On the Use of Causal Models to Build Better Datasets

On the Use of Causal Models to Build Better Datasets

Fabio Garcea
Dep. of Control and Computer Eng.
Politecnico di Torino
Italy, Turin
fabio.garcea@polito.it

Lia Morra
Dep. of Control and Computer Eng.
Politecnico di Torino
Italy, Turin
lia.morra@polito.it

Fabrizio Lamberti
Dep. of Control and Computer Eng.
Politecnico di Torino
Italy, Turin
fabrizio.lamberti@polito.it

***Abstract*—In recent years, Machine Learning and Deep Learning communities have devoted many efforts to studying ever better models and more efficient training strategies. Nonetheless, the fundamental role played by dataset bias in the final behaviour of the trained models calls for strong and principled methods to collect, structure and curate datasets prior to training. In this paper we provide an overview on the use of causal models to achieve a deeper understanding of the underlying structure beneath datasets and mitigate biases, supported by several real-life use cases from the medical and industrial domains.**

***Index Terms*—deep learning, machine learning, causal models, dataset bias, causal analysis**

I. Introduction

Machine Learning (ML) and Deep Learning (DL) represent nowadays the high performance solutions for a large number of classification tasks from autonomous driving [1] to the natural sciences [2], to disease diagnosis in the medical domain [3]. It is well known that the quality of a trained ML model is a direct reflection of the underlying dataset, as summarized by the popular expression *Garbage In, Garbage Out*. Yet, while the bulk of ML/DL research is model-centric and focused on enhancing neural networks by introducing novel training strategies, or diagnosing ML models after training, only recently the research community started to adopt a more data-centric approach [4].

Dataset bias may affect the performance of a ML model during its operational life, causing the model to exploit spurious correlations and preventing generalization to unseen data [5]–[9]. Mounting evidence gathered from the scientific literature and the general press highlights how biased ML models may lead to the perpetuation of social and racial biases, exacerbating discrimination and unfair outcomes [9], [10], or may have serious implications in domains in which robustness and safety are critical [11], [12].

Deep neural networks (DNNs), due to their high nonlinearity and black-box nature, are particularly prone to dataset biases. In fact, DNNs tend to learn – and consequently rely on – shortcuts to solve a specific task, a behavior which has been traced to several properties of DNNs [13]. As a simple example, if a model that recognizes cats and dogs is exclusively trained on images of cats with an overlayed text, it may learn an association between the cat class and the presence of text, and even rely solely on the latter. This example may seem far fetched, but has actually occurred in real medical problems, as hospital archival systems often superimpose textual information on X-ray scans [13].

Another crucial phenomenon to consider is that of *concept drift* [14], which occurs when the statistical properties of the training data and the deployment data diverge over time. Imagine, as an illustrative scenario, the appearance over time of new species of cats and dogs that were not originally included in the training data. Would the cats vs dogs classifier still be able to generalize?

These considerations strongly suggest that a more systematic and robust approach should be adopted when collecting, structuring and characterizing the development datasets used to train a DL model. At the state of the art, many techniques try to diagnose ML models after they have been trained [9]. However, assessing the quality of the dataset *prior to training* could bring multiple advantages. It would allow all stakeholders to be aware of the characteristics and potential pitfalls of a dataset, as well as document the underlying assumptions in the data collection and generation process in a clear and transparent fashion that can be easily validated or integrated by domain experts. It would also help practitioners to select more effective training, collection, annotation and data augmentation strategies. Most importantly, it would enable them to detect biases and other limitations in the current dataset, and possibly resolve or at least anticipate any issues in the resulting ML models. This approach would ultimately lead to models that are less prone to biases and enhance their robustness and generalization capabilities.

