

Data Quality Challenges in Artificial Intelligence: Impacts of Missing, Mislabeled and Biased Data on Machine Learning

Zhiyuan Niu

*College of Arts and Sciences, New York University, New York, USA
zackniu00@gmail.com*

Abstract. Artificial Intelligence (AI) is developing at a high speed and is now used in many places. AI models need a large amount of data to some extent; therefore, problems in the data can lead to errors in the model. The three main types of harmful data that are the subject of this study are biased, mislabelled and missing data. If a large part of the data set is missing, it will result in information loss and a lower-performing model. In addition, the model may be incorrectly trained due to mislabelled data and thus unreliable. If the data are biased, then there will be discrimination against some groups. Some possible countermeasures have also been mentioned in the discussion. Imputation or Deletion Methods are used to address missing values. Data cleaning, label verification and noise-robust learning can all reduce mislabelled data. Bias in the data can be addressed before, during and after model training. Generally speaking, there is no single way to handle all the data problems. Different datasets and problems have different solutions, so they will be different.

Keywords: Data Quality, Machine Learning, Artificial Intelligence

1. Introduction

Artificial intelligence (AI) is a system that can make computers and other machines exhibit some human-like traits, such as learning, reasoning, decision-making and behaviour. It is a popular topic in contemporary society, and this new technology is developing rapidly and is being applied in the lives. For example, ChatGPT, Claude, DeepSeek and Doubao are widely used by people in their daily lives and various fields such as biology, medicine, sports, etc. People's expectations of AI are rising, and many now believe that AI will be the main force driving change in human society in the future. At the same time, Big Data serves as the foundation for artificial intelligence (AI), and many people are now using AI [1]. Big Data refers to a large amount of complex data that is collected at a high speed from many sources in different places, and through organisation, management and analysis of this data, new ideas and improved business strategies can be obtained. Before AI was widely used, big data was applied to social media to help people find more videos of interest by means of related searches. Big data is used to build large language models and improve the user experience of AI through self-learning from big data [2].

Although AI and Big Data have been widely talked about and used in many places, some research deficiencies still exist. A large number of existing studies have pointed out the possible advantages

of AI, such as increased efficiency, decision support and personalised services; for example, improving efficiency, supporting decision-making and providing personalised services. However, the performance of AI is also directly related to the quality, accuracy and diversity of the training data [1]. If the data are missing, biased, old or incorrectly collected, the AI may produce inaccurate or unfair results [1, 3, 4]. There are also privacy and data security risks in the collection and use of a large amount of personal information [2]. Although AI and big data are closely related, most of the research on them has focused on the technical or positive aspects. Therefore, more research is needed to explore how big data can support the development of AI, as well as the problems of data quality, bias, privacy and security that may affect the reliability and responsible use of AI in people's daily lives and business activities [5].

Toxic data in this paper refers to data that are missing, incorrectly labelled, biased, duplicated, outdated, underrepresented, or intentionally poisoned. The above problems have been addressed in previous studies. Missing values, inaccurate features and incorrect labels will reduce the performance of a machine-learning model [1]. Label errors can also make the model test unreliable and lead to the selection of an incorrect model [6]. Biased and unrepresentative datasets will have different levels of accuracy for different groups of people [3]. Duplicate data may cause the model to overfit the training data and create an overlap between the training and test sets [2]. When the old data no longer reflects the present, the model's performance will decline over time [4]. Another problem is that, intentionally or unintentionally, data poisoning can be carried out to harm the model [7].

These data problems are generally studied in isolation. Some papers address missing data or incorrect labels; others mainly discuss bias, duplication, old data and data poisoning. Therefore, it is difficult to know how these factors affect the construction of AI models as a whole, and the solutions proposed by different scholars cannot be compared. Collectively, the above studies have investigated how all kinds of toxic data affect the performance, fairness, reliability, robustness and evaluation of AI models. It also considers the above measures: data cleaning, label correction, dataset balancing, deduplication, model update and defence against data poisoning. The following sections will first introduce the main kinds of toxic data, then discuss the effect of these on the development of AI models, compare the solutions proposed in previous research, and finally present the limitations and future directions of research.

2. Missing data

Missing data refers to samples or indices that should have been provided for training, testing or generating an AI model but could not be obtained or recorded. The three types of missing data mechanisms are: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) [8]. MCAR is the case where the observed and unobserved measurements are independent of the missing data. MAR is only the observable data that are related to the probability of a missing value in MAR. This is to say, the missing data are neither MCAR nor MAR. The observed and missing values have the same weight in the missing data. Since this way does not use the original data, handling missing values is not feasible.

