Monitoring bias and fairness in machine learning models: A review

Marley Oliveira Bacelar

UNICAMP - Universidade Estadual de Campinas

May, 2021

1. Introduction

Machine learning algorithms are quickly gaining traction in both the private and public sectors for their ability to automate both simple and complex decision-making processes. The vast majority of economic sectors, including transportation, retail, advertisement, and energy, are being disrupted by widespread data digitization and the emerging technologies that leverage it. Computerized systems are being introduced in government operations to improve accuracy and objectivity, and AI is having an impact on democracy and governance [1].

Numerous businesses are using machine learning to analyze massive quantities of data, from calculating credit for loan applications to scanning legal contracts for errors to analyzing employee interactions with customers to detect inappropriate behavior. New tools make it easier than ever for developers to design and deploy machine-learning algorithms [2] [3] .

Due to the availability of massive data sets, computers have simplified the process of extracting new insights. As a result, algorithms have developed into more sophisticated and pervasive methods for automating decision-making [4]. Algorithms are a collection of sequential instructions that computers follow in order to perform a task. Humans and organizations made decisions about hiring, advertising, criminal punishment, and lending in the pre-algorithm period. These decisions were often governed by federal, state, and local laws that established requirements for decision-making fairness, transparency, and equality. Today, some of these decisions are made completely or heavily affected by machines, whose scale and statistical rigor promise previously unimaginable efficiencies. Algorithms are using massive quantities of macro- and micro-data to manipulate human behavior across a range of domains, from movie recommendations to assisting banks in assessing a person's creditworthiness. Machine learning algorithms rely on multiple data sets, or

training data, that specify the appropriate outputs for particular people or objects. It then develops a model that can be applied to other individuals or objects and predicts their appropriate outputs based on the training data [5].

However, since machines respond differently to similar circumstances involving people and objects, research is beginning to reveal some concerning instances in which algorithmic decision-making falls short of our expectations. As a result, some algorithms risk reproducing and amplifying human biases, particularly those affecting vulnerable groups. Automated risk assessments used by judges in the United States to determine bail and penalty limits, for example, can produce incorrect results, resulting in significant cumulative effects on certain populations, such as longer prison sentences or higher bails imposed on people of color [6] [7].

2. Machine learning models' biases and fairness

By and large, a procedure or judgment is deemed to be fair if it does not discriminate against individuals on the basis of their inclusion in a protected group, such as gender or race. Discrimination is the product of bias. Bias is a deliberate departure from an objective reality. A statistical estimator is biased when it encounters a systemic error that prevents it from convergeing to the true value it is attempting to estimate [8]. Bias can manifest in humans as distorted vision, reasoning, remembering, or judgment, resulting in decisions and results that vary for individuals based on their membership in a protected community. There are various types of bias, including subjective bias on the part of individuals, data bias, developer bias, and institutionalized prejudices that are rooted in the decision's underlying social context. If not addressed, bias will result in unfairness in automated decision-making systems [9].