Several techniques have been proposed to assist practitioners in this task [7], [15]. One of most interesting techniques is represented by DAGs (directed acyclic graphs) that offer a formal and visual representation of the causal relationships between the variables related to a problem or domain [16]–[19]. A causal diagram does not imply assumptions about the mathematical properties of these relationships. It provides an explicit and principled way to specify assumptions, enabling transparent scrutiny of their plausibility and validity. By accounting for factors that affect the underlying data generation and collection process, causal models can help preventing biases in the dataset, and anticipating the nature of concept

shift. Causal models have been proposed to analyze both structured [20] and unstructured data [19]. A recent paper by Castro and colleagues provides a comprehensive analysis of the role of causal models in medical imaging [19].

Moving from this inspiring body of literature, our goal is to investigate the application of causal models in dataset design, extending the work beyond the medical domain with a specific focus on DL and Computer Vision (CV) applications. We thus analyze causal diagrams for different use cases, both taken from the literature and from real-life industrial R&D projects. Furthermore, we connect causal diagrams with other techniques that have been proposed to characterize datasets.

The rest of the paper is organized as follows. In Sections II III we briefly summarize the principles of causal analysis that are recalled throughout the paper. In Section IV we illustrate two scenarios in which causal models have been used to characterize the dataset and the task. In Section V we connect the proposed strategy to other works focusing on dataset characterization in the ML/DL field.

II. Background

In this section, the main principles of causality theory are briefly introduced. We closely follow the terminology introduced by previous seminal works, to which the reader is referred for further details [17]–[19], [21].

Formally, a causal model is a DAG consisting of nodes (representing variables or factors) and arrows, also known as edges (representing causal relationships between variables). A direct causal relationship between A and B ($A \rightarrow B$) represents the notion that a direct experimental manipulation of A would change the likelihood of B, holding everything else constant. An important principle in causal inference is that the distribution of the cause $P(A)$ does not influence the conditional distribution $P(B \mid A)$, a principle known as independence of cause and mechanism [21]. An *intervention* is defined as any forced change to the value or distribution of a node, regardless of its direct cause, and results in a modified graph wherein this node is disconnected from its parents (or, in other words, the corresponding path is blocked), though crucially all other mechanisms are unaffected. Typical interventions used in experimental sciences and ML/DL are randomization, stratification and controlling by a given variable when building a statistical model.

Causality theory introduces three main canonical relationships between three (or more) variables: *confounders*, *mediators* and *colliders*. For large graphs, one should reason in terms of paths [19]. In particular, a mediator B is a variable which connects an indirect cause A to the final effect C ($A \rightarrow B \rightarrow C$). Here, controlling for B (e.g., by conditioning a statistical model on the value of B) removes the link between A and B, and thus completely screens off the effect of A. A confounder C is a common cause of two variables A and B (formalized as $A \leftarrow C \rightarrow B$). Consider a variable A that is at the same time the effect of multiple independent causes B and C: A is said to be a collider of B and C (formalized $B \rightarrow A \leftarrow C$).

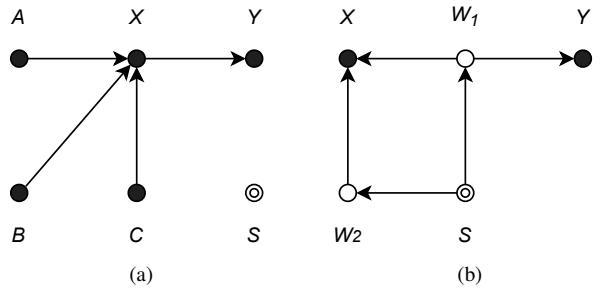

Fig. 1: Example causal models for a causal (a) an anticausal (b) task. X and Y represent the input and output variables, respectively. The selection variable is represented with a double line circle. Empty circles indicate hidden (latent) variables.

III. Causal Models for Datasets

In this section, recent literature on i) how to represent a task in causal terms and ii) how to model the data generation and selection processes will be reviewed. We will refer primarily to DNNs, but many considerations apply also to other classes of ML algorithms.

