IBIMA Publishing
Communications of International Proceedings
https://ibimapublishing.com/p-articles/46AI/2025/4625025/
Vol. 2025 (34), Article ID 4625025

Analysis of Factors Contributing to Limited and Imbalanced Datasets in Machine Learning*

Maciej KACZYŃSKI and Zbigniew PIOTROWSKI

Military University of Technology, Warsaw, Poland

Correspondence should be addressed to: Maciej KACZYŃSKI, maciej.kaczynski@wat.edu.pl

* Presented at the 46th IBIMA International Conference, 26-27 November 2025, Ronda, Spain

Abstract

The development of reliable machine learning (ML) models fundamentally depends on the availability of large, diverse, and balanced datasets. However, in practice, limited and imbalanced data pose common challenges in both scientific research and business applications. The aim of this paper is to analyse the fundamental financial, regulatory, ethical, and technical constraints that contribute to dataset limitations, with a particular focus on their impact on model robustness and generalizability. High acquisition costs, intellectual property restrictions, and inadequate labelling practices limit data availability, while regulatory frameworks impose strict constraints on data usage and cross-border transfer. Technical challenges, including insufficient computational resources, label noise, and integration difficulties, further expand the problem of small datasets. For business applications in fields such as finance, healthcare, and manufacturing, these constraints not only hinder predictive accuracy but also impact decision-making efficiency. Understanding these factors is of key importance for developing strategies that reduce dataset limitations, ensure the preparation of correct and sufficient datasets, and support efficient ML.

Keywords: Datasets, imbalanced data, small data, machine learning.

Introduction

The origins of data shortage are diverse and often interdependent, encompassing financial, ethical, regulatory, and technical factors. These constraints typically lead to incomplete or non-representative training samples, thereby undermining the robustness and generalizability of models. Knowledge of the causes of limited datasets supports prevent or reduce such problems, given the inherent challenges in obtaining proper training data. This paper discusses the primary sources of limited datasets, demonstrating how each factor constrains data availability and hinders advancements in research and practical applications. Figure 1 provides a systematic categorization of the key factors contributing to limited and imbalanced datasets. Each factor is discussed in detail in the subsequent sections of this paper.

Cite this Article as: Maciej KACZYŃSKI and Zbigniew PIOTROWSKI, Vol. 2025 (34) "Analysis of Factors Contributing to Limited and Imbalanced Datasets in Machine Learning " Communications of International Proceedings, Vol. 2025 (34), Article ID 4625025, https://doi.org/10.5171/2025.4625025

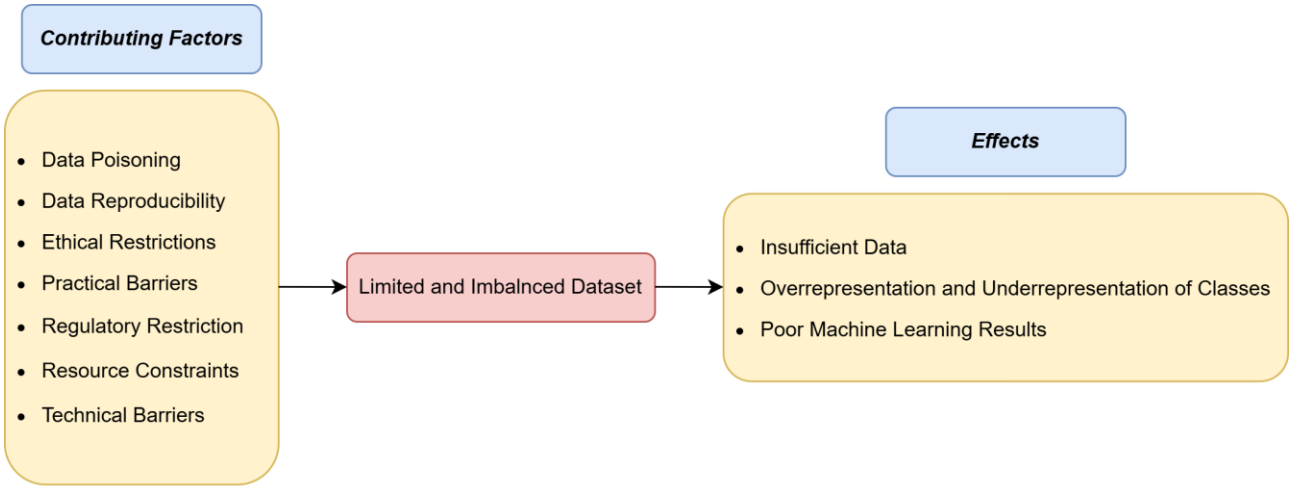

Fig. 1. Systematic categorization of key factors contributing to limited and imbalanced datasets.