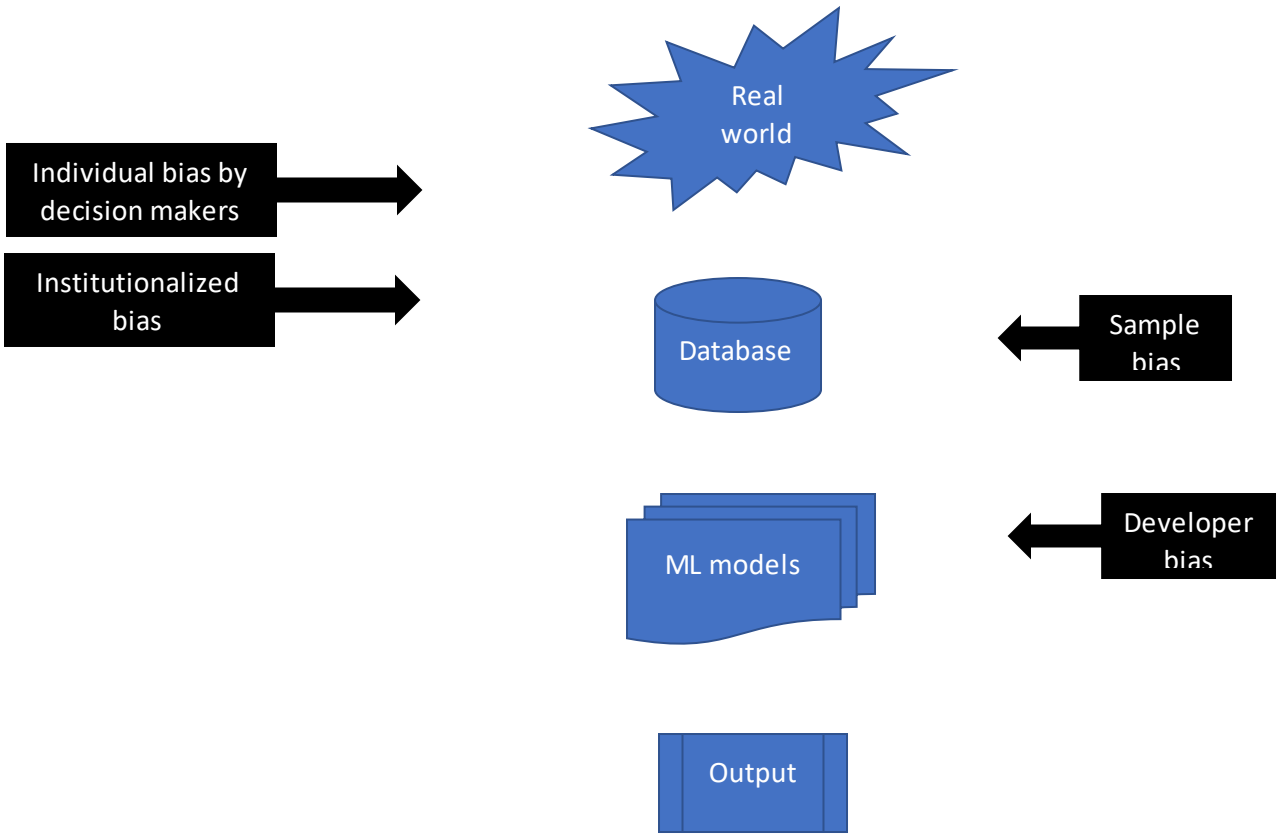

Figure 1. Bias monitoring in different phases

Figure 1 depicts the different types of biases at different stages of Machine learning applications.

a) Institutionalized bias

Frequently, it is the underlying data, rather than the algorithm, that is the source of the problem. Models may be trained on data derived from human decisions or on data derived from the second-order effects of social or historical inequities. For instance, word embeddings (a collection of natural language processing techniques trained on news articles) can represent societal gender stereotypes [10].

Bias can also be introduced into the data through how they are collected or selected for use. In criminal justice models, oversampling some areas due to excessive policing will result in more crime being recorded, which results in increased policing.

b) Individual bias

Due to the inherent existence of human prejudices, biases are expressed in all available data. With the growing popularity and accessibility of Machine Learning and big data, the environment now contains an unfathomable amount of data and readily available resources. In all of this data, there are many ways in which bias can affect the data and our inferences [11].

Biases do not occur solely in the data; they also exist in the scientist's cognition as he or she conducts analysis, experiments, or implements algorithms. Simple errors, such as failing to account for certain parameters or features during testing, may have serious consequences.

c) Sample bias

Another issue is sample bias, which happens when the data sample used to train the algorithm is not representative of the entire population due to systemic data collection errors. Any decision-making system based on this sample will be biased in either direction, favoring or opposing the over- or underrepresented party[12]. There are numerous explanations for a group's over- or under-representation in the results.

d) Algorithm development bias

Apart from data bias as a source of unfairness, there are prejudices introduced during the algorithm's preparation. Throughout the process of developing an algorithm, researchers are confronted with a plethora of decisions that can steer the outcome in a variety of directions, including the selection of the dataset, the selection and encoding of features extracted from the dataset, the selection and encoding of the outcome variable, the rigor in which sources of bias in the data are identified, and the selection and specification of specific auxiliary variables [13]. Each decision is based on an implicit assumption. These assumptions are silent because they appear to be concealed behind an emphasis on mathematical precision and procedural rigor during the algorithm's development. Additionally, in many situations, the underlying assumptions are normative in nature.

The process of designing algorithms that adhere to some of the definitions of algorithmic fairness can be criticized as biased if it fails to consider the social and moral background of the decision being made.

