2021; Lemaître et al., 2017), i.e., a balanced dataset is the one that contains an equal number of inputs from all output classes. However, as seen from our motivational example, such a dataset does not necessarily aid in the alleviation of robustness bias.
3. They primarily focus on *large datasets* (Nam et al., 2020; Kim et al., 2019; Le Bras et al., 2020; Gat et al., 2020; Zhang et al., 2019), which provide a large pool of training samples to learn the input features from as well as to handpick a subset of inputs that favor an unbiased NN. However, such large datasets may not always be available.
4. Some works focus on adding new input samples to the training dataset or at deeper network layers (Zhang et al., 2019). However, the *heuristics for adding new inputs* do not always favor a *balanced dataset*.
5. The addition and deletion of input samples (Bagui & Li, 2021) may also lead to *overfitting or reduction* of the training dataset, respectively.
6. The works also often focus on *visual datasets*, like colored MNIST or the IMDB dataset, where the existence of bias is perceptually easy to detect and comprehend (Wang et al.,

2020; Zhao et al., 2017). However, the *robustness bias problem may stretch beyond visual datasets, albeit often being difficult to (perceptually) detect* in non-visual datasets.

1.2 Our novel contributions

To address the aforementioned limitations and challenges, this paper proposes the *UnbiasedNets* framework,¹ which facilitates the detection and reduction (ideally elimination) of bias in a trained NN by addressing the bias at the root level, i.e., by reducing the bias within the training data, rather than relying on training algorithms to unlearn biases. Our framework is generic and hence can be implemented along with any training algorithm, using any programming language (including MATLAB, Python, C++, etc.). The novel contributions of the work are as follows:

1. This work deals with *robustness bias*, which results from having an imbalanced dataset (which may in turn be a consequence of either dataset bias or representation bias or both), to alleviate bias from datasets where the bias may not always be apparent in the absence of noisy inputs.
2. We redefine the notion of *balanced dataset* to provide a more precise explanation of the extent to which the number of inputs from each output class is, or is not, essential for training unbiased NNs.
3. Unlike the state-of-the-art approaches, *UnbiasedNets* can work efficiently to diversify the dataset even in the *absence of a large dataset* using K-means clustering and the noise tolerance of a NN previously trained on the dataset.
4. Our novel framework can identify the practical bounds for generating synthetic input samples using clusters of input features obtained via K-means and the noise tolerance bounds of the trained network. To the best of our knowledge, *UnbiasedNets* is the only framework exploiting noise tolerance to obtain realistic bounds for synthetic inputs. We also make use of feature correlation from real-world inputs to ensure that the *synthesized inputs are realistic*.
5. *UnbiasedNets* combines synthetic input generation with redundancy minimization to diversify and generate potentially balanced and equally-represented datasets, *with not necessarily an equal number of inputs from all output classes*.
6. The framework is applicable in diverse application scenarios. We demonstrate this using *UnbiasedNets* on two real-world datasets, where the *bias in the dataset is not always visually detectable*, and hence may not be straightforward to address.

1.3 Paper organization

The rest of the paper is organized as follows. Section 2 gives an overview of the existing works for bias alleviation in NNs. Section 3 elaborates on the notions of balanced datasets, robustness, robustness bias, metric for bias estimation and noise tolerance, while also providing the relevant formalism. Section 4 then explains our novel data diversification framework, *UnbiasedNets*, to alleviate robustness bias from the training dataset. Sections 5 and 6 show the application of *UnbiasedNets* on real-world datasets, providing details of experiments, results, and analysis. Section 7 discusses the open future directions for the

¹ <https://github.com/Mahum123/UnbiasedNets.git>

Table 1 Comparison of the state-of-the-art bias alleviation approaches with our proposed *UnbiasedNets* framework

Recent work	Small dataset	Non-visual dataset	Approach	Dataset Aug./Del	Leverages Δx_{max}	X_{new} validation
Alvi et al. (2018)	×	×	AC	N/A	×	N/A
at et al. (2020)	×	×	AC	N/A	×	N/A
Kim et al. (2019)	×	×	AC	N/A	×	N/A
Li & Vasconcelos (2019)	×	×	AC	Del	×	N/A
Nam et al. (2020)	×	×	AC	N/A	×	N/A
Sanh et al. (2020)	×	✓	AC	N/A	×	N/A
avani et al. (2020)	×	×	AC	N/A	×	N/A
hao et al. (2017)	×	×	AC	N/A	×	N/A
Xu et al. (2021)	×	×	AC	N/A	×	N/A
Benz et al. (2021)	×	×	AC	N/A	×	N/A
Zhang et al. (2019)	×	×	AC+DC	Aug	×	×
Chawla et al. (2002)	✓	✓	DC	Aug	×	✓
He et al. (2008)	✓	✓	DC	Aug	×	✓
Lemaître et al. (2017)	✓	✓	DC	Aug./Del.	×	✓
Le Bras et al. (2020)	×	✓	DC	Del	N/A	N/A
Li et al. (2018)	×	×	DC	Del	N/A	N/A
UnbiasedNets	✓	✓	DC	Aug.+Del.	✓	✓

Aug., Input Augmentation; Del., Input Deletion; N/A, technique not applicable for scenario; Δx_{max} , Noise Tolerance; X_{new} , Synthetic Data

improvements in data diversification for alleviating robustness bias. Finally, Sect. 8 concludes the paper.

2 Related work

This section provides an overview of the current state-of-the-art on reducing bias in NNs. The summary of state-of-the-art, including approach categorization, their predominant focus on non-visual datasets, and their comparison to our novel *UnbiasedNets* approach, is given in Table 1. The bias alleviation approaches can be broadly classified into two major categories: (1) unbiased training algorithms (i.e., algorithm-centric (AC) approaches), and

(2) bias reduction via dataset manipulation (i.e., data-centric (DC) approaches). Towards the end of the section, we also provide an overview of the current and on-going works targeting the recently discovered problem of robustness bias.

2.1 Algorithm-Centric (AC) approaches

Training unbiased NN via AC approaches often involves splitting the network model into two separate but connected networks (Alvi et al., 2018; Kim et al., 2019; Nam et al., 2020). The first network aims at either identifying key input features or amplifying the bias present in the dataset. The second network, in turn, uses these features or accentuated bias to unlearn the bias from the network. Learning features at deeper NN layers during training for data augmentation (Zhang et al., 2019) has also been shown to aid unbiased training. In addition, knowledge of known biases in the dataset and a NN trained using standard cross-entropy loss has also been leveraged to develop a more robust NN (Sanh et al., 2020). Other AC bias reduction approaches include the incorporation of additional constraints during training to guide the NN in order to avoid learning unwanted correlations in data (Zhao et al., 2017).

For biases specific to multi-modal datasets (like colored MNIST (Kim et al., 2019), where the dataset contains two kinds of information: the colors and the numerals), the use of a training algorithm based on functional entropy is shown to perform better (Gat et al., 2020). A recent work (Li & Vasconcelos, 2019) also explores inputs in the dataset to identify the weights² that the inputs must be encoded with before training, to successfully reduce the bias. The determination of invariants in inputs has also been proposed (Arjovsky et al., 2019) to enable unbiased training of a NN. In addition, recent work (Savani et al., 2020) also explores algorithms where instead of training an unbiased network from scratch, a trained NN and dataset (not used during training) are used to fine-tune the network to be devoid of biases specific to a certain application.

However, as indicated earlier, these works are tailored for minimizing data and representation biases, generally for large datasets. The biases are often explored in visual datasets. In contrast, NNs deployed in the real-world often also deal with non-visual inputs, like patient's medical data, where the existence of a bias (even the data and representation biases) may not always be easy to detect and hence may go unnoticed. Hence, bias alleviation poses a challenge in cases where the detection of bias is beyond visual perception. Moreover, the exploration of robustness bias is a fairly new research direction, and hence, the success of these AC approaches for minimizing robustness bias remains largely unexplored.

2.2 Data-centric (DC) approaches

The orthogonal direction to minimize bias is by manipulating the training dataset via DC approaches, to potentially eliminate the bias at its core. Among the simplest and most popular DC bias alleviation approaches are random over-sampling (ROS), i.e., random replication of inputs from the class with less number of input samples, or random under-sampling (RUS), i.e., random deletion of inputs from the class with a significantly larger portion of

² Note that the weights for encoding inputs in Li & Vasconcelos (2019) are not same as the parametric weights of NN layers.

available inputs (Bagui & Li, 2021; Leevy et al., 2018). The idea is to obtain a dataset with an equal number of inputs from each class. However, RUS is known to reduce the number of input samples available for NN to learn, while ROS may lead to overfitting the training data.