Resource Constraints

Data may be rare, proprietary, or extremely expensive to acquire. Additionally, resource constraints, such as budget constraints for computing infrastructure, regulatory restrictions, and stringent compliance requirements, can make it difficult to collect large, high-quality datasets. Data collection can require significant financial and human resources. Limited funding or staffing can limit the scope and depth of data collection efforts. Limited access to data sources increases the challenges of having a small or imbalanced dataset in financial or other heavily regulated domains.

Even if an organization knows which data would be most relevant to a particular problem (e.g., a stock market forecast), it may not be able to obtain it due to a variety of factors. Prohibitive subscription costs to proprietary data feeds (Subrahmanyam, 2019) or data usage restrictions imposed by regulators and data providers can impact the dataset. Incomplete or delayed access to real-time data streams can also impact the collection and value of data. Additionally, internal interest groups within large organizations may store valuable data in isolation. These limitations can reduce the size and richness of the dataset, making it difficult to develop robust neural networks (NN) models. Potentially valuable data may exist but remain inaccessible due to ownership, organizational policies, or competition, limiting the amount of open or shared data.

Regulatory and Ethical Restrictions

Legal and regulatory restrictions often shape how data can be collected, stored, and used in machine learning (ML) applications. For organizations operating in tightly regulated sectors (e.g., finance, healthcare), these rules can severely limit the volume and variety of data available to train NNs. Even when it is technically possible to gather robust datasets, regulations may prohibit or constrain data sharing and cross-border transfers, leading to smaller, fragmented datasets that can hinder model performance. Examples of key regulations and their impact on data availability:

- Cybersecurity Law (CSL) (DigiChina, 2017);
- General Data Protection Regulation (GDPR) (Regulation - 2016/679 - EN - gdpr - EUR-Lex, 2016);
- Health Insurance Portability and Accountability Act (HIPAA) (CDC, 2024).

CSL of the People's Republic of China is a comprehensive national regulation governing network security and data protection in China. The law establishes strict requirements on data collection, storage, and cross-border transfer, mandating that vital information infrastructure operators store personal and important data within China. It also enforces obligations for real-name registration, security assessments, and government access to data. These restrictions significantly influence the availability and usability of datasets, limiting the inclusion of certain demographic, transactional, or behavioural variables in ML applications.

GDPR is a major data protection law recognized in the European Union (EU). Strict rules on personal data processing, transfer, and storage are imposed under GDPR. The use of personally identifiable information is heavily restricted by this law, thereby excluding certain demographic or behavioural variables from ML models.

HIPAA is a United States (U.S.) law designed to protect the privacy of patient health information. Although HIPAA is healthcare-focused, its data protection principles are often extended to financial or insurance applications involving medical records alongside financial records. Access to detailed patient information which could benefit risk assessment is reduced under this law, leading to smaller and less informative datasets for model training.

Technical and Practical Barriers

Even when the data itself is available and permitted for use, the training of NNs on small datasets is frequently complicated by practical and technical limitations. For instance, a shortage of high-quality labelled data, suboptimal computational resources, and challenges in model deployment or maintenance can cause the derailment of ML initiatives. Such barriers are further amplified by domain-specific complexities (e.g., in manufacturing, finance, healthcare).

NNs are often subject to high computational costs during training, particularly when utilizing advanced architectures such as transformer-based models or deep recurrent networks. When data scarcity is present, the addition of complex structures (e.g., attention layers, large embedding dimensions) may not provide significant benefits unless robust hardware is available to properly optimize the training process. Prolonged training times, risk of suboptimal hyperparameter tuning, and difficulty experimenting with complex architectures. Insufficient or incorrect data preprocessing can lead to model overfitting, poor generalization, and misleading performance metrics, especially in small-data regimes. The risk of NNs memorizing training samples (overfitting) is increased when there are too many parameters relative to the available data. This issue is particularly problematic for small datasets, where the ratio of parameters to data points becomes disproportionately high. The model performs well on the training dataset but poorly in production, leading to wasted time and resources. When datasets are small, they may be impacted by label noise or contain partial labels, thereby further reducing the effective sample size available for training. The model may learn incorrect associations or fail to capture the minority classes effectively (Frenay and Verleysen, 2014).

Proper interpretation of small datasets is considered vital, with domain expertise playing a crucial role. Without expert guidance, events may be mislabelled, crucial features may be ignored or domain-specific transformations may fail to be utilized by data scientists. A NN might focus on spurious correlations, underutilizing the limited data's real informative cues. Even when the technical capability to store large amounts of data is possessed by an organization, sufficient financial or human resources might not be allocated to properly label or maintain those datasets. As a result, data projects may be left underfunded or understaffed, ultimately rendering the limited dataset in practice. The final dataset remains small or partially labelled causing lower NN performance. Datasets may be stored in different formats or governed by separate systems, leading to technical integration challenges. The merging of data from diverse sources can increase small data problems if large portions are found to be incompatible or if key linkages are missing.