When modelling a ML task, defined as a mapping function $f : X \mapsto Y$ between an input space X and an output space Y, a key question that has to be answered is whether the task is *causal* or *anticausal* [18]. A task is defined as causal when the goal is to estimate $P(Y \mid X)$ when $X \rightarrow Y$, in other words, if we try to estimate the conditional distribution of an output Y which is an effect of the input X. Conversely, a task is defined as anticausal if the goal is to estimate $P(Y \mid X)$ when $Y \rightarrow X$. Anticausal problems, as differently as it may seem, are ubiquitous in ML/DL [18]. A special case of anticausal task occurs when there is not a direct relationship between X and Y, but both have a common unobserved common cause (confounded). The distinction between causal and anticausal tasks may not be trivial, depending also on the level of information available about the data collection process. Toy examples of causal and anticausal tasks are reported in Fig. 1a and 1b, respectively. Some practical examples will be provided in Section IV.

Causal diagrams allow practitioners to link the characteristics of their problem to properties established by ML and statistical theory [17]–[19], [22], and to select the most appropriate training and statistical methods accordingly. For instance, SSL (semi-supervised learning) techniques should give little benefits for causal tasks [18], meaning that other training strategies should be prioritized. This fact stems from the observation that, in SSL, we have access only to unlabelled data, hence to the distribution $P(X)$, which for the principle of independence of cause and mechanism should be uninformative with respect to $P(Y|X)$ if $X \rightarrow Y$. On the other side, SSL could work for anticausal tasks.

A causal model should not be limited to the input and target variables, but should also include all factors (either observed or hidden) that may influence their distribution. Causal diagrams

allow us to define and assess the role of confounders/colliders in a robust way. The confounding influence can introduce what is often called a *spurious* correlation, i.e., when the considered variables are statistically correlated through another variable but have no causal relationship between each other. Controlling for a confounder blocks the corresponding path, effectively removing the spurious correlation. This may be achieved in several ways, e.g., by introducing the confounder as an additional variable to the ML model, or by stratified sampling. On other hand, controlling for a collider (or its descendants) introduces an association between the two otherwise independent causes, and hence should be avoided. A vast body of literature in the epidemiological sciences exploits causal models for determining which variables should be controlled for [23].

Colliders play a fundamental role in the appearance of dataset bias [22], [24]. Let us consider again our toy example in the introduction, in which one latent variable represents the class A (cat vs. dog) and a second latent variable B represents the confounder (e.g., whether the image contains text or not). The image X is a collider of both A and B and, depending on how the dataset is selected, when computing $P(Y \mid X)$ a strong confounding signal is introduced. Since the confounding signal may be easier to learn than the true signal, as postulated by several authors, collider bias may deeply impact the learning process and result in unfair or non-generalizing models [22], [24], [25]. This scenario can be well represented by the causal model depicted in Fig. 1b.

Causal diagrams are also useful in understanding how dataset shift could manifest and clarify the role played by confounders. Statistical ML is based on the implicit assumption that data samples are independent and identically distributed. Machines, however, often perform poorly when faced with problems that violate these assumptions. In ML/DL, we define *domain* as a combination of a feature space $\mathcal{X}$ and the related marginal distribution $\mathcal{D} = \{\mathcal{X}, P(X)\}$. A dataset or domain shift occurs when the training set distribution $\mathcal{D}^t = \{\mathcal{X}^t, P(X)^t\}$ is different from the testing set $\mathcal{D}^s = \{\mathcal{X}^s, P(X)^s\}$.

Finally, a central aspect to consider is the *selection* scheme, indicated by one or more selection variables S, as selection may block or unblock paths and thus dampen or amplify the effect of colliders. Generally speaking, the selection may be directed by one or more variables composing the causal model, including the input X and the output Y, or may be performed randomly. The selection could be either *explicit*, e.g., made by the ML practitioner, or *implicit*, e.g., driven by the availability of data. Likewise, selection variables can be used to model how data is divided in the training, validation, and testing sets, when for practical reasons the training set is not representative of the entire training set.