The synthetic minority over-sampling (SMOTE) (Chawla et al., 2002) and adaptive synthetic sampling (ADASYN) (He et al., 2008) techniques provide an improvement over ROS by synthesizing new points in the class with less number of samples using the available inputs as reference for the synthesis of new input samples (Lemaître et al., 2017). However, the general assumption in these works is that having an equal number of inputs for each of the classes ensures a balanced dataset, and in turn ensures an absence of bias (Bagui & Li, 2021; Picek et al., 2019). As such, the approaches deploy data manipulation for the output class with a smaller number of inputs only. As observed in the motivating example in Sect. 1, this assumption provides a limited notion of balanced datasets. In addition, neither do these works have the means to ensure if the new inputs generated in fact belong to the minority class (i.e., output class with less number of inputs), nor the sophistication to analyze the number of inputs required to be added to the class to alleviate bias.

Other works explore heuristics to identify the inputs that must be removed from the training dataset (Le Bras et al., 2020; Li et al., 2018) for obtaining an unbiased NN. However, for most real-world applications, large labeled datasets may not always be available, except to a few tech giants. This leaves limited scope for tasks relying on limited dataset for bias alleviation.

In summary, the DC approaches again focus on alleviating representation and data bias, i.e., the biases pertaining to faulty data acquisition and lack of data generalizing well to all output classes. Alleviation of robustness bias remains an unexplored research direction in the existing works. The notion of a balanced dataset often used in these works is too naive. For the approaches relying on the deletion of inputs from the training dataset, the approaches are ideal only for large datasets to ensure sufficient inputs remain for NN training. For the augmentation approaches (like ROS, SMOTE and ADASYN), i.e., the approaches where synthetic inputs are added to the training dataset (henceforth referred to as *data augmentation*), the location for the new inputs is chosen to be in the close proximity around existing “randomly” selected inputs. The new inputs may or may not be realistic for the real-world input domain. The validation of these generated synthetic inputs relies solely on them being a part of NN training, and how well the trained NN works with the testing dataset.

2.2.1 Bias and the focus on visual datasets

As highlighted in Sect. 1, NNs are deployed in a diverse range of applications. These include networks performing classification and decision-making tasks for visual inputs (Vu et al., 2022; Li et al., 2021). Yet, a large portion of NN applications, for instance, banking (Asha & KR, 2021), environmental forecast (Benali et al., 2019), finance (Calvo-Pardo et al., 2020) and spam filtering (Barushka and Hajek, 2018), accept non-visual inputs. However, most literature pertaining to bias analysis (Alvi et al., 2018; Gat et al., 2020; Kim et al., 2019; Nam et al., 2020; Li et al., 2018; Li & Vasconcelos, 2019; Zhang et al., 2019; Zhao et al., 2017) focus (often solely) on NNs working on visual datasets—this comes to no surprise since a bias in these datasets is visually perceptible to human analysts, who are inclined to perceive visual queues better than the non-visual ones (for instance, consider

the case of visual capture, where visual senses are observed to dominate over auditory senses (Welch, 1999)).

The NNs using non-visual inputs often deploy similar network architectures as those using visual inputs. Intuitively, these NNs are likely to be as biased as their counterparts used in visual applications. Yet, the difficulty in perceiving the bias in non-visual datasets makes their bias analysis a scarcely explored research area, as evident in the lack of existing works in the domain.

Such dominant focus on visual datasets is not unique to the study of bias but is, in fact, also observed in fields like visual analytics, where non-visual aspects of the system are transformed into visual aspects. For example, the neuron activations are presented graphically (visually) in the research on network interpretability (Becker et al., 2020) and security (Liu et al., 2018), which enables problem identification (detection). This in turn motivates deeper research/solutions.

2.3 Current and ongoing efforts

The vulnerability of NNs to robustness bias has only been recently discovered (Nanda et al., 2021). Hence, the efforts to resolve this particular category of bias are still limited. Nevertheless, a few AC approaches have been proposed within the last year to alleviate such bias. This includes a multi-objective training algorithm, which ensures that the standard error (which dictates the classification accuracy of the networks) and boundary error (since the inputs from class(es) closer to the decision boundary are expected to be more vulnerable under noise) (Xu et al., 2021) are minimal, thereby minimizing the bias. However, later work (Nayak et al., 2022) comes to a contrary conclusion, i.e., even the inputs with the same distance to the classification boundary may have different vulnerabilities to the noise. A re-weighting approach has also been proposed (Benz et al., 2021), which aims to update parameter values during training whenever the accuracy of a particular output class deviates too much from the average accuracy of the network.

Recent work (Benz et al., 2021) also notes that the bias in the NNs exists due to the dataset (and its features) itself, rather than depending on the NN model or its optimization factors. Yet, to the best of our knowledge, no DC effort has been proposed to alleviate bias from the dataset itself. It is interesting to note that adversarial training, a popular approach found successful in ensuring the robustness of NN against noise (concept explained later in Sect. 3.2), is found to aggravate the bias (Tian et al., 2021).

3 Preliminaries

This section describes the notions and provides the relevant formalism for balanced datasets, robustness, robustness bias, bias estimation and noise tolerance (Nanda et al., 2021; Naseer et al., 2020), which form the basis of *UnbiasedNets*. The terminology and notations introduced in the section will be used throughout the rest of the paper.

3.1 Balanced datasets

Contrary to the popular notion, i.e., a balanced dataset (Bagui & Li, 2021; Lemaître et al., 2017) consists of an equal number of inputs from all output classes, we define *balanced dataset* to be the dataset where all output classes are *equally-represented*.

Definition 1 (*Balanced Dataset*) Given a dataset X with \mathcal{L} output classes (i.e., $Y_1, Y_2, \dots, Y_{\mathcal{L}}$), the dataset is said to be balanced/the output classes are equally represented iff density ρ of inputs from each class in the input hyperspace is (approximately) equal, i.e., $\rho(Y_1) \approx \rho(Y_2) \approx \dots \approx \rho(Y_{\mathcal{L}})$. Note that density ρ of input here refers to the average number of input samples contained within the unit hypervolume of the valid input domain for an output class.

This implies that a network trained on such a balanced dataset would potentially be equally likely to identify inputs from all the classes, without a bias (explained in Sect. 3.3).

3.2 Robustness

Robustness is the property of NN that signifies how the application of noise Δx to the inputs does not change what the trained NN originally learned about the inputs.

Definition 2 (*Robustness*) Given a trained network $N : X \rightarrow Y$, N is said to be robust against the noise Δx if the application of an arbitrary noise $\eta \leq \Delta x$ to the input $x \in X$ does not change network's classification of x , i.e., $\forall \eta \leq \Delta x : N(x + \eta) = N(x)$.

It must be noted that x corresponds to inputs that the network N does not originally misclassify, i.e., $N(x)$ corresponds to the true output class for input x . For the purpose of this work, we assume the noise η to be bounded within the L^∞ space around input x , with the radius of Δx —this is one of the most popular noise used in NN analysis literature. Nevertheless, it is fairly straightforward to opt for any other type of (L^p -norm bounded) noise for the framework.

3.3 Robustness bias

Section 1 highlighted the well-studied NN biases in literature, i.e., data and representation bias. This paper instead deals with *robustness bias* (henceforth referred to as only *bias*) proposed by Nanda et al. (2021) and Joshi et al. (2022), which is a property of the dataset where a specific output class may or may not be robust under the application of noise. More specifically, it can be defined as follows:

Definition 3 (*Robustness Bias*) Given a dataset X with \mathcal{L} output classes (i.e., $Y_1, Y_2, \dots, Y_{\mathcal{L}}$), and $\mathcal{D}_{Y_1}, \mathcal{D}_{Y_2}, \dots, \mathcal{D}_{Y_{\mathcal{L}}}$ as the of input sub-domain representing each output class. X is said to exhibit robustness bias iff the sub-domains $\mathcal{D}_{Y_1}, \mathcal{D}_{Y_2}, \dots, \mathcal{D}_{Y_{\mathcal{L}}}$ are not equidistant from the decision boundary.

Naturally, the sub-domains $\mathcal{D}_{Y_1}, \mathcal{D}_{Y_2}, \dots, \mathcal{D}_{Y_{\mathcal{L}}}$ may be disjoint or overlapping. However, as long as the sub-domains are equidistant from the decision boundary, the dataset is said to be free from a robustness bias. A NN trained on such a dataset is said to be unbiased, since intuitively, for a NN with a decision boundary equidistant from all input sub-domains, all output classes must be equally robust to noise.

However, given the large number of input features (forming an input hyperspace) in practical datasets, it is not easy to visualize the bias in the dataset itself. Hence, we define

the notion of biased NN, which aids in identifying the robustness bias in the dataset via analyzing the NN trained on the dataset:

Definition 4 (*Biased Network*) Given a trained network $N : X \rightarrow Y$, N is said to be biased if the application of an arbitrary noise $\eta \leq \Delta x$ to any (correctly classified) input from class $X_i \subset X$ does not change network's output classification, $\forall \eta \leq \Delta x, x_i \in X_i : N(x_i + \eta) = N(x_i)$. However, application of the same noise to any input from another class $X_j \subset X$ makes the network misclassify the originally correctly classified input from the class $\forall \eta \leq \Delta x, x_j \in X_j : N(x_j + \eta) \neq N(x_j)$.

