Chapter 3
Shedding Light on Dataset Influence for More Transparent Machine Learning

Venkata Surendra Kumar Settibathini
https://orcid.org/0009-0000-6091-2632
Intellect Business Solutions, USA

Ankit Virmani
https://orcid.org/0009-0000-9290-8056
Google Inc., USA

Manoj Kuppam
https://orcid.org/0009-0006-4696-5280
Independent Researcher, USA

Nithya S.
Dhaanish Ahmed College of Engineering, India

S. Manikandan
Dhaanish Ahmed College of Engineering, India

Elayaraja C.
Dhaanish Ahmed College of Engineering, India

ABSTRACT

From healthcare to banking, machine learning models are essential. However, their decision-making processes can be mysterious, challenging others who rely on their insights. The quality and kind of training and evaluation datasets determine these models' transparency and performance. This study examines how dataset factors affect machine learning model performance and interpretability. This study examines how data quality, biases, and volume affect model functionality across a variety of datasets. The authors find that dataset selection and treatment are crucial to transparent and accurate machine learning results. Accuracy, completeness, and relevance of data affect the model's learning and prediction abilities. Due to sampling practises or historical prejudices in data gathering, dataset biases can affect model predictions, resulting in unfair or unethical outcomes. Dataset size is also important, according to our findings. Larger datasets offer greater learning opportunities but might cause processing issues and overfitting. Smaller datasets may not capture real-world diversity, resulting in underfitting and poor generalisation. These views and advice are useful for practitioners. These include ways for pre-processing data to reduce bias, assuring data quality, and determining acceptable dataset sizes. Addressing these dataset-induced issues can improve machine learning model transparency and effectiveness, making them solid, ethical tools for many applications.

DOI: 10.4018/979-8-3693-1355-8.ch003

INTRODUCTION

MACHINE learning models, with their vast applications in domains like recommendation systems, autonomous vehicles, healthcare, and finance, have become pivotal in the landscape of modern technology (Cao et al., 2018; Radha et al., 2020). However, the inherent opacity of these models often raises significant concerns about their reliability, trustworthiness, and fairness (Xie et al., 2019; Aceto et al., 2018). A central element in addressing these concerns lies in the transparency of these models, which hinges heavily on the datasets used for their training and evaluation (Abbassy, 2020). This paper delves deeply into how the choice of datasets critically impacts machine learning outcomes, aiming to illuminate the intricate and often challenging aspects of dataset selection (Abbassy & Abo-Alnadr, 2019).

Transparency in machine learning is not merely a technical requirement but a vital aspect that enables models to provide results that are understandable, interpretable, and justifiable (Ahmed Chhipa et al., 2021). The need for transparent models escalates, particularly in high-stakes sectors like healthcare and finance, where the implications of decisions can be profound (Amer & Shoukry, 2023). Transparent models foster trust among users and stakeholders, aid in regulatory compliance, and support more effective and responsible decision-making processes (Angeline et al., 2023). Conversely, models lacking transparency or opaque models often conceal the logic behind their predictions, posing challenges in gaining trust and undergoing scrutiny (Bose et al., 2023).

The pivotal role of dataset selection in the development of transparent machine-learning models cannot be overstated (Chakrabarti & Goswami, 2008). The nature of the dataset can introduce inadvertent biases, influence the model's ability to generalize across different scenarios and affect its competence in handling atypical or edge cases (Cirillo et al., 2023). Crucial factors that determine the efficacy and integrity of a dataset include its quality, the volume of data, the diversity of the samples, and the representation of different groups within the data (Das et al., 2023). This paper investigates these factors in detail, exploring how each aspect of dataset selection can significantly sway the behavior and performance of machine learning models (Devi & Rajasekaran, 