It is easy to understand how a certain amount of domain knowledge is necessary to include as many variables as possible in the causal model for a target task and to correctly postulate the causal correlation between them. As such, the causal model building phase should be conducted or, at least, supervised by a domain expert.

In conclusion, to draw a complete causal model to represent the data generation, collection and annotation process, the following methodology (adapted from [19]), can be followed:

- gather information about the data collection, annotation, and selection processes to reconstruct a complete model, including relevant mediators, confounders and colliders;
- determine whether the task is causal or anticausal;
- identify possible mismatches between the training and testing set: consider applying data augmentation, domain adaptation, or resampling strategies to tackle them, depending on the nature of the domain shift;
- determine whether the data collection was biased with respect to the input X, the target Y or any other variable;
- draw the causal model: include all factors and draw the causal relationships between the them; particular attention should be paid to the emergence of collider biases;
- decide if/how further selection (randomization or stratification) should be conducted to control for confounders.

IV. Use Cases

In this Section, we present two scenarios, one taken from the medical domain and originally proposed by [19], one developed for an industrial CV application.

A. Medical Imaging Field

The seminal work by Castro et al. presents a comprehensive view of the application of causal diagrams to the characterization of medical imaging tasks [19]. In fact, many medical imaging analysis tasks share a common workflow, and are characterized by a relatively constrained pool of factors, and their role in the construction, sampling and annotation of datasets has been extensively studied in literature [11], [24].

Building on this body of literature, Castro and colleagues propose a general causal model structure (reported in Fig.2) that likely suits the vast majority of medical imaging ML tasks. By including selection and domain variables, it is possible to model different types of dataset biases and domain shifts. Particularly interesting is the introduction of a selection variable D which models factor(s) that possibly differ between the training and testing population. Different types of domain shifts thus arise depending on the resulting casual model and the type of task (anticausal vs. causal), for which appropriate countermeasures were suggested [19].

Let us see how this general model can be adapted to a sample medical application, such as the diagnosis of cancer (e.g., prostate cancer), from an image (e.g., a magnetic resonance image). The task is thus to predict the probability of the presence or absence of cancer (Y) from the input image X. Typically, the training set for such a model would be based on datasets collected from one or, ideally, multiple institutions.

The image X is the effect of several causes, namely presence of the disease, patient anatomy, and acquisition conditions. Patient anatomy is an internal hidden variable

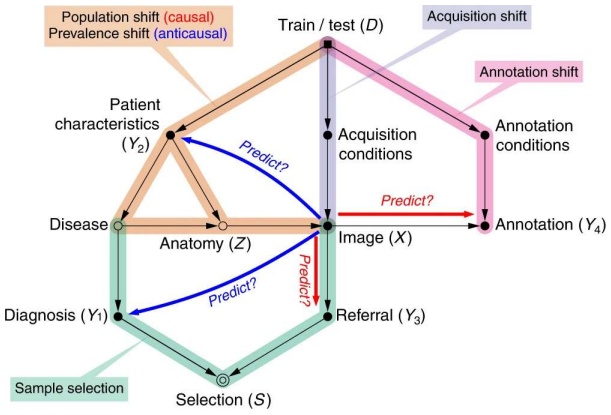

Fig. 2: Causal model template for dataset analysis in the medical imaging domain. Filled circles represent measured values, empty circles hidden variables, and double-line empty circle selection variables. Reproduced from [19].

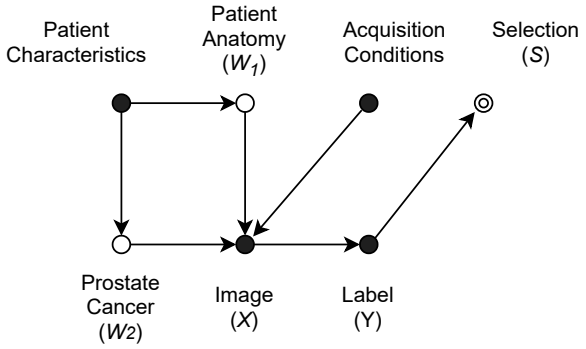

Fig. 3: Causal model representation for prostate cancer classification. Hidden factors of interest are depicted with empty circles.

which accounts for inter-patient variability, whereas acquisition conditions represent all factors (type of scanner, acquisition protocol, etc.) that may influence image appearance.