3. Monitoring biases

Generally, approaches for addressing and monitoring algorithmic biases fall into three categories [6]:

(1) Preparation.

Pre-processing methods attempt to turn the data in such a way that any inherent discrimination is eliminated. Pre-processing can be used if the algorithm is allowed to change the training data.

(2) The in-processing stage.

In-processing techniques attempt to adjust and alter cutting-edge learning algorithms in order to eliminate discrimination during the model training process. If the learning protocol for a machine learning model can be changed, then in-processing can be used during model training—either by adding changes to the objective function or implementing a constraint.

(3) The post-processing stage.

After preparation, post-processing is done by referencing a holdout collection that was not used during the model's training. If the algorithm can only handle the learned model as a black box with no ability to alter the training data or learning algorithm, then only post-processing can be used, in which the labels allocated by the black-box model are initially reassigned based on a function.

Bias and fairness measurement methods

a) Metrics Focused on Parity

This metric, one of the earliest descriptions of fairness, describes fairness as an equal probability of receiving a positive mark. That is, each category has an equal probability of being classified as positive. However, a downside of this notion is that it ignores possible variations between classes [14].

As with statistical parity, this classification considers the likelihood of being graded as positive. In comparison to parity, "Disparate Effect" takes the ratio of disadvantaged to privileged classes into account. Its roots are in legal fairness considerations for selection processes, which often use the 80 percent rule to determine whether or not a process has a differential effect (ratio less than 0.8) [6].

b) Metrics Based on Confusion Matrix

Additionally, confusion matrix-based metrics include True Positive Rate (TPR), True Negative Rate (TNR), False Positive Rate (FPR), and False Negative Rate (FNR) (FNR). The benefit of these metrics is that they allow for the inclusion of underlying discrepancies between groups that are not captured by parity-based approaches [15].

		Predicted Classification		
		$\hat{Y} = 1$	$\hat{Y} = 0$	
Outcome	$Y = 1$	True Positives (TP)	False Negatives (FN)	False Negative Rate (FNR) $FN/(TP+FN)$
	$Y = 0$	False Positives (FP)	True Negatives (TN)	False Positive Rate (FPR) $FP/(FP+TN)$
		False Omission Rate (FOR) $FP/(TP+FP)$	False Discovery Rate (FDR) $FN/(FN+TN)$	

Table 1. Confusion matrix

Due to the fact that parity and unequal effect do not account for possible disparities between the groups being compared, consider additional metrics that account for FPR and TPR differences between groups. To be precise, an algorithm is considered fair under equal opportunity if its TPR is consistent across classes. As with equal opportunity, equalized odds considers both TPR and FPR, i.e., the percentage of real negatives expected as positive.

c) Metrics focused on Calibration

In contrast to previous metrics, which were described on the basis of expected and actual values, calibration-based metrics take into account the predicted likelihood, or score. Calibration seeks to ensure that the likelihood of y = 1 remains constant for a given score. Calibration by well is a subset of normal calibration in which the probability of being in the positive class must also match the particular score [16].

Bayesian Fairness extends the principle of equilibrium to situations in which model parameters are unknown. Bayesian fairness takes into account situations in which the perceived utility of a decision maker must be measured against the decision's fairness. The model considers the likelihood of various outcomes (probabilities of model parameters) and the resulting fairness / unfairness.

d) Approaches to Binary Classification

The majority of strategies for mitigating unfairness, bias, or discrimination are focused on the concept of protected or sensitive variables (we will use the words interchangeably) and (un)privileged classes: groups (often identified by one or more sensitive variables) that are statistically (less) more likely to be rated positively. Before delving into the critical components of the fairness system, it is necessary to address the essence of protected variables. Protected

variables identify the characteristics of data that are socio-culturally precarious for machine learning applications. Sex, ethnic origin, and age are often used examples (as well as their synonyms) [17]. However, the term "protected variable" may refer to any aspect of data that includes or concerns humans.

The literature is dominated by methods for eliminating bias and unfairness in machine learning as applied to the binary classification problem class.

There are several explanations for this, but the most important are as follows:

1) Some of the domain's most contested application areas are binary.