It must be noted that even though unbiasedness (i.e., the property of a trained NN to be unbiased) and classification accuracy may intuitively seem similar, they are not identical. Obtaining an accurate NN involves identifying the decision boundary that *separates* the output classes in the dataset. In contrast, obtaining an unbiased NN involves identifying a decision boundary that is *equidistant* from all the sub-domains encapsulating the different output classes. The resulting unbiased network, in turn, may or may not have the highest classification accuracy. However, all the output classes will likely be equally robust to noise in an unbiased network.

3.4 Metric for robustness bias

In practice, it is often impossible to obtain a completely unbiased NN. Hence, a metric is required to quantify and analyze the bias in the network. Let R_i be the ratio of misclassified to correctly classified inputs from class i , which defines the average tendency of inputs from output class i to be misclassified. We define the metric to estimate robustness bias (\mathcal{B}_R) as follows:

$$\mathcal{B}_R = \max \left(\text{abs} \left(R_i - \frac{\sum_{j \in \mathcal{L} \setminus i} R_j}{|\mathcal{L}| - 1} \right) \right)$$

where \mathcal{L} is the set of all output classes. Having a \mathcal{B}_R of zero indicates an equal R_i across all output classes, and therefore an unbiased NN. Consequently, larger \mathcal{B}_R implies higher bias. It must also be noted that the (absolute) difference in ratios R_i and R_j is generally different across the different pairs of output classes. In order not to reduce (nullify) the impact of the differences (and hence that of the bias in the network), the maximum difference, rather than the average, is used to estimate the bias in NN.

Contrary to the formal notion of robustness bias, as provided in Definition 3, \mathcal{B}_R uses the inputs to quantify bias rather than the decision boundary of the NN. This is a viable approach since the exact decision boundary of the NN is often hard to visualize for the multi-dimensional input space. The metric \mathcal{B}_R , instead, makes use of the measurable/quantifiable entity, i.e., the input classification, to estimate the bias. As stated earlier, the ratio R_i provides the tendency of the boundary to misclassify the inputs from class i . This is compared to the average tendency of misclassification of inputs from the other network classes R_j —this is analogous to comparing the distance of inputs to the decision boundary for different classes. Hence, if the ratio R_i for all classes is equal (analogously all classes are equidistant from the decision boundary), \mathcal{B}_R computes to zero. The NN is then ought to be unbiased.

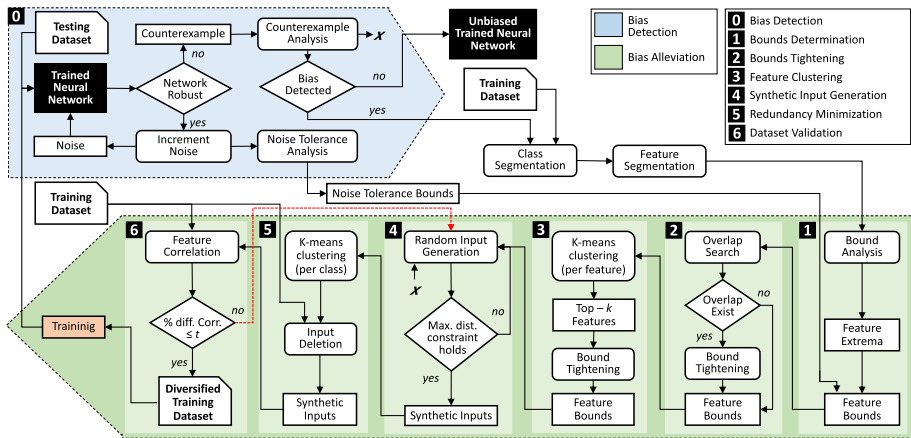


Fig. 2 Overview of the *UnbiasedNets* framework incorporating the proposed methodology starting with a trained NN undergoing bias detection, followed by bias alleviation, ultimately leading to a diversified dataset and potentially unbiased trained NN

3.5 Noise tolerance

Similar to robustness, noise tolerance also checks the classification performance of a NN for inputs under the application of noise. However, it is a stronger property than robustness (i.e., noise tolerance to a specific noise implies robustness to the noise as well) such that it provides the bounds within which the addition of noise does not change the classification of the inputs by a trained NN.

Definition 5 (Noise Tolerance) Given a trained network $N : X \rightarrow Y$, noise tolerance is defined as the *maximum* noise Δx_{max} , which can be applied to a correctly classified input $x \in X$ such that N does not misclassify the input. Hence, for any arbitrary noise $\eta \leq \Delta x_{max}$, the application of noise to an input $x \in X$ does not change network’s classification of x , i.e., $\forall \eta \leq \Delta x_{max} : N(x + \eta) = N(x)$.

Alternatively, noise tolerance can be viewed as the largest δ -ball (l^∞ norm ball) around the inputs, such that $\delta = \Delta x_{max}$ and any input within this ball is correctly classified by the NN. Consequently, this knowledge can in turn be used to estimate the region around seed inputs where the realistic synthetic inputs may reside and still be correctly identified by a trained NN.

4 UnbiasedNets: framework for bias alleviation

We categorize *UnbiasedNets* into two major tasks: *bias detection* using a trained NN to identify the existence of robustness bias followed by *bias alleviation* to diversify the training dataset to eliminate the bias at its core. Figure 2 provides an overview of our proposed methodology.

4.1 Bias detection

The first step here is the application of noise η , bounded by the small noise bounds Δx to the inputs present in the testing dataset $x \in X$ (shown as Block 0 in Fig. 2) to obtain the noisy inputs x_η .

$$x_\eta = x + \eta \quad \text{s.t.} \quad \eta \leq \Delta x \quad (1)$$

The noisy inputs are then supplied to the trained NN, and their output classifications are compared to the classifications of inputs in the absence of noise. For the network to be robust (see Definition 2), the NN's classification must not change under the influence of noise. The noise is then iteratively increased, beyond the maximum noise at which the NN does not misclassify the inputs, i.e., beyond the NN's noise tolerance (see Definition 5). Such iterative increment of noise provides the noise tolerance bounds of the network.

The application of noise larger than the noise tolerance bounds of the NN entails that the NN misclassifies some or all the noisy inputs. These misclassifying noise patterns (i.e., the counterexamples) act as inputs for the counterexample analysis. These noise patterns can be collected either using a formal framework [such as the ones based on model checking used by Naseer et al. (2020) and Bhatti et al. (2022)] or an empirical approach [like the Fast Gradient Sign Method (FGSM) attack (Goodfellow et al., 2015)].

During counterexample analysis, the collected noise patterns, and in turn the misclassified inputs, are used to compute the \mathcal{B}_R of the network to detect the presence and severity of robustness bias in the trained NN. A non-zero \mathcal{B}_R implies a robustness bias in the network. Additionally, the number of misclassified inputs from each class is also used to determine the number of synthetic inputs required in the training dataset (elaborated in Sect. 4.2.4) to alleviate the bias.

4.2 Bias alleviation

Using the noise tolerance available from the bias detection and the feature extremum of the inputs from the training dataset, we provide the step-by-step bias alleviation methodology. The aim of the methodology is to identify the valid input domain for the generation of synthetic data and provide a diversified training dataset for the training of a potentially unbiased NN. The details of each step in the methodology are as follows.

4.2.1 Bounds determination

For each input feature in every output class, the feature extremum, i.e., the maximum and minimum value of the feature as per the available training data, is first identified (as shown in Block 1 of Fig. 2). As discussed earlier, the inputs with noise, less than the allowed noise tolerance, are still likely to be correctly classified by a trained NN. Hence, the feature bounds are relaxed using Δx_{max} , to provide a larger input space for the diversified inputs (also shown in Fig. 3a), as follows:

Theorem 1 (Bound Relaxation using Noise Tolerance) *For input domain X , let $[x_i, \bar{x}_i]$ represent the bounds of inputs belonging to X_i (where $X_i \subset X$) and Δx_{max} be the noise*

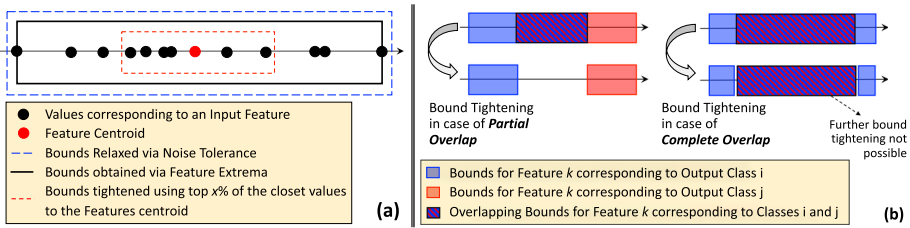


Fig. 3 **a** Realistic bounds determination for individual feature bounds using available training inputs, K-means clustering and noise tolerance, **b** Bound tightening to eliminate/reduce bound overlap for synthetic input generation

tolerance of the network. From Definition 4, we know that the application of noise within the tolerance of the network does not change the output classification. Hence, more realistic input bounds $[x'_i, x'_i]$ can be obtained using the laws of interval arithmetic as:

$$x'_i = \min((x_i - \Delta x_{max}), (x_i + \Delta x_{max}), (\bar{x}_i - \Delta x_{max}), (\bar{x}_i + \Delta x_{max})),$$

$$\bar{x}'_i = \max((x_i - \Delta x_{max}), (x_i + \Delta x_{max}), (\bar{x}_i - \Delta x_{max}), (\bar{x}_i + \Delta x_{max}))$$

It must be noted that due to the scalability of underlying bias detection framework [for instance (Naseer et al., 2020)], where the application of large noise to NN inputs may lead to very large formal models, not suitable for analysis, noise tolerance may not always be available for bound relaxation. A similar challenge is encountered for NNs with a very low noise tolerance. Consider the example of a NN trained on an image dataset, where the addition of noise leading to a magnitude change of even 1.0 in the pixel value of an image may still lead to misclassification (Ma et al., 2021). This indicates a very low noise tolerance. Under these conditions, *UnbiasedNets* assumes the noise tolerance to be zero, and proceeds with feature extremum as the feature bounds obtained during bound determination.