To determine whether the task is causal or anticausal, it is crucial to understand how the reference standard is established. Whenever possible, presence or absence of the disease should be defined by means of biopsy and/or follow-up for a suitable period of time [11]. Hence, the presence or absence of the disease is known, and manipulating or changing the image would not alter the label: in this case, the task would be anti-causal. In fact, many computer-aided diagnosis tasks are anticausal in nature.

On the other hand, if the labels are established based solely on the radiologist report, then the task would be considered causal, as depicted in Fig. 3. In this case, the true disease status W_2 is considered a hidden variable, and the image X is a mediator between the disease W_2 and the output label Y. In this case, manipulating the image could alter the radiologist perception, and hence the resulting labels. Practical situations could, however, be more nuanced. For instance, biopsy is typically performed only for cases which the radiologist deems suspicious. This factor can be represented by adding a selection variable *biopsy* which is an effect of the label Y.

Finally, the input image X depends on the acquisition conditions: different acquisition systems may, for example, produce a diverse set of inputs in terms of resolution, contrast and other visual features. Several opportunities for *collider bias* emerge when the acquisition conditions are not randomly distributed across the patient population, e.g., if different acquisition protocols are used for high-risk or low-risk populations.

B. Industrial Field

In this subsection, we illustrate the causal diagram for the task of estimating road conditions, in particular to identify the presence of wet road conditions form video frames captured by surveillance Full-HD cameras. The frames have been collected from areas characterized by different features in terms of illumination, road morphology, and point of view, to name a few. The causal diagram reported in Fig. 4 clarifies the role played by different factors in the data collection, sampling, and annotation process.

The main observed variables are the image (X) and the manually determined label (Y), a binary value indicating whether the road appears to be dry or wet. The phenomenon of interest is the presence of water W, which in turn is caused by the weather (rain/snow) or, possibly, by other causes (e.g., floods): these are the crucial *hidden* variables that we need to indirectly estimate. We consider these variables to be hidden because they are not directly measured in our experimental setup (i.e., on-the-road sensors were not available). In causality terminology, the input image X is a *mediator* between the phenomenon of interest W and the label Y.

The camera site, time of day and month (or season) are confounders, as they are indirect causes of both W and X (highlighted in light blue in Fig. 4). For simplicity, the site variable embeds multiple information such as geographical location, road morphology, the type of asphalt, the frequency of car passing, and many other characteristics. These factors influence, directly or indirectly, the type (population shift) and frequency (prevalence shift) of the wet road events [19]. At the same time, images taken at different sites have distinctive visual features due to the different road morphology, illumination, and so forth: hence the site also contributes to the domain or acquisition shift. These considerations entail possible ways to improve generalization during training and testing: for instance, data resampling is a possible strategy to mitigate prevalence and population biases [19].

The classification task in principle modeled as a causal task since, in the absence of an independent reference standard, the images were manually labelled. The assumption that Y is caused by X, and not vice versa, stems from the fact that labels are generated through manual annotation without explicit knowledge of the hidden variable W and, thus, any substantial modification of the image may change the value of