2) Quantifying fairness on a binary dependent variable is more mathematically convenient; solving multi-class problems will at the very least add terminology to the fairness quantity.

e) Debilitating/ Blinding

Blinding is a technique that makes a classifier "immunize" itself against one or more sensitive variables. For example, if there is no discernible outcome distinction based on the variable gender, a classifier is gender blind. For instance, suppose we want to train a gender blind classifier so that the loan rate given to each of the two gender groups is equivalent for both genders. Blinding has been used in other works to refer to the exclusion of critical variables from training results. It has been shown that omission reduces model accuracy and does not improve model discrimination. Both omission and immunity ignore relationships with proxy variables, which may result in an increase in rather than a decrease in bias and discrimination, or in concealment of discrimination indirectly [18]. Additionally, it ignores the fact that prejudice may not be caused by a single variable, but rather by a combination of several variables; as such, evaluating which combination(s) of variables to blind is not easy, and the phenomenon of omitted variable bias should not be minimized. Similarly, researchers continue to use omission in their assessment methodologies in order to equate their findings to prior work and serve as a benchmark. Additionally, omission can be used in particular components of the fairness approach, such as temporarily omitting sensitive variables prior to transforming the training data. Blinding (or partial blinding) has also been used as a tool for conducting fairness audits [19]. More precisely, these approaches investigate the effect of partially blinding features (sensitive or otherwise) on model results. This is analogous to causal models and can assist in identifying problematic sensitive or proxy variables through black-box-like analysis of a machine learning model.

e) Modification/transformation

Transformation approaches learn a new representation of the data, frequently as a mapping or projection function, that ensures fairness while maintaining the fidelity of the machine learning task. Currently available transformation techniques are limited to numeric data, which is a major limitation [20]. There are several approaches to data transformation: operating on the dependent variable, operating on non-sensitive numeric variables, mapping individuals to an input space that

is independent of safe subgroupings, and transforming the distribution of model predictions to meet unique fairness objectives. There are connections between blinding (in the immunity sense) and independence mappings, since both methods aim to achieve the same goal: independence from one or more unique sensitive variables. Other types of transformation include re-labelling and perturbation, which we classify separately. To illustrate the concept of transformation, we will turn the distribution of SAT scores toward the median in order to "degender" the original distribution into one that preserves only the rank order of individuals regardless of gender. This is equivalent to obfuscating knowledge about protected variables from a set of covariates. To maintain predictive ability, transformation approaches often aim to preserve rank orders within transformed variables.

The degree of transformation (fairness repair) and the impact on classifier performance are inextricably linked. To address this, approaches mostly rely on partial repair: transforming the data into a target distribution, but not entirely, in order to offset this trade-off. While transformation is primarily a pre-processing technique, it can also be used during the post-processing phase.

When applying transformation techniques, there are a variety of considerations to make:

1) The transformed data should be identical to the original data; otherwise, the amount of "repair" will reduce the utility of the generated classifier and, in general, result in data loss.

2) Understanding the relationships between sensitive and possible proxy variables; as such, machine learning researchers may wish to first understand these relationships using causal methods before applying transformation methods;

3) Choosing "fair" target distributions is not simple;

4) Finding a "optimal" transformation often involves an optimization stage, which can be computationally costly in cases of high dimensionality, even under convexity assumptions.

5) Missing data presents unique challenges for transformation methods, as it is unclear how such data samples should be treated. Many resolve this by omitting these samples, but this can pose additional methodological concerns;

6) Transformation reduces the interpretability of the model, which can conflict with existing data protection legislation; and

7) There is no guarantee that a transformed data set has "corrected" for potentially unequal latent relationships between proxy variables.

f) re-weighing

Unlike transformation, which modifies (certain instances of) the data, reweighing assigns weights to instances of the training data without altering the data itself.

Weights may be used for a variety of purposes [21]:

1) to denote the frequency of occurrence of an instance kind,

2) To assign a lower/higher weight to "sensitive" training samples, or

3) To increase the stability of the classifier.