4.2.2 Bound tightening

Bounds obtained from the previous step identify the regions in the input space where real inputs from the training dataset exist, and hence provide an estimate for the generation of valid synthetic data. However, it is possible for the feature bounds for different output classes to overlap, as shown in Fig. 3b. The overlap can be either *partial* or *complete*. This provides a means for tightening the feature bounds (shown as Block 2 in Fig. 2), hence leading to smaller, yet realistic, input space for the generation of synthetic data. This in turn ensures that a lesser number of iterations are required for realistic synthetic input generation in the later steps of the framework. The generation of tighter feature bounds in the case of partial feature can be seen as follows:

Theorem 2 (Bound Tightening in case of Partial Overlap) *Given the bounds of input feature a for inputs belonging to class i and j to be $[x'_i, x'_i]$ and $[x'_j, x'_j]$, respectively, the bounds can be tightened to $[x''_i, x''_i]$ and $[x''_j, x''_j]$ provided that $x'_i < \bar{x}'_j$ and $\bar{x}'_i < x'_j$ (i.e., the bounds overlap partially). Then, any input belonging to the new bounds also belongs to the original feature bounds as well.*

$$\forall i, j. (([\underline{x}_i^a, \overline{x}_i^a] \in X_i^a \wedge [\underline{x}_j^a, \overline{x}_j^a] \in X_j^a) \implies ([\underline{x}_i^a, \underline{x}_j^a] \in X_i^a \wedge [\overline{x}_i^a, \overline{x}_j^a] \in X_j^a))$$

$$s.t. \underline{x}_i^a < \underline{x}_j^a < \overline{x}_i^a < \overline{x}_j^a$$

However, the same cannot be generalized for complete overlap since the bounds of one label form a subset of the other. As such, tightening is possible for a single label only.

Theorem 3 (Bound Tightening in case of Complete Overlap) *Given the bounds of input feature a for inputs belonging to class i and j to be $[\underline{x}_i^a, \overline{x}_i^a]$ and $[\underline{x}_j^a, \overline{x}_j^a]$, respectively, the bounds for feature a of class i , X_i^a , can be tightened to $[\underline{x}_i^a, \underline{x}_j^a]$ and $[\overline{x}_j^a, \overline{x}_i^a]$ provided that $\underline{x}_i^a < \underline{x}_j^a$ and $\overline{x}_j^a < \overline{x}_i^a$. Then, any input belonging to the new bounds for X_i^a also belongs to the original feature bounds as well.*

$$\forall i, j. (([\underline{x}_i^a, \overline{x}_i^a] \in X_i^a \wedge [\underline{x}_j^a, \overline{x}_j^a] \in X_j^a) \implies ([\underline{x}_i^a, \underline{x}_j^a] \in X_i^a \wedge [\overline{x}_j^a, \overline{x}_i^a] \in X_i^a))$$

$$s.t. \underline{x}_i^a < \underline{x}_j^a < \overline{x}_j^a < \overline{x}_i^a$$

Motivating Example Consider an arbitrary feature a with valid input values in the range $[0, 10]$. Let the inputs from class i have the bounds $[2, 8]$ and those from class j have the bounds $[7, 10]$, for the feature a . Without bound tightening, any input $7 < x^a < 8$ can belong to either class i or j (but not both). On the contrary, bound tightening reduces the bounds of the feature a for classes i and j to $[0, 7]$ and $[8, 10]$, respectively. This reduces the valid input domain for feature a such that it is impossible to pick a sample for feature a that may belong to more than a single output class, hence simplifying the task of generating realistic synthetic input samples.

4.2.3 Feature clustering

The previous steps in the framework make use of the entire training dataset to obtain realistic feature bounds. But intuitively, real-world inputs often contain outliers that may be part of the training dataset, which do not occur frequently in practical case scenarios. To subsume this characteristic into the synthetic inputs generated, further bound tightening is carried out (shown as Block 3 in Fig. 2) on the *top-k* input features, i.e., the k features with the smallest distance from cluster centroid to the farthest input.

4.2.4 Synthetic input generation

Using the feature bounds obtained from the previous step, the random input values are chosen within the available bounds (shown as Block 4 in Fig. 2). The number of inputs to be added to each output class χ_i is determined on the basis of the ratio of percentage of misclassified inputs from class i (i.e., μ_i) and the percentage of misclassified inputs from the class with minimum misclassifications (i.e., $\min(\mu_L)$) using counterexamples recorded during the bias detection. Hence, the class with higher μ_i gets the most synthetic inputs added to the dataset.

Algorithm 1 outlines the entire synthetic data generation process, starting from the training dataset and noise tolerance bounds. Function `classSegment` (Line: 3) splits the dataset into non-overlapping subsets of inputs belonging to each class, `globalExt`

(Line: 5) provides feature bounds using feature extremum, `nonOverlapping` (Line: 8) performs bound tightening on basis of Theorems 2 and 3, `minDist` (Line: 10) identifies the *top-k* features based on k-means clustering, `boundsFinal` (Line: 12) performs further bound tightening based on the top features, and `randInp` (Line: 15) finally generates the synthetic inputs for each output class.

Algorithm 1 Synthetic Data Generation

Input: Training Inputs (X), Number of Output Classes (N), Noise Tolerance (Δx_{max}), Number of top Features to use for Bound Tightening k , Number of Inputs to add to each Class (χ)
Output: Augmented Input Matrix (X'), Vector of Output Classes (L')

```

1: function SYNTHGEN( $X, N, \Delta x_{max}, \chi$ )
2:    $n = \text{size}(X, 2)$  ▷ Number of Input Features
3:    $(X_1, \dots, X_N) = \text{classSegment}(X, N)$ 
4:   for  $i = 1:N$  do
5:      $(f'_{min_i}, f'_{max_i}) = \text{globalExt}(\Delta x_{max}, X_i)$  ▷ Block 1 in Fig. 2
6:   end for
7:   for  $j = 1:n$  do
8:      $(f'_{min}^j, f'_{max}^j) = \text{nonOverlapping}(f'_{min}^j, \dots, f'_{max}^j)$  ▷ Block 2 in Fig. 2
9:   end for
10:   $(T_1, \dots, T_k) = \text{minDist}(X)$ 
11:  for  $m = 1:k$  do
12:     $(f''_{min}^{T_m}, \dots, f''_{max}^{T_m}) = \text{boundsFinal}(f'_{min}^{T_m}, f'_{max}^{T_m})$  ▷ Block 3 in Fig. 2
13:  end for
14:  for  $i = 1:N$  do
15:     $X_{new} = \text{randInp}(f''_{min_i}, \dots, f''_{max_i}, \chi_i)$  ▷ Block 4 in Fig. 2
16:  end for
17: end function

```

It must be noted that the above input generation assumes an implicit hyperrectangular distribution of the input domain. This means, each input feature may take any input value (from within the defined input bounds), with equal likelihood. However, it is also possible for the input features to have non-rectangular distributions. Assuming these distributions to be known a priori, the random input generation could be modified to select input values, from within the input bounds, according to their probability of occurrence in their exact input distributions, i.e., with the more probable values having higher likelihood of selection and vice versa.

4.2.5 Redundancy minimization

Oversampling may lead an NN to overfit to the training samples. Moreover, the existence of similar inputs, after the addition of synthetic inputs, does not add to the diversity of the dataset. Existing works also indicate that training the NNs on smaller datasets—for instance, those obtained by eliminating input instances leveraging different distance matrices—may reduce the timing overhead for training while providing comparable classification accuracy (Fuangkhan, 2022; Kotsiantis et al., 2006; Wang et al., 2009). (Also see Appendix A for case studies indicating how redundancy minimization using K-means deletion reduces the bias of the actual NNs).