Missing values generally occur in all the data fields, and they can cause all sorts of problems, such as a drop in performance, difficulties with data processing, and biased results due to differences between missing and complete information. In addition, the quantity of missing data, the pattern of missing data and the reason for the missing data all affect the extent of the missing-value problem [8]. Therefore, before building a model, the missing values in the data need to be addressed; otherwise, there may be a bias or inaccuracy in the analysis caused by these missing values.

Therefore, the choice of approach for handling missing data must also be reasonable to avoid errors in the results [8]. If the missing values are not random, bias will occur due to the exclusion of incomplete data. In addition, after excluding many incomplete cases, listwise deletion may lead to the loss of important information [8].

Based on the results shown in the classification, regression and clustering tasks, the missing value performance loss across all of them exceeded 25 per cent, with some reaching as high as 40 per cent. Therefore, when there was a large amount of missing data, completeness was considered to have a high impact [1]. However, the results also show that the model's performance is not significantly affected by having less than 40% missing values during the training stage [1]. Changes in completeness and feature correctness had the most significant impact on the density-based OPTICS algorithm for clustering, but the autoencoder technique was less affected [1]. In any case, the impact of missing data varies by situation and pattern, and all datasets have some such factors.

3. Mislabeled data

Then, Mislabelled Data. It means that some examples in the dataset have been assigned to the wrong category and do not match their actual labels [9]. Several possible reasons may lead to mislabelled data. 1. The information collected by the annotators is lacking. First, if the descriptive language is too limited or the data quality is low, there may not be sufficient information available for accurate labelling [9, 2]. An error occurs in the labelling stage. Second, even professionals frequently make mistakes when labeling [9]. In addition, the performance of a machine-learning model is directly related to how much high-quality, labelled data is available; otherwise, it cannot be used. To reduce the high cost of labelling, some non-expert sources, such as Amazon Mechanical Turk and the surrounding text of collected data, have been used regularly; however, the use of these sources often results in unreliable labels. Therefore, although the labels provided by non-experts are inexpensive and easy to obtain, they are generally less reliable. Labeling is inherently subjective to some extent. Third, there is no norm; therefore, the experts' judgments may differ. For example, the pattern boundaries provided by the two experts for the segmentation of ECG signals are often slightly different [9, 4]. Encoding or Transmission may be faulty at this time. Label noise can also be caused by communication or data-encoding errors. Real-world databases are also expected to have about 5% encoding errors [9].

Schafer and Graham have proposed a statistical taxonomy of label noise, and there are three kinds of it [9]. Noisy Completely at Random (NCAR): A labelling error occurs independently of the sample's features and the true class [9]. Noisy at Random (NAR): The probability of a labelling error is independent of the sample attributes but depends on the actual class [9]. 3. Noisy Not at Random (NNAR): Some classes and feature regions are more likely to be mislabelled because the probability of a labelling error depends on both the true class and the sample features [9].

Label noise is relatively common in real-world datasets and has several negative effects [9]. When the target accuracy of the training data is reduced, the model's performance drops below that of a majority classifier if the target accuracy is less than one divided by the number of classes [1]. However, the results show that classification models can be trained with noisy data and have little loss in performance if at least 80% of the data are correctly classified [1]. In addition, the results of this study show that the test data need to be correctly labelled; otherwise, the performance of the model will be underestimated to some extent [1]. Finally, from a theoretical standpoint, learning algorithms are rarely entirely robust to label noise except in a few simple cases. However, some algorithms are more robust in practice than others [9].

4. Biased data and its effect on AI

Biased data represent another important source of harm in machine learning systems. The term "bias" can refer to a wide range of problems, but in this context it mainly describes data-related factors that may lead to unfair or systematically distorted model outcomes. These issues can be understood more clearly by separating broad concepts such as "algorithmic bias" or "data bias" into more specific and practical sources of harm throughout the machine learning process [10]. The majority of AI algorithms and systems are data-driven and need training data. As a result, the operation of these algorithms and systems is closely related to data. If there are biases in the training data, then the algorithm that is trained on this data will learn to recognise and incorporate those biases in its predictions. Therefore, the original bias in the data may be carried over by the algorithm and result in a biased outcome. Even existing biases in the data can be amplified and perpetuated by algorithms [11].