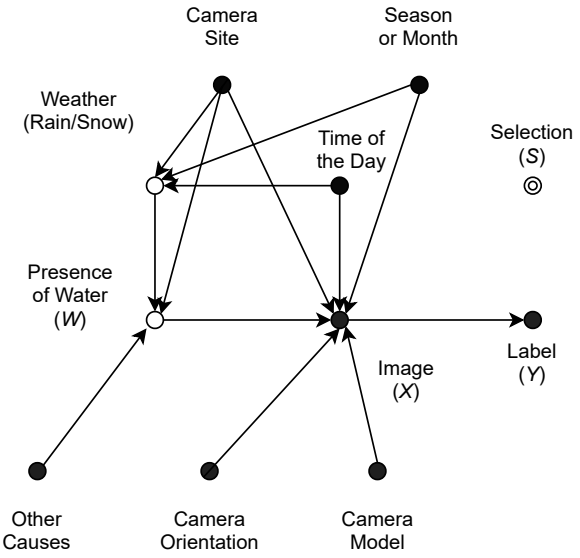

Fig. 4: Causal model for the prediction of wet road conditions. Hidden variables are depicted with empty circles. The selection scheme is random sampling.

the assigned label. To further illustrate the difference, consider the case where the presence of water is not visible due to occlusions or poor visibility: in such a case, any rater would give a negative label, regardless of the actual road condition.

However, as in the medical use case, the distinction between causal or anticausal tasks is not entirely clear-cut. What if the image-derived labels determined by an expert are nearly identical to the hidden variable? Could then the labels serve as proxies for the ground truth, possibly configuring an anticausal relationship? It is also worth noting that the annotators were aware of the site, date and time of acquisition; the weather, even if not explicitly recorded in the dataset, could be inferred from the input image. Hence, from this viewpoint, the task can be seen as confounded, and hence anti-causal. This perspective is supported by our empirical observation that SSL was effective to improve performance. More investigations are needed to clarify the role of SSL in ML tasks for which the causal or anticausal nature is not easily determined.

The domain $\mathcal{D}$ is determined by the following variables: site, season, camera orientation and camera model. Camera orientation and camera model have a direct influence on the image and hence are contributors to the so-called domain or acquisition shift. From a causal perspective, the image X is a collider of many variables, including W and camera parameters previously mentioned, thus care must be taken to prevent collider biases affecting the training process. The distribution of all known factors was studied and corrected, if needed, by sampling and data augmentation. For instance, when data is acquired continuously during the day and for the whole year, the distribution may be balanced with respect to time of day and season, but it is unlikely to account for all possible sites, which we identified as the most critical factor for generalization. Camera orientation has a profound impact on the presence of reflections, myrages, as well as under- and over-exposure. Data augmentation strategies that simulate different camera orientations and illumination conditions, thus artificially balancing the dataset, are useful in this case to prevent the appearance of spurious correlations as a result of imperfect data selection or small dataset sizes.

Finally, a predictive model may be controlled by confounders (e.g., the season) by adding them as additional inputs, thus statistically conditioning the prediction on their value; however, this is possible only if the confounder is observed (measured) at training and inference time, and if the training set covers the entire support of the confounder distribution.

V. Related Works and Discussion

Several recent works have focused on the data collection and annotation phases, in an effort to increase its reliability and reproducibility, as well as promote awareness of the risks and perils of dataset biases.

Dataset Datasheets [7] and Data Nutrition Labels [15] are two high-profile initiatives to *standardize the way datasets are collected and reported*. A Dataset Datasheet includes discursive descriptions like the motivation behind the creation of the dataset, its composition and other methodological information such as any applied preprocessing and the recommended usage. The Data Nutrition Labels provide similar information but in a more concise format inspired by food nutritional labelling. Both initiatives enhance the transparency of the data collection process, and have had a mitigating effect on undesired or undetected dataset biases. Causal diagrams, on the other hand, represent a formal tool that can be used during, and after, the creation of a dataset to reason about the relationships between different variables.

A few authors worked on *quantifying biases in ML datasets* by employing statistical techniques [6], [20], [26]. For instance, Beretta et al. studied the intrinsic discriminatory risk by assessing the degree of dependency between a protected attribute (e.g., race or gender) and the target variable [26]. To the best of our knowledge, the majority of these techniques work on structured datasets, and are not directly applicable to typical DL models for unstructured data, e.g., images or text.