Hence, $x\%$ closely resembling inputs from each class are removed to minimize the redundancy in the diversified training dataset (shown as Block 5 in Fig. 2). This is done by generating $\frac{1}{x}$ clusters for each output class and then retaining a single input from each

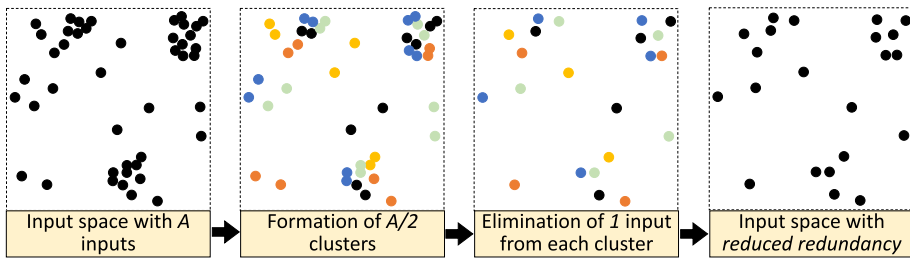


Fig. 4 Redundancy minimization by 50% in a two-dimensional input space

cluster. The result is a dataset with input samples covering diverse input space, without densely populating any specific region of the input space (as realized in Fig. 4).

4.2.6 Dataset validation

Up until the previous step, *UnbiasedNets* used real-world inputs to identify valid input space within which the inputs exist, used knowledge of the percentage of misclassified inputs from each output class to identify the number of synthetic inputs to generate, and minimized the redundancy in the generated input samples to obtain a diversified dataset. However, features in the real-world data may be correlated, and the synthetic input features, despite lying in the valid input domain, may not follow the correlation of real-world data. Hence, this step aims to validate the synthetic inputs by comparing their feature correlation with that of the original training data. If the percentage difference between the correlation coefficients is within $t\%$, the new inputs are deemed suitable for training a potentially unbiased NN. Otherwise, the process of synthetic data generation is repeated until the feature correlation of the synthetic inputs resembles that of the original training dataset (shown as Block 6 in Fig. 2).

The choice of t is made on the basis of the percentage difference between the correlation coefficients of training and testing datasets. However, if this difference is too large, the features may simply be independent, or obtaining appropriate correlations may require some input pre-processing (Zhao et al., 2006). The use of *only* simple Pearson correlation coefficient, on such raw data, may not be an appropriate statistical measure to ensure the synthetic inputs to be realistic here. (Check Appendix B for more insights into this.)

5 Experiments

This section describes our experimental setup, and details of NNs and datasets used in our experiments.

5.1 Experimental setup

All experiments were carried out on CentOS-7 system running on a 3.1GHz 6 core Intel i5-8600. Our *UnbiasedNets* framework was implemented on MATLAB. The NN training was carried out using Keras.

However, the setup did not make use of any special libraries and, hence, can be easily re-implemented using any programming language(s). Bias detection (and counterexample generation) was carried out using SMV models with applied noise in the range of 1–40% of the actual input values, using a timeout of 5 minutes for each input.

5.2 Datasets and neural network architecture

We experimented on the Leukemia dataset (Golub et al., 1999), which is composed of the genetic attributes of Leukemia patients classified between Acute Lymphoblast Leukemia (ALL) and Acute Myeloid Leukemia (AML). The training dataset consists of 38 input samples (with 27 and 11 inputs indicating ALL and AML, respectively), while the testing dataset contains 34 inputs (with 20 and 14 ALL and AML inputs, respectively). We trained a single hidden layer (20 neurons), fully-connected ReLU-based NN, using the top-5 most essential genetic features from the dataset extracted using Minimum Redundancy and Maximum Relevance (mRMR) feature selection technique (Khan et al., 2018). A learning rate of 0.5 for 40 epochs followed by another 40 epochs with a learning rate of 0.2 were used during training.

We also experimented on the Iris dataset (Dua & Graff, 2017; Fisher, 1936), which is a multi-label dataset, with characteristics of three iris plant categories as input features. The dataset has an equal number of inputs from all output classes. We split the dataset into training and testing datasets, with 120 and 30 inputs, respectively, while ensuring an equal number of inputs from all classes in each dataset. A fully-connected ReLU-based two-hidden layer (15 neurons each) NN was trained with a learning rate of 0.001 for 80 epochs, using a training to validation split of 4:1.

Since *UnbiasedNets* is a data-centric bias alleviation framework, we compare the framework to well-acknowledged open-source state-of-the-art data-centric approaches: RUS, ROS, SMOTE (Chawla et al., 2002) and ADASYN (He et al., 2008). The Python toolbox *imbalanced-learn* implements all of the aforementioned techniques, except RUS, and was used for the generation of testing datasets. Since these approaches require the number of inputs to be different in each class, 50% of the inputs from the Iris dataset were randomly selected to create a sub-dataset with an unequal number of inputs for the classes. RUS was implemented on MATLAB, removing inputs from class with more inputs to ensure both classes have the same number of inputs in the case of the Leukemia dataset and removing 25% samples from each class in the case of the Iris dataset. To avoid overfitting during retraining of NNs using augmented datasets, the number of training epochs was reduced proportionally to the increase in the size of datasets.

All NNs considered in the experiments were trained to the training and testing accuracies of over 90%. In addition, the experiments for each bias alleviation approach were repeated 10 times to ensure conformity.

6 Results and analysis

This section elaborates on the empirical results obtained from our experiments followed by comparison and analysis of *UnbiasedNets* to the data-centric bias alleviation approaches.

Table 2 Comparison of \mathcal{B}_R values (average \pm standard deviation) obtained for the NNs trained on original and diversified datasets, using open-source state-of-the-art approaches and *UnbiasedNets*

Approach	Robustness bias (\mathcal{B}_R) of Networks trained on:	
	Leukemia dataset	Iris dataset
Original	0.2228	0.4732
RUS	0.1710 \pm 0.07	0.5042 \pm 0.11
ROS	0.2213 \pm 0.07	0.8059 \pm 0.36
SMOTE	0.1452 \pm 0.08	0.7709 \pm 0.72
ADASYN	0.2434 \pm 0.06	<i>ADASYN not suited for dataset</i>
<i>UnbiasedNets</i>	0.1236 \pm 0.05	0.4906 \pm 0.15

The bias of the network trained on original dataset is given in bold, and that of the network trained diversified dataset is given in bold italics.

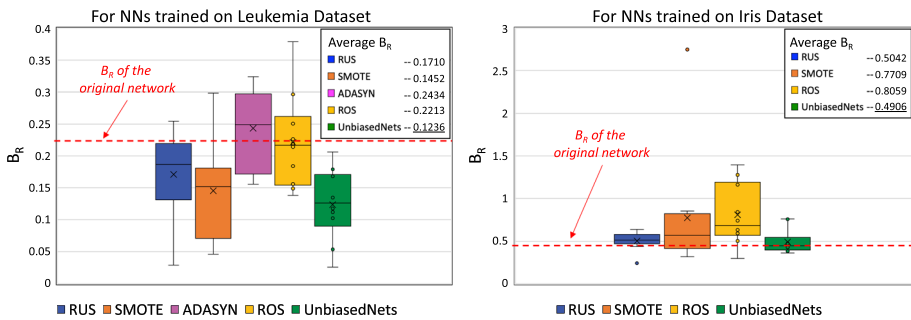


Fig. 5 Variation in \mathcal{B}_R results for NNs trained on RUS, SMOTE, ADASYN, ROS, and the diversified *UnbiasedNets* datasets

6.1 Observations

As the number of output classes increases, ensuring an unbiased NN becomes a more challenging task. This was clearly observed in our experiments (Table 2), wherein the multi-label classifiers had a higher bias and at the same time, their bias reduction was substantially less effective in all bias alleviation approaches.

As discussed in Sect. 3, lower \mathcal{B}_R indicates that the difference in the ratio of misclassified to correctly classified inputs is low, implying that the NN is less biased towards any output class. As summarized in Table 2, our *UnbiasedNets* framework outperformed all the DC bias alleviation techniques while obtaining optimum \mathcal{B}_R values for both binary and multi-label datasets. Moreover, in the case of the Iris dataset, using classical data-centric approaches to generate dataset with an equal number of inputs from each class seems to exacerbate the robustness bias. Although *UnbiasedNets* may not always reduce the robustness bias, the data diversification ensures that the dataset remains balanced.

This success of biased can also be seen in Fig. 5, which shows the variation in \mathcal{B}_R values over the repeated experiments. It is clearly evident that the individual experiments leading to a decrease in average robustness bias are far more compared to vice versa. Hence, we advocate executing several instances of experiments in order to obtain dataset instances that offer the best bias alleviation.

Additionally, it can be seen from the box plots that NNs trained using the *UnbiasedNets* datasets demonstrate considerably low interquartile ranges and the lowest average \mathcal{B}_R values. Even though RUS illustrates competitive \mathcal{B}_R values, the use of RUS is not appropriate for small datasets, since the approach involves the deletion of real input samples and may hence diminish the learning capability of the NN. The remaining approaches, i.e., ROS, SMOTE, and ADASYN, present a large variation in \mathcal{B}_R results, deeming the approaches less effective for alleviation of robustness bias.