There are many types of bias, and some of them may lead to unfairness in all the following learning tasks [11]. Even after careful measurement and sampling of the data, historical bias can still occur if the world as it is or was has resulted in a model with negative consequences. A population can be harmed by such a system even if it accurately reflects the world. One type of harm to be avoided in the evaluation of representation is reinforcing stereotypes about a particular group [10]. Additionally, when selecting, gathering, or calculating features and labels for a prediction problem, measurement bias takes place. A feature or label is generally a proxy (an observable indicator) used to represent a construct (an abstract concept) that cannot be directly observed or directly recorded [10].

Aggregation bias also occurs when a single method is used to process data with different groups or examples that need to be treated differently. The assumption that the mapping from inputs to labels is constant over subsets of the data is the source of aggregation bias. This is often not the case in reality. People or groups with different histories, cultures, or conventions may be included in a particular dataset, and a given variable can have very different meanings for these groups. A model that is well-suited to the dominant population (e.g., if there is also representation bias) or that is not ideal for any group can result from aggregation bias [10]. If the benchmark data for a particular task do not cover the actual user base, then there will be an evaluation bias. Among all kinds of bias, this problem is more serious: a misleading benchmark encourages the creation and application of models that perform well only on the part of the data covered by the benchmark data [10]. The indicators of performance in the report may also be biased. For instance, these individual indicators are frequently used to compare different models and determine which one is "better", even if the overall results do not show a significant difference in subgroups. Disparities in other kinds of errors, such as the false negative rate, may also be hidden if only one type of indicator, such as accuracy, is used [10].

The research found that lighter-skinned people accounted for the majority of these datasets (79.6% in IJB-A and 86.2% in Adience), and a new facial analysis dataset was created that is balanced in terms of skin colour and gender. Based on this dataset, the three commercial gender category algorithms were tested and found that, at an error rate of 34.7%, darker-skinned females are the most frequently misclassified group. The upper bound of the error rate for light-skinned boys is 0.8% [3]. Similarly, the experiment shows that a popular algorithm used by millions of patients and typical of this industry-wide strategy has significant racial bias: signs of serious illness are much more pronounced in Black patients than in White patients at the same risk score. Reducing this difference would increase the proportion of Black patients receiving extra assistance from 17.7% to 46.5%. Due to the uneven distribution of medical facilities, fewer Black patients have been seen by

doctors and thus treated. This bias is due to the algorithm's prediction of healthcare expenses rather than illness. Therefore, serious racial bias still exists even if healthcare costs are considered an effective proxy for health by some metrics of prediction accuracy [12]. However, not all the problems should be attributed to the data. Many choices and steps in the ML pipeline occur from the model definition to the user interface in deployment. Decisions at all levels may have adverse effects [10]. Although the above categories are not mutually exclusive, they are more convenient to understand and manage if they can be identified and described separately. There is no single standard for fairness; a particular application and the perspectives of the people concerned need to be considered [10].

5. Solutions

As all sorts of toxic data are harmful to AI models, it is required to find ways to reduce or handle such problems. No single way can address all kinds of toxic data because the sources and effects of different data problems vary. Therefore, the answers provided in the earlier studies will be reviewed in this section, starting with methods for handling missing data.

Solutions to Missing Data. Some methods are to delete the corresponding instances and replace them with possible or estimated values to address this problem by imputation [8]. A number of traditional statistical and machine learning imputation methods have been proposed to address missing values in the research, such as mean, regression, K-nearest neighbour, ensemble-based methods, etc. [8]. In terms of how missing data are handled, some scholars have chosen to omit or exclude the relevant data, while others have used multiple imputation. Most of the problems with single imputation have been addressed by multiple imputation; that is, several values that reflect the uncertainty in the true value are estimated using the distribution of the observed data [8]. Although several methods for handling missing data have been employed in recent times via multiple imputation, a suitable imputation method needs to be selected for this experiment to ensure the reliability of the results [8].

The KNN method for machine-learning-based imputation uses a distance function between samples to find the nearest neighbours of missing values and then uses these neighbours to perform imputation [8]. Random Forest missForest is a non-parametric imputation method based on Random Forest, which was used next. MissForest predicts the missing values of each variable by building a random forest model on the observed data [8]. Random Forest, discriminating DL and k-NN are good choices in most experimental circumstances, and according to the experimental comparison [13], they have the best imputation performance. Random Forest is a relatively good Imputation method. It performs well in most of the experimental conditions and also has a good training balance that is independent of the size of the training set, as well as optimised hyperparameters and inference time [13].