Causal and counterfactual reasoning are also increasingly used to quantify and/or mitigate biases in trained ML models [9], [27], [28]. However, most existing causality-based algorithms require knowledge of the underlying causal graph. The examples presented in this paper show how causal diagrams can connect these two lines of research by linking unstructured data (e.g., images) to latent factors, whose distribution can be modelled and analyzed by either statistical methods or counterfactual reasoning.

An open question is how to *evaluate the validity* of the causal model when used in this fashion: first, the correctness and completeness of the causal model cannot be checked against the ground truth, as the latter is unknown. Recent work has tackled the problem of verifying causal models learned from observational data, but the method is not readily applicable to unstructured data [29]. Second, the benefit of

integrating causal models into dataset construction, while strongly supported by intuition, has not been experimentally linked to an increase in performance and generalization of the trained ML models.

The *nature, type and presence of biases* in CV datasets has also been extensively studied, starting from the seminal work by Torralba and colleagues, most commonly by studying the cross-dataset generalization performance of trained DNNs [5], [8], [13], [30]. Torralba et al. suggested general strategies to avoid common biases such as selection bias, capture bias, and negative set bias [5]. Biases due to gender and ethnicity have been addressed by collecting larger, more diverse datasets [28]. These strategies are complementary to causal diagrams, which however are more easily applied to specialized datasets in which the domain is well defined (e.g., industry), and the data generation process can be precisely and accurately modelled.

VI. Conclusions

Incorporating causality is a fundamental challenge in ML research. In this work, we investigated causal analysis as a highly effective technique to characterize the properties of ML datasets. Previous works have proven its effectiveness in the medical domain [19], [24], we here extended the methodology to a real-life example from an industrial research project to prove its feasibility and potential benefits. In future works, we plan to perform a more in-depth evaluation of the impact of modelling the dataset based on causal diagrams on the performance and generalization ability of the trained models. Nonetheless, we hope with this work to encourage practitioners adopting this systematic approach to the analysis of data collections in other domains, as this approach may help to manage dataset biases and concept drift in the training and deployment of neural networks.

Acknowledgements

The authors gratefully acknowledge financial support from Waterview srl.

References

[1] K. Muhammad, A. Ullah, J. Lloret, J. Del Ser, and V. H. C. de Albuquerque, "Deep learning for safe autonomous driving: Current challenges and future directions," *IEEE Transactions on Intelligent Transportation Systems*, 2020.

[2] A. Ziletti, D. Kumar, M. Scheffler, and L. M. Ghiringhelli, "Insightful classification of crystal structures using deep learning," *Nature communications*, vol. 9, no. 1, pp. 1–10, 2018.

[3] L. Morra, S. Delsanto, and L. Correale, *Artificial intelligence in medical imaging: From theory to clinical practice*. CRC Press, 2019.

[4] R. Sagar. Big data to good data: Andrew Ng urges ML community to be more data-centric and less model-centric. Available: https://tinyurl.com/3t6dp5n3. [Online]. Available: https://tinyurl.com/3t6dp5n3

[5] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *CVPR 2011*. IEEE, 2011, pp. 1521–1528.

[6] M. Mecati, F. E. Cannavò, A. Vetrò, and M. Torchiano, "Identifying risks in datasets for automated decision–making," in *International Conference on Electronic Government*. Springer, 2020, pp. 332–344.

[7] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford, "Datasheets for datasets," *arXiv preprint arXiv:1803.09010*, 2018.

[8] T. Tommasi, N. Patricia, B. Caputo, and T. Tuytelaars, "A deeper look at dataset bias," in *Domain adaptation in computer vision applications*. Springer, 2017, pp. 37–55.

[9] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *arXiv preprint arXiv:1908.09635*, 2019.

[10] T. Calders, "Machine-learning discrimination: bias in, bias out," in *2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS)*, 2019, pp. 27–27.