6.2 Analysis

Our work focuses on robustness bias, which is exhibited by a trained NN in the presence of inputs having higher robustness to noise for certain output classes as compared to others. From our experiments, we confirm the hypothesis that having an equal number of inputs (as in the case of Iris dataset) is in fact *insufficient* to ensure an unbiased network.

In the case of the datasets where the number of inputs in each class is different, the known approaches like RUS, ROS, SMOTE, and ADASYN may reduce the bias. But for most datasets, they may be inadequate for robustness bias alleviation mainly for two reasons: (1) they rely on the naive definition of balanced datasets and only ensure the number of inputs for each class is equal, which overlooks the requirement of each class to be *equally-represented* (concept explained in Sect. 3.1) in the input, and (2) during data augmentation, new inputs are only added in between the existing inputs, which neither diversifies the dataset sufficiently nor ensures that the new inputs are valid candidates for the augmented dataset. *UnbiasedNets*, on other hand, uses counterexample analysis from the bias detection stage to obtain the required number of inputs in each class for a potentially *equally-represented* dataset. It also uses noise tolerance, which allows us to diversify the data beyond the bounds of the existing training dataset, which is subsequently validated by leveraging feature correlations, to alleviate bias in NN.

In the case of the Iris dataset, ROS and SMOTE were observed to significantly worsen robustness bias. This may be partially due to the deletion of inputs from the dataset to create an unequal number of inputs in the classes, which reduces the data available for NN training. However, RUS retained the \mathcal{B}_R value close to the original dataset, even though the approach also employs input deletion. This suggests that the data augmentation by ROS and SMOTE may actually contribute to an exacerbation of bias rather than alleviation. In the case of *UnbiasedNets*, even though the improvement in bias is often small, the results clearly suggest that diversifying the training dataset by adding realistic synthetic inputs and reducing redundancy in dataset is a potential direction to alleviate bias in NNs, unlike the other approaches.



Fig. 6 Inputs from one output class may resemble inputs from other classes, as observed in the MNIST dataset

7 Discussion

UnbiasedNets aims to diversify the dataset so as to (potentially) achieve a balanced dataset. While the diversification goal for obtaining a completely unbiased network may not always be achieved, *UnbiasedNets* rarely aggravates the bias due to its precise perception of balanced datasets, unlike existing DC techniques. This section discusses the various aspects of NNs, which contribute to the challenge of data diversification and ultimately the persisting bias in trained networks.

7.1 Input resemblance

As seen from Table 2, the greater the number of output classes, the higher the robustness bias in the NN. This implies that the higher the number of output classes, the more likely is the dataset imbalanced, and the more unlikely it is to obtain a trained NN that is equally robust for all output classes. A likely explanation for this could in fact be a close resemblance of inputs from the different classes, for datasets with a higher number of output classes.

For instance, consider the case of hand-written digits (from the MNIST dataset), which comprises of 10 output classes. As shown in Fig. 6, it is possible for inputs from some classes to closely resemble inputs from other classes—for example, digit 0 may resemble a 6, and digit 2 may resemble a 3. With inputs having likely resemblance to multiple classes, it is challenging to generate realistic synthetic inputs, and hence obtain successful data diversification for reducing the bias.

A more careful study of the example provided above also reveals that the difference between the closely resembling inputs blur when their semantic distance is smaller (Kenett, 2019), as shown in Fig. 6. Yet, the syntactic rules for output classification stay intact even for these closely resembling inputs. For instance, a single loop forms the digit 0, while an arc of a length comparable to half the circumference of the loop is required in addition to the loop to syntactically define the digit 6. Hence, the addition of such syntactic rules for the generation of synthetic inputs [similar to the approach taken in neuro-symbolic learning (Sarker et al., 2021)] may improve the data diversification.

7.2 Curse of dimensionality

Another challenge to data diversification is the large number of input neurons comprising the NN inputs—a challenge often referred to as the “curse of dimensionality” in the NN analysis literature (Wu et al., 2020). This implies that as the number of input neurons for the NN increase, the computational requirements for its analysis increase exponentially.

To understand this from the perspective of data diversification, let us consider the example of an image dataset. Data diversification determines input feature bounds directly from the raw input data to generate inputs such that the synthesized inputs x belong to the valid input \mathcal{D} , i.e., $x \in \mathcal{D}$. However, various transformations, like affine, homographic and photometric transforms associated with image inputs may tremendously change the inputs, while still keeping the inputs realistic (Pei et al., 2017). Hence, for a practical image dataset, inputs belonging to even a single output class will have individual inputs that have undergone different transformations. As a result, the bounds of each input feature obtained from the inputs, for such a dataset, will be very large. This hinders the generation of synthetic data using these bounds, in turn making the data diversification halt at the data validation step since the search input space is too large for the randomly generated inputs to be realistic. (See Appendix B for details on the experimental analysis carried out to test the stated hypothesis on a real-world image dataset, MNIST.)

Towards this end, appropriate input pre-processing and the use of feature correlation knowledge to determine the bounds of the correlated input features (rather than raw input features) could potentially extend the applicability of *UnbiasedNets* framework to a larger variety of datasets.

8 Conclusion

The overall performance of Neural Networks (NNs), particularly those relying on supervised training algorithms, is largely dependent on the training data available. However, the data used to train NNs may often be biased towards specific output class(es), which may propagate as robustness bias in the trained NN. But, unlike checking the testing accuracy of the trained NN, determining the bias in a NN is not a straightforward task. Existing works often rely on large datasets and aim at addressing biases by ensuring an equal number of inputs from each output class. However, as shown by our detailed experiments, such approaches are not always successful. This paper proposes a novel bias alleviation framework *UnbiasedNets*, which initially detects and quantifies the extent of bias in a trained NN and then uses a methodological approach to diversify the training datasets by leveraging the NN’s noise tolerance and K-means clustering. To the best of our knowledge, this is the first framework specifically addressing the *robustness bias* problem. We show the efficacy of *UnbiasedNets*, using both binary and multi-label classifiers in our experiments, and also demonstrate how the existing bias alleviation may rather exacerbate the bias instead of alleviating it. We also discuss the challenges in robustness bias alleviation in certain datasets, and elaborate on the potential future research direction for addressing the robustness bias problem in trained NNs.

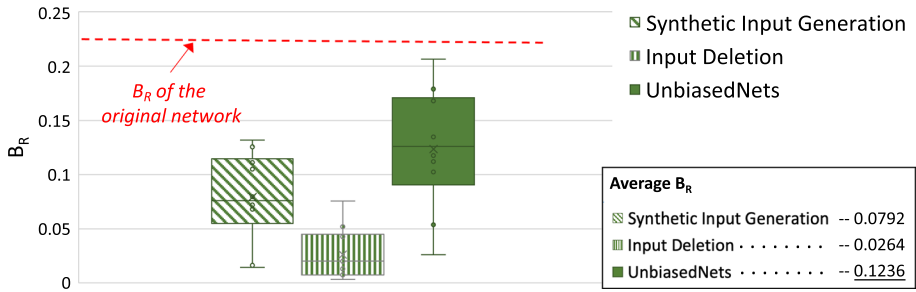


Fig. 7 *UnbiasedNets*, along with both its constituent components, i.e., synthetic input generation and input deletion, successfully diversifies the training dataset. This consequently reduces the bias in trained NN, as indicated by the lower B_R values

Appendix A: Ablation studies

UnbiasedNets provide an overall framework for data diversification, leveraging K-means clustering and the noise tolerance of the trained NNs. As elaborated in Sect. 4, the methodology involves the perceptive addition of synthetic inputs to the existing dataset and consequently, minimizing the redundancy in the dataset using input deletion based on K-means clustering. This section provides ablation studies for both our novel synthetic input generation and deletion (i.e., redundancy minimization), to show the individual effectiveness of each component of the *UnbiasedNets*. The studies make use of the NN trained on the Leukemia dataset (described in detail in Sect. 5.2). This is followed by a discussion to highlight the strengths and weaknesses of the components, motivating the sequential use of components, as adapted in our framework.

A.1 Synthetic input generation

Here, the feature bounds are determined for individual input features using the bounds from the training dataset and the noise tolerance of the trained NN for synthetic input generation. These details of the process are elaborated in Sects. 4.2.1–4.2.4 and Algorithm 1. Synthetic inputs are generated for genetic attributes of AML leukemia (i.e., the output class with less number of samples in the training dataset). The updated dataset is validated (as elaborated in Sect. 4.2.6) to ensure realistic input generation. Consequently, to avoid overfitting during training using this updated dataset, the number of training epochs is reduced to 56, using the learning rates of 0.5 and 0.2 for 28 epochs each, respectively.