In addition, the results show that imputation can enhance the predictive power of the following machine-learning model. Over 75% of the experiments showed improvements of 10–20% in the downstream task [13]. However, the advantage of imputing missing values in the test data was reduced to a certain extent, and in some cases, it was even negative after the imputation technique was trained on incomplete data [13]. Finally, none of the methods were universal solutions, and the results of several approaches differed according to the dataset [8].

Solutions to Mislabeled data. Label noise, or incorrectly labelled data, is the second problem after missing values. Three different approaches to mislabelled data are available in the literature: label noise-tolerant learning algorithms, data cleansing techniques, and label noise-robust models [9]. First, in the presence of label noise, robust methods often avoid overfitting [9]. To reduce label

noise, one can simply eliminate cases that appear to be incorrectly labelled. Cases can also be filtered based on model predictions; for instance, one might exclude training cases that a classifier has misidentified, but this may result in losing too many instances [9]. In addition, when all or almost all of the learners in an ensemble agree to delete an instance, voting filtering removes it [9].

Crowdsourced label curation employs many human workers to verify or correct errors in the labels of a test set in a dataset; however, it is very expensive for large-scale datasets [6]. First, determine which samples are more likely to be mislabelled, and then only validate subsets of the datasets to address this problem [6]. Confident Learning is a model-agnostic method that determines whether the labels are incorrect in all datasets for robustness to class imbalance and theoretical support for precise uncertainty quantification [6]. Confident Learning automatically detects potential label problems and can thus identify label errors in the test set without manual examination of the entire test set [6]. Crowdsourcing can be employed to verify the identified label inaccuracies by people [6]. Confident learning does, however, occasionally identify data items that are not actually incorrect [6].

More recently, there are several ways to improve deep learning's resilience [14]. The above methods include robust regularisation to reduce the overfitting of a DNN to false-labelled data, either explicitly or implicitly; robust architecture, which adds a noise-adaptation layer at the top of an underlying DNN to learn the label-transition process or develops a dedicated architecture to reliably support various types of label noise; robust loss function to improve the loss function; loss adjustment, which modifies the loss value based on the confidence of a particular loss or label through loss correction, loss reweighting, or label refurbishment; and sample selection to use multi-network or multi-round learning to identify true-labelled samples from noisy training data [14]. The above methods will also enhance the noise immunity of the system. However, an increase in computing cost is unavoidable, and the hyperparameters provided by these methods make the DNN more susceptible to variations in data and noise types [14]. Finally, the comparison results showed that none of the methods could meet all the required attributes; instead, the supported qualities varied according to the category of the method [14].

Solutions to Biased Data. The third problem to be addressed is bias in the data, not missing or miscategorised data. Generally speaking, the three kinds of approaches to algorithmic bias are as follows [11]. Preprocessing methods first attempt to alter the data in order to reduce the original bias. Preprocessing is possible if the algorithm is allowed to modify the training set [11]. Second, in-processing methods aim to avoid discrimination in model training by modifying or altering cutting-edge learning algorithms. In-processing can be used during model training if it is allowed to change the learning process of a machine learning model by adding a constraint or modifying the objective function [11]. Finally, after training, post-processing is performed on a holdout set that was not used for training to gain access to it. The labels initially assigned by the black-box model are reallocated based on a function in the post-processing stage if the learned model can only be used as a black box without altering the training data or learning method [11].

Every dataset is the result of several design choices made by the data curator at the data level. The above selections will affect the fairness of the final dataset and thus impact the algorithms that will be developed [11]. Some general strategies to support good practices in data use have been proposed to reduce the impact of bias in data [11]. One such way is to use the datasheet as a reference for the data and to report on the attributes, reasons for collection and distribution of the dataset. In addition, to improve the accountability and transparency of AI, inclusive benchmark datasets and subgroup accuracy reports will be released [3]. Improvements in the architecture of

neural networks, feature selection and dataset construction can all be achieved by means of intersectional phenotypic and demographic error analysis [3].