Reweighing as a technique straddles the line between pre- and post-processing. For instance, assigning weights that account for the probability of an instance belonging to a particular class and sensitive value pairing (pre-processing). A related approach is to define sensitive training instances (pre-processing), but then to learn weights for these instances (in-processing) in order to optimize for the selected fairness metric. Reweighing, when done properly, will achieve a high level of precision when opposed to re-labelling and blinding (omission) methods. However, classifier robustness and stability may be a problem. Thus, machine learning researchers must carefully analyze how reweighting methods are implemented and conduct effective model stability checks. Additionally, reweighing subtly alters the data composition, rendering the process less transparent.

g) Regularization and Optimization of Constraints

Regularization has traditionally been used in machine learning to penalize the difficulty of the learned hypothesis in order to prevent over-fitting [22]. Regularization techniques, when applied to fairness, apply one or more penalty words that penalize the classifier for discriminatory practices. Thus, it is not hypothesis-driven (or learned model-driven), but data-driven and founded on the considered notion(s) of justice. Most of the literature extends or augments the classifier's (convex) loss function with fairness terms, usually in an attempt to strike a balance between fairness and accuracy. Frequently, approaches to fair machine learning are not stable, i.e., small changes in the training data have a major effect on results (comparatively high standard deviation). During model training, constraint optimization approaches often incorporate notions of fairness into the classifier loss function operating on the confusion matrix.

The following are significant obstacles to regularization approaches:

1) they are often naturally non-convex or achieve convexity at the expense of probabilistic interpretation;

2) Not all tests of justice are affected equally by the strength of daily

3) Different regularization terms and penalties have varying effects on different data sets, implying that this option may have a qualitative impact on the accuracy-fairness trade-off.

It can be challenging to reconcile conflicting constraints in constraint optimization, resulting in more complicated or unstable preparation.

h) Thresholding

Thresholding is a post-processing technique inspired by the fact that discriminatory judgments are often taken near to decision-making limits as a result of the decision maker's bias and that humans make decisions using threshold laws.

Threshold methods also aim to identify regions of a classifier's posterior probability distribution where favored and safe classes are categorized positively and negatively. Such cases are deemed vague, and therefore susceptible to bias. To address this, researchers have developed methods for determining threshold values for various protected groups using measures such as equalized odds in order to strike a balance between true and false positive rates and thus reduce predicted classifier loss. The underlying concept is to incentivize superior performance across all classes and groups (in terms of both fairness and accuracy). Calculating the threshold value(s) can be done manually to allow a consumer to express expectations about the fairness-accuracy trade-off, or by the use of other statistical methods [23].

i) reabeling Changes and Perturbation

Relabelling and perturbation approaches are a subset of transformation approaches in that they explicitly alter the distribution of one or more variables in the training data by flipping or modifying the dependent variable [24]. Relabeling, also known as data-massaging, is the process of modifying the labels of training data instances in order to ensure that the proportion of positive instances is comparable for all covered classes. Additionally, it can be applied to test data using the techniques or probabilities gained from the training data. Sometimes, but not always, methods aim to preserve the overall class distribution, i.e. to maintain the same number of positive and negative instances.

j) Adversarial learning [25]

The goal of adversarial learning is for an adversary to decide if a model training algorithm is sufficiently robust. As applied to applications of fairness in machine learning, an adversary instead tries to ascertain the fairness of the training process, and when it is not, the adversary's input is used to refine the model.

A benefit of adversarial methods is that they can take into account several fairness constraints, often by treating the paradigm as a blackbox. However, it has been documented that adversarial methods often lack stability, making them difficult to train consistently and, more precisely, in certain transfer learning situations where the protected variable is established for a limited number of samples.

4. Conclusion

There are two promising strategies for avoiding inequity caused by skewed data. One is to gain a thorough understanding of the data used. The other objective is to enhance the data's quality. Causal inference techniques will become more important as a means of comprehending the underlying mechanisms of a decision process and identifying sources of bias in data.

This is consistent with increasing researchers' knowledge of the data they use and raising awareness of the critical nature of understanding the data generation process. Additionally, a focus on diversity of data is critical in addressing issues of discrimination against minorities.

Additional research on real-world applications of automated decision-making systems is required to better understand how human decision makers interact with machine support. This also implies that algorithms will need to be reevaluated on a regular basis. No algorithm can remain fair in perpetuity without periodic adjustments. Increased diversity becomes critical once again when considering subtle representational harms. Not only the diversity of the data is critical in this case, but also the diversity of the group of developers of automated decision making systems.

It's tempting to believe that once educated, a machine-learning model can operate autonomously. In practice, the world in which the model operates is constantly evolving, and managers must retrain models on new data sets on a periodic basis.

Machine learning is one of the most groundbreaking technological capabilities to emerge in the last decade, bringing real-world business value. When combined with big data technology and the vast computing power available via public cloud computing, machine learning has the potential to transform how people communicate with technology, and potentially entire industries. However, as exciting as machine learning technology is, careful preparation is needed to prevent unintended biases.