A.2 Input deletion

In this study, the inputs are deleted from the output class with more input samples in the training dataset, as described in Sect. 4.2.5. The objective here is to leverage K-means clustering to reduce the redundancy in the training dataset, thereby potentially obtaining a balanced dataset for training. Again, the number of epochs used for training is updated to avoid overfitting, i.e., 69 epochs are used each with the learning rates of 0.5 and 0.2, respectively, during training.

A.3 Results and discussion

The experiments based on studies provided earlier in the section were repeated 10 times to ensure conformity. The resulting \mathcal{B}_R from the experiments are summarized in Fig. 7. It is clearly evident that both components of *UnbiasedNets*, i.e., synthetic input generation and input deletion, aid in the reduction of bias in the trained network.

Additionally, input deletion appears to provide the best overall reduction in the bias of the trained NN. However, it must be noted that the learning capability of NNs is data-driven, i.e., the more the training data available, the more likely is the trained NNs to perform well in real-world applications (Mayer et al., 2016; Xu et al., 2018). Hence, to ensure optimal classification performance and generalization capabilities for trained NNs, the use of input deletion, standalone, is counter-intuitive.

Bias reduction using synthetic input generation, on the other hand, appears only slightly better than that using *UnbiasedNets*. However, since the selection of synthetic input samples is made using the existing data samples, it is possible for the new inputs to closely resemble the existing samples. Hence, not all synthetic inputs may add to the diversity of data.

UnbiasedNets leverages the strengths of both synthetic input generation (by providing a larger training dataset and potentially adding diversity to the data) and input deletion (by removing closely resembling input samples), to provide an overall data diversification framework proposed in Sect. 4. Hence, the framework not only diversifies the training dataset, but also reduces the bias in the trained NN (albeit not as significantly as its individual components), as observed in Fig. 7.

Appendix B: Robustness bias alleviation for image datasets

To test the hypothesis (given in Sect. 7) that input feature bounds obtained directly from the raw input data are often too large for realistic synthetic input generation, we consider the problem of diversifying the MNIST (image) dataset. Details of the experiment, results, and analysis are as follows:

Experimental setup

We trained a LeNet-5 model on the MNIST dataset, which comprises of 60,000 training and 10,000 testing inputs, using Keras. Training and testing accuracies of 99.23% and 98.78% respectively, were achieved in 50 epochs using a batch size of 1024. The FGSM attack was implemented once for all inputs of the testing dataset using Adversarial Robustness Toolbox (ART) (Nicolae et al., 2018). The adversarial noises were recorded for the counterexample analysis (as shown earlier in Fig. 2) and ultimately bias detection.

Results and analysis

The number of inputs from each class in the MNIST testing dataset varies only slightly, as depicted by Fig. 8a. Yet, as highlighted in Sect. 6, the larger the number of output classes, the higher the chances of large robustness bias in the trained network, despite having an equal number of inputs from each class. This was observed in our neural network trained

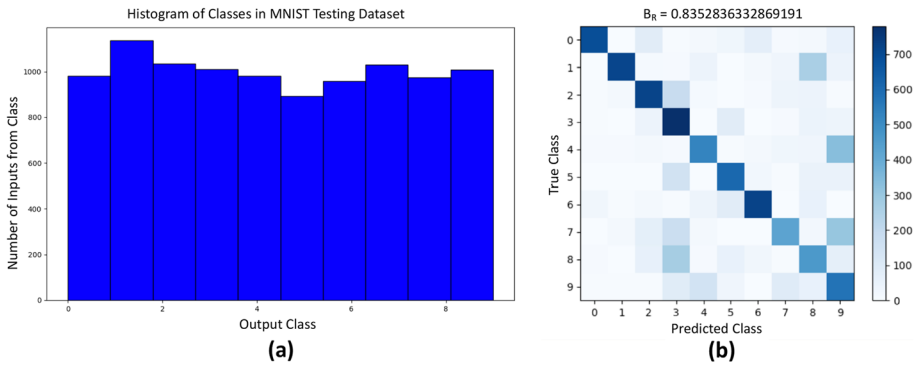


Fig. 8 **a** Number of inputs from each class is approximately equal for MNIST’s testing dataset, **b** the confusion matrix and B_R for LeNet-5 model trained on the original MNIST dataset, with FGSM attack used as a measure for bias detection

on the MNIST dataset (having 10 output classes), which achieved a robustness bias B_R of 0.84, as shown in Fig. 8b.

Moreover, the percentage difference between the correlation coefficients of training and testing datasets in MNIST is in the factor of 10^4 . This is often evident in datasets like imaging datasets, which involve significant affine, homographic and photometric transformations leading to significantly different feature correlations in training and testing datasets. As hinted in Sect. 4.2.6, this suggests that the input features are either independent or the input requires some pre-processing to obtain appropriate feature correlation.

In the case of image inputs, the input features are already known to have spatial correlation (Berryman, 1985), albeit its determination is a non-trivial problem, particularly using only raw inputs (Zhao et al., 2006). As expected, the resulting input feature bounds for the dataset are too large to enable the generation of realistic synthetic inputs to allow diversification. Hence, the use of appropriate input pre-processing, more sophisticated feature correlation measures (Zhao et al., 2006) and the inclusion of the correlated feature correlation knowledge during the bound determination process can potentially allow robustness bias alleviation for a wider range of machine learning datasets.

Acknowledgements This work was partially supported by Doctoral College Resilient Embedded Systems which is run jointly by TU Wien’s Faculty of Informatics and FH-Technikum Wien, and partially by Moore4Medical project funded by the ECSEL Joint Undertaking under grant number H2020-ECSEL-2019-IA-876190. The authors also acknowledge TU Wien Bibliothek for financial support through its Open Access Funding Programme.

Author contributions All authors contributed to the study conception and design of the presented work. Framework implementation, analysis, interpretation of results, and comparison to state-of-the-art were performed by Mahum Naseer and Bharath Srinivas Prabakaran. The first draft of the manuscript was prepared by Mahum Naseer. All authors contributed to the critical review, iterative improvement and approval of the final version of the manuscript.

Funding Open access funding provided by TU Wien (TUW). This work was partially supported by Doctoral College Resilient Embedded Systems which is run jointly by TU Wien’s Faculty of Informatics and FH-Technikum Wien, and partially by Moore4Medical project funded by the ECSEL Joint Undertaking under grant number H2020-ECSEL-2019-IA-876190.

Availability of data and materials All the datasets used in the study are open source. The papers contributing to the inception of the datasets have been cited in the paper (Fisher, 1936; Golub et al., 1999).

Code availability Complete source code to reproduce the framework will be made open source, and the link will be provided with the camera-ready version of the paper.

Declarations

Conflicts of interest No conflict of interest competing interests are applicable to any of the authors.

Ethics approval The submitted work does not require the approval of an Internal Review Board (IRB) as the datasets used in this work are openly accessible for widespread use.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.


References

- Alvi, M., Zisserman, A., & Nellåker, C. (2018). Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *Proceedings of the European conference on computer vision (ECCV) workshops*.
- Arjovsky, M., Bottou, L., Gulrajani, I., & Lopez-Paz, D. (2019). Invariant risk minimization. arXiv preprint [arXiv: 1907.02893](https://arxiv.org/abs/1907.02893).
- Asha, R., & Suresh Kumar, K. R. (2021). Credit card fraud detection using artificial neural network. In *Global Transitions Proceedings* (Vol. 2(1), pp. 35–41).
- Bagui, S., & Li, K. (2021). Resampling imbalanced data for network intrusion detection datasets. *Journal of Big Data*, 8(1), 1–41.
- Barushka, A., & Hajek, P. (2018). Spam filtering using integrated distribution-based balancing approach and regularized deep neural networks. *Applied Intelligence*, 48(10), 3538–3556.
- Becker, F., Drichel, A., Müller, C., & Ertl, T. (2020). Interpretable visualizations of deep neural networks for domain generation algorithm detection. In *Symposium on visualization for cyber security (VIZSEC)* (pp. 25–29).
- Benali, L., Notton, G., Fouilloy, A., Voyant, C., & Dizene, R. (2019). Solar radiation forecasting using artificial neural network and random forest methods: Application to normal beam, horizontal diffuse and global components. *Renewable Energy*, 132, 871–884.
- Benz, P., Zhang, C., Karjauv, A., & Kweon, I. S. (2021). Robustness may be at odds with fairness: An empirical study on class-wise accuracy. In *NeurIPS 2020 Workshop on Pre-registration in Machine Learning NeurIPS 2020 workshop on pre-registration in machine learning* (pp. 325–342).
- Berryman, J. G. (1985). Measurement of spatial correlation functions using image processing techniques. *Journal of Applied Physics*, 57(7), 2374–2384.
- Bhatti, I. T., Naseer, M., Shafique, M., & Hasan, O. (2022). A formal approach to identifying the impact of noise on neural networks. *Communications of the ACM*, 65(11), 70–73.
- Calvo-Pardo, H. F., Mancini, T., & Olmo, J. (2020). Neural network models for empirical finance. *Journal of Risk and Financial Management*, 13(11), 265.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357.
- Dua, D., & Graff, C. (2017). UCI machine learning repository. <http://archive.ics.uci.edu/ml>
- Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, 25(1), 24–29.