If an invalid proxy is used to obtain biased predictions, the training labels may also be biased. Instead, the algorithm's input data—more especially, the labels applied to it—must be modified [12]. The sample, predictors, and training procedure were all part of the current model architecture, but the label was altered as a first step. An index variable combining cost and health prediction was developed instead of a future cost. Given the risk score, this strategy reduced the number of excess active chronic illnesses among Black patients to 7,758 and thus lowered the bias by 84% [12]. Therefore, the bias in the labels can be corrected [12]. However, unless there are very limited circumstances, it is difficult to meet all the fairness requirements simultaneously [11]. Moreover, because the reasons for harm differ from place to place, there is no single set of suitable countermeasures that can be applied everywhere [10]. There is no single norm of fairness; therefore, in a particular case, one must consider all the parties concerned to determine what is fair [10].

6. Conclusion

The three main types of harmful data in this paper are biased, mislabelled and missing data. The performance, reliability and fairness of the AI model may all be affected by these problems, although they originate from different places. Depending on the amount and pattern of missing values, there will be a loss of information or incorrect results. In addition, the evaluation results will be unreliable because of mislabelled data in the training set. If the data are biased, different groups will receive different results and this will further discriminate against certain people.

All the above will help to solve these problems. Statistical and machine-learning-based imputation methods can be used to handle missing data by substitution or deletion. Data cleaning, label checking, confident learning, and models that are more robust to label noise can all be employed to handle mislabeled data. Modify the data before training, alter the learning process, or correct the model's output after training to address bias in the data. To reduce bias in the results, more representative datasets can be used, labels improved, and separate performance reports for different groups provided.

But all of them have some defects. Imputation may be inaccurate; deletion will result in a loss of data. Robust learning techniques may require more computing power, and human label checking is relatively expensive. The place where the problem occurs and the people it affects will also impact the plans to reduce bias. Therefore, there is no single way that can be applied to all datasets and AI systems. Pay attention to data quality at all times in the construction of the AI model. Select a method according to the specific data problem to improve the accuracy, reliability and fairness of the AI system.

References

- [1] Mohammed, S., Budach, L., Feuerpfeil, M., Ihde, N., Nathansen, A., Noack, N., Patzlaff, H., Naumann, F., & Harmouch, H. (2025). The effects of data quality on machine learning performance on tabular data. *Information Systems*, 132, 102549. <https://doi.org/10.1016/j.is.2025.102549>
- [2] Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., & Carlini, N. (2022). Deduplicating training data makes language models better. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 8424–8445. <https://doi.org/10.18653/v1/2022.acl-long.577>
- [3] Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 81, 77–91.
- [4] Guo, L. L., Pfohl, S. R., Fries, J., Posada, J., Fleming, S. L., Aftandilian, C., Shah, N. H., & Sung, L. (2021). Systematic review of approaches to preserve machine learning performance in the presence of temporal dataset

- shift in clinical medicine. *Applied Clinical Informatics*, 12, 808–815. <https://doi.org/10.1055/s-0041-1735184>
- [5] Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P. K., & Aroyo, L. M. (2021). "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Article 39, 1–15. <https://doi.org/10.1145/3411764.3445518>
- [6] Northcutt, C. G., Athalye, A., & Mueller, J. (2021). Pervasive label errors in test sets destabilize machine learning benchmarks. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 1.
- [7] Steinhardt, J., Koh, P. W. W., & Liang, P. (2017). Certified defenses for data poisoning attacks. *Advances in Neural Information Processing Systems*, 30, 3517–3529.
- [8] Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. *Journal of Big Data*, 8, 140. <https://doi.org/10.1186/s40537-021-00516-9>
- [9] Frénay, B., & Kabán, A. (2014). A comprehensive introduction to label noise. *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2014)*, 667–676.
- [10] Suresh, H., & Gutttag, J. V. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. *Proceedings of the 2021 ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, Article 17, 1–9. <https://doi.org/10.1145/3465416.3483305>
- [11] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54, Article 115, 1–35. <https://doi.org/10.1145/3457607>
- [12] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366, 447–453. <https://doi.org/10.1126/science.aax2342>
- [13] Jäger, S., Allhorn, A., & Bießmann, F. (2021). A benchmark for data imputation methods. *Frontiers in Big Data*, 4, 693674. <https://doi.org/10.3389/fdata.2021.693674>
- [14] Song, H., Kim, M., Park, D., Shin, Y., & Lee, J.-G. (2023). Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 34, 8135–8153. <https://doi.org/10.1109/TNNLS.2022.3152527>