Data Reproducibility and Poisoning

Data can be obtained from one-time or non-repeatable scenarios such as historical events, catastrophic system failures, or scientific studies conducted over time. Once these events have concluded or the conditions have shifted, capturing new data can become expensive or impossible. Consequently, the original dataset may remain small and fixed, and the NNs and their training process must be optimized given the limited amount of data.

Rare events or difficult-to-observe phenomena such as rare diseases, major failures in manufacturing, or extreme weather conditions are often encountered in many real-world problems. By definition, these events occur infrequently, so sufficient training data is not easily gathered. Even when data is collected over a long period, only a small fraction of the samples may contain the crucial signal. Certain phenomena are inherently infrequent, making the capture of enough correct examples difficult. Expensive equipment, specialized expertise, or ethically sensitive procedures are required by some domains to collect data, thereby limiting the amount of available data (Hanke et al., 2009). The capturing of rare events might require continuous monitoring over an extended period, which is not always possible.

The reliability and truthfulness of data are influenced by the planned data collection process as well as the reliability and experience of the staff creating the dataset. The dataset may contain incorrect data resulting from unintentional or intentional poisoning of the dataset. In ML, data poisoning is characterized by the insertion or modification of malicious data within the training process aiming to influence the model's learned parameters and outcomes. Redundancy, which could support average out the effect of corrupted samples, is typically lacking in small datasets.

One example of a dataset poisoning attack scenario is a backdoor attack when a trigger (e.g., a particular pixel pattern in images) is embedded by adversaries into some training samples to cause the model to misclassify only when the trigger is present (Chen et al., 2017). Another type of poisoning attack is targeted data poisoning, in this case, specific training samples are manipulated by an attacker so that the model fails on certain inputs or classes (Biggio et al., 2013). In common dataset poisoning scenario instead of inducing targeted misclassifications, the overall model performance is intended to be degraded to the point where it is no longer practically usable. Potentially successful cleansing of the dataset from poisoned data may lead to a further reduction in size and expansion of imbalance in the dataset.

Summary

The analysis identifies financial constraints, regulatory and ethical considerations, technical barriers, and reproducibility challenges as primary causes for limited and imbalanced datasets in ML. Strict regulations, including GDPR, HIPAA, and CSL, impose restrictions on data access and cross-border exchange, while high acquisition costs, inadequate labelling, and limited computational resources further constrain dataset quality. Additionally, the limited availability of rare-event data and vulnerability to data poisoning amplify these limitations, undermining model robustness and generalizability. Addressing these issues is essential for properly creating a dataset that will impact the final performance of the trained model.

Acknowledgment

This research work was funded by the project of the Military University of Technology titled: "New Neural Network Architectures for Signal and Data Processing in Radiocommunications and Multimedia." Project No. UGB 22-054/2025.

Bibliography

- Biggio, B., Nelson, B. and Laskov, P. (2013) "Poisoning Attacks against Support Vector Machines." arXiv. Available at: https://doi.org/10.48550/arXiv.1206.6389.
- CDC (2024) Health Insurance Portability and Accountability Act of 1996 (HIPAA), Public Health Law. Available at: https://www.cdc.gov/phlp/php/resources/health-insurance-portability-and-accountability-act-of-1996-hipaa.html (Accessed: September 26, 2025).
- Chen, X. et al. (2017) "Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning." arXiv. Available at: https://doi.org/10.48550/arXiv.1712.05526.
- DigiChina (2017) "Translation: Cybersecurity Law of the People's Republic of China (Effective June 1, 2017)". Available at: https://digichina.stanford.edu/work/translation-cybersecurity-law-of-the-peoples-republic-of-china-effective-june-1-2017/ (Accessed: September 26, 2025).
- Frenay, B. and Verleysen, M. (2014) "Classification in the Presence of Label Noise: A Survey," IEEE Transactions on Neural Networks and Learning Systems, 25(5), pp. 845–869. Available at: https://doi.org/10.1109/TNNLS.2013.2292894.
- Hanke, M. et al. (2009) "PyMVPA: a Python Toolbox for Multivariate Pattern Analysis of fMRI Data," Neuroinformatics, 7(1), pp. 37–53. Available at: https://doi.org/10.1007/s12021-008-9041-y.
- Regulation - 2016/679 - EN - gdpr - EUR-Lex (2016). Available at: https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng (Accessed: September 26, 2025).
- Subrahmanyam, A. (2019) "Big data in finance: Evidence and challenges," Borsa Istanbul Review, 19(4), pp. 283–287. Available at: https://doi.org/10.1016/j.bir.2019.07.007.