[11] N. Petrick, B. Sahiner, S. G. Armato III *et al.*, "Evaluation of computer-aided detection and diagnosis systems," *Medical physics*, vol. 40, no. 8, p. 087001, 2013.

[12] M. Borg, C. Englund, K. Wnuk *et al.*, "Safely entering the deep: A review of verification and validation for machine learning and a challenge elicitation in the automotive industry," *arXiv preprint arXiv:1812.05389*, 2018.

[13] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," *Nature Machine Intelligence*, vol. 2, no. 11, p. 665–673, Nov 2020. [Online]. Available: http://dx.doi.org/10.1038/s42256-020-00257-z

[14] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang, "Learning under concept drift: A review," *IEEE Transactions on Knowledge and Data Engineering*, p. 1–1, 2018. [Online]. Available: http://dx.doi.org/10.1109/TKDE.2018.2876857

[15] S. Holland, A. Hosny, S. Newman, J. Joseph, and K. Chmielinski, "The dataset nutrition label: A framework to drive higher data quality standards," *arXiv preprint arXiv:1805.03677*, 2018.

[16] J. M. Rohrer, "Thinking clearly about correlations and causation: Graphical causal models for observational data," *Advances in Methods and Practices in Psychological Science*, vol. 1, no. 1, pp. 27–42, 2018.

[17] J. Pearl and E. Bareinboim, "External validity: From do-calculus to transportability across populations," *Statistical Science*, pp. 579–595, 2014.

[18] B. Schoelkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij, "On causal and anticausal learning," 2012.

[19] D. C. Castro, I. Walker, and B. Glocker, "Causality matters in medical imaging," *Nature Communications*, vol. 11, no. 1, pp. 1–10, 2020.

[20] L. Zhang, Y. Wu, and X. Wu, "Achieving non-discrimination in data release," in *Proc. of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 1335–1344.

[21] J. Peters, D. Janzing, and B. Schölkopf, *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[22] L. Darlow, S. Jastrzebski, and A. Storkey, "Latent adversarial debiasing: Mitigating collider bias in deep neural networks," *arXiv preprint arXiv:2011.11486*, 2020.

[23] M. Etminan, G. S. Collins, and M. A. Mansournia, "Using causal diagrams to improve the design and interpretation of medical research," *Chest*, vol. 158, no. 1, pp. S21–S28, 2020.

[24] G. J. Griffith, T. T. Morris, M. J. Tudball, A. Herbert, G. Mancano, L. Pike *et al.*, "Collider bias undermines our understanding of covid-19 disease risk and severity," *Nature communications*, vol. 11, no. 1, pp. 1–12, 2020.

[25] Z. Xu, J. Liu, D. Cheng, J. Li, L. Liu, and K. Wang, "Assessing the fairness of classifiers with collider bias," *arXiv preprint arXiv:2010.03933*, 2020.

[26] E. Beretta, A. Vetrò, B. Lepri, and J. C. D. Martin, "Detecting discriminatory risk through data annotation based on bayesian inferences," in *Proc. of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 794–804.

[27] S. Verma, J. Dickerson, and K. Hines, "Counterfactual explanations for machine learning: A review," *arXiv preprint arXiv:2010.10596*, 2020.

[28] V. Sharmanska, L. A. Hendricks, T. Darrell, and N. Quadrianto, "Contrastive examples for addressing the tyranny of the majority," *arXiv preprint arXiv:2004.06524*, 2020.

[29] B. Butcher, V. S. Huang, C. Robinson, J. Reffin, S. K. Sgaier, G. Charles, and N. Quadrianto, "Causal datasheet for datasets: An evaluation guide for real-world data analysis and data collection design using bayesian networks," *Frontiers in Artificial Intelligence*, vol. 4, p. 18, 2021.

[30] A. Kortylewski, B. Egger, A. Schneider, T. Gerig, A. Morel-Forster, and T. Vetter, "Analyzing and reducing the damage of dataset bias to face recognition with synthetic data," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019.