- Fink, M., Liu, Y., Engstle, A., & Schneider, S.-A. (2019). Deep learning-based multi-scale multi-object detection and classification for autonomous driving. In *Fahrerassistenzsysteme 2018* (pp. 233–242). Berlin: Springer
- Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2), 179–188.
- Fuangkhon, P. (2022). Effect of the distance functions on the distance-based instance selection for the feed-forward neural network. *Evolutionary Intelligence*, 15(3), 1991–2015.
- Gat, I., Schwartz, I., Schwing, A., & Hazan, T. (2020). Removing bias in multi-modal classifiers: Regularization by maximizing functional entropies. *Advances in Neural Information Processing Systems*, 33, 3197–3208.
- Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., et al. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439), 531–537.
- Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *International conference on learning representations (ICLR)*.
- He, H., Bai, Y., Garcia, E. A., Li, S. (2008). Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)* (pp. 1322–1328).
- Joshi, A. R., Cuadros, X. S., Sivakumar, N., Zappella, L., & Apostoloff, N. (2022). Fair SA: Sensitivity analysis for fairness in face recognition. In *Algorithmic fairness through the lens of causality and robustness workshop* (pp. 40–58).
- Kenett, Y. N. (2019). What can quantitative measures of semantic distance tell us about creativity? *Current Opinion in Behavioral Sciences*, 27, 11–16.
- Khan, S., Ahmad, J., Naseem, I., & Moinuddin, M. (2018). A novel fractional gradient-based learning algorithm for recurrent neural networks. *Circuits, Systems, and Signal Processing*, 37(2), 593–612.
- Kim, B., Kim, H., Kim, K., Kim, S., & Kim, J. (2019). Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9012–9020).
- Kotsiantis, S., Kanellopoulos, D., Pintelas, P., et al. (2006). Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, 30(1), 25–36.
- Le Bras, R., Swayamdipta, S., Bhagavatula, C., Zellers, R., Peters, M., Sabharwal, A., & Choi, Y. (2020). Adversarial filters of dataset biases. In *International conference on machine learning* (pp. 1078–1088).
- Levy, J. L., Khoshgoftaar, T. M., Bauder, R. A., & Seliya, N. (2018). A survey on addressing high-class imbalance in big data. *Journal of Big Data*, 5(1), 1–30.
- Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research*, 18(1), 559–563.
- Li, G., Yang, Y., Qu, X., Cao, D., & Li, K. (2021). A deep learning based image enhancement approach for autonomous driving at night. *Knowledge-Based Systems*, 213, 106617.
- Li, Y., Li, Y., & Vasconcelos, N. (2018). Resound: Towards action recognition without representation bias. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 513–528).
- Li, Y., & Vasconcelos, N. (2019). Repair: Removing representation bias by dataset resampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9572–9581).
- Liu, K., Dolan-Gavitt, B., & Garg, S. (2018). Fine-pruning: Defending against backdoor attacks on deep neural networks. In *International symposium on research in attacks, intrusions, and defenses* (pp. 273–294).
- Ma, X., Niu, Y., Gu, L., Wang, Y., Zhao, Y., Bailey, J., & Lu, F. (2021). Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recognition*, 110, 107332.
- Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., & Brox, T. (2016). A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4040–4048).
- Moody, G. B., & Mark, R. G. (2001). The impact of the MIT-BIH arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine*, 20(3), 45–50.
- Nam, J., Cha, H., Ahn, S., Lee, J., & Shin, J. (2020). Learning from Failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33, 20673–20684.
- Nanda, V., Dooley, S., Singla, S., Feizi, S., & Dickerson, J. P. (2021). Fairness through robustness: Investigating robustness disparity in deep learning *FACCT* (pp. 466–477).

- Naseer, M., Minhas, M. F., Khalid, F., Hanif, M. A., Hasan, O., & Shafique, M. (2020). FANNet: Formal analysis of noise tolerance, training bias and input sensitivity in neural networks. In *2020 design, automation & test in Europe conference & exhibition (date)* (pp. 666–669).
- Nayak, G. K., Rawal, R., Lal, R., Patil, H., & Chakraborty, A. (2022). Holistic approach to measure sample-level adversarial vulnerability and its utility in building trustworthy systems. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4332–4341).
- Nicolae, M.-I., Sinn, M., Tran, M. N., Buesser, B., Rawat, A., Wistuba, M., & Edwards, B. (2018). Adversarial robustness toolbox v1.2.0. CoRR 1807.01069. [arxiv: org/pdf/1807.01069](https://arxiv.org/pdf/1807.01069).
- Pei, K., Cao, Y., Yang, J., & Jana, S. (2017). Deepxplore: Automated whitebox testing of deep learning systems. In *Symposium on operating systems principles* (pp. 1–18).
- Picek, S., Heuser, A., Jovic, A., Bhasin, S., & Regazzoni, F. (2019). The curse of class imbalance and conflicting metrics with machine learning for side-channel evaluations. *IACR Transactions on Cryptographic Hardware and Embedded Systems, 2019*(1), 1–29.
- Rothe, R., Timofte, R., & Gool, L. V. (2018). Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision, 126*(2–4), 144–157.
- Sanh, V., Wolf, T., Belinkov, Y., & Rush, A. M. (2020). Learning from others’ mistakes: Avoiding dataset biases without modeling them. arXiv preprint [arXiv:2012.01300](https://arxiv.org/abs/2012.01300).
- Sarker, M.K., Zhou, L., Eberhart, A., & Hitzler, P. (2021). Neuro-symbolic artificial intelligence: Current trends. arXiv preprint [arXiv:2105.05330](https://arxiv.org/abs/2105.05330).
- Savani, Y., White, C., & Govindarajulu, N. S. (2020). Intra-processing methods for debiasing neural networks. In *Advances in neural information processing systems 33*.
- Tian, Q., Kuang, K., Jiang, K., Wu, F., & Wang, Y. (2021). Analysis and applications of class-wise robustness in adversarial training. In *Proceedings of the conference on knowledge discovery and data mining* (pp. 1561–1570).
- Vu, H. N., Nguyen, M. H., & Pham, C. (2022). Masked face recognition with convolutional neural networks and local binary patterns. *Applied Intelligence, 52*(5), 5497–5512.
- Wang, S., Tang, K., & Yao, X. (2009). Diversity exploration and negative correlation learning on imbalanced data sets. In *2009 international joint conference on neural networks* (pp. 3259–3266).
- Wang, Z., Qinami, K., Karakozis, I. C., Genova, K., Nair, P., Hata, K., & Russakovsky, O. (2020). Towards fairness in visual recognition: Effective strategies for bias mitigation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8919–8928).
- Welch, R. B. (1999). Meaning, attention, and the “unity assumption” in the intersensory bias of spatial and temporal perceptions. In *Advances in psychology* (Vol. 129, pp. 371–387). Elsevier.
- Wu, H., Ozdemir, A., Zeljić, A., Julian, K., Irfan, A., Gopinath, D., & Barrett, C. (2020). Parallelization techniques for verifying neural networks. In *Proceedings of FMCAD* (pp. 128–137).
- Xu, H., Liu, X., Li, Y., Jain, A., & Tang, J. (2021). To be robust or to be fair: Towards fairness in adversarial training. In *International conference on machine learning* (pp. 11492–11501).
- Xu, Z., Yang, W., Meng, A., Lu, N., Huang, H., Ying, C., & Huang, L. (2018). Towards end-to-end license plate detection and recognition: A large dataset and baseline. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 255–271).
- Zhang, Y., Wu, H., Liu, H., Tong, L., & Wang, M. D. (2019). Improve model generalization and robustness to dataset bias with bias-regularized learning and domain-guided augmentation. arXiv preprint [arXiv: 1910.06745](https://arxiv.org/abs/1910.06745).
- Zhao, F., Huang, Q., & Gao, W. (2006). Image matching by normalized cross-correlation. In *International conference on acoustics speech and signal processing proceedings* (Vol. 2, pp. II–II).
- Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Conference on empirical methods in natural language processing* (pp. 2979–2989). Copenhagen, Denmark Association for Computational Linguistics.

Authors and Affiliations

Mahum Naseer¹  · Bharath Srinivas Prabakaran¹ · Osman Hasan² ·
Muhammad Shafique³

Bharath Srinivas Prabakaran
bharath.prabakaran@tuwien.ac.at

Osman Hasan
osman.hasan@seecs.nust.edu.pk

Muhammad Shafique
muhammad.shafique@nyu.edu

¹ Technische Universität Wien (TU Wien), 1040 Vienna, Austria

² School of Electrical Engineering and Computer Science (SEECS), National University of Sciences and Technology (NUST), Sector H-12, Islamabad 44000, Pakistan

³ Division of Engineering, New York University Abu Dhabi (NYUAD), Abu Dhabi, United Arab Emirates