The designers of the machine-learning models that will power the future must understand and monitor the effects of bias on the effectiveness of the decisions made by the machines. Otherwise,

decision makers risk undermining the possible benefits of machine learning by creating models with biases.

References

[1] R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke, "Explainable Machine Learning for Scientific Insights and Discoveries," *IEEE Access*, 2020.

[2] C. Reis, P. Ruivo, T. Oliveira, and P. Faroleiro, "Assessing the drivers of machine learning business value," *J. Bus. Res.*, 2020.

[3] A. Kollár, "Betting models using AI: A review on ANN, SVM, and Markov Chain," 2021.

[4] J. Qiu, Q. Wu, G. Ding, Y. Xu, and S. Feng, "A survey of machine learning for big data processing," *Eurasip Journal on Advances in Signal Processing*. 2016.

[5] E. Lee, "How do we build trust in machine learning models?," *Available SSRN 3822437*, 2021.

[6] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *arXiv*. 2019.

[7] K. Żbikowski and P. Antosiuk, "A machine learning, bias-free approach for predicting business success using Crunchbase data," *Inf. Process. Manag.*, 2021.

[8] B. Hutchinson, V. Prabhakaran, E. Denton, K. Webster, Y. Zhong, and S. Denuyl, "Unintended machine learning biases as social barriers for persons with disabilitiess," *ACM SIGACCESS Access. Comput.*, 2020.

[9] M. Thelwall, "Gender bias in machine learning for sentiment analysis," *Online Inf. Rev.*, 2018.

[10] H. Jiang and O. Nachum, "Identifying and correcting label bias in machine learning," *arXiv*. 2019.

[11] C. Wang, B. Han, B. Patel, F. Mohideen, and C. Rudin, "In pursuit of interpretable, fair and accurate machine learning for criminal recidivism prediction," *arXiv*. 2020.

[12] A. Završnik, "Criminal justice, artificial intelligence systems, and human rights," *ERA Forum*, 2020.

[13] A. Yapo and J. Weiss, "Ethical Implications of Bias in Machine Learning," in *Proceedings of the 51st Hawaii International Conference on System Sciences*, 2018.

[14] P. Besse, E. del Barrio, P. Gordaliza, J. M. Loubes, and L. Risser, "A survey of bias in machine learning through the prism of statistical parity for the Adult data set," *arXiv*. 2020.

[15] Abhishek Sharma, "Confusion Matrix in Machine Learning," *Www.Geeksforgeeks.Org*, 2018.

[16] A. Cappello, D. Lenzi, and L. Chiari, "Periodical in-situ re-calibration of force platforms: A new method for the robust estimation of the calibration matrix," *Med. Biol. Eng. Comput.*, 2004.

[17] S. A. Abdullah and A. Al-Ashoor, "An artificial deep neural network for the binary classification of network traffic," *Int. J. Adv. Comput. Sci. Appl.*, 2020.

[18] S. Chen, N. Carlini, and D. Wagner, "Stateful Detection of Black-Box Adversarial Attacks," in *SPAI 2020 - Proceedings of the 1st ACM Workshop on Security and Privacy on Artificial Intelligent, Co-located with AsiaCCS 2020*, 2020.

[19] P. Saleiro *et al.*, "Aequitas: A bias and fairness audit toolkit," *arXiv Prepr. arXiv1811.05577*, 2018.

[20] R. K. E. Bellamy *et al.*, "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM J. Res. Dev.*, vol. 63, no. 4/5, pp. 1–4, 2019.

[21] P. Fox-Roberts and E. Rosten, "Unbiased generative semi-supervised learning," *J. Mach. Learn. Res.*, 2014.

[22] J. Chen *et al.*, "A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide," *Environ. Int.*, 2019.

[23] R. Zhao, R. Yan, Z. Chen, K. Mao, P. Wang, and R. X. Gao, "Deep learning and its applications to machine health monitoring," *Mechanical Systems and Signal Processing*. 2019.

[24] E. Wong and J. Z. Kolter, "Learning perturbation sets for robust machine learning," *arXiv*. 2020.

[25] Y. Vorobeychik and M. Kantarcioglu, "Adversarial machine learning," *Synth. Lect. Artif. Intell. Mach. Learn.*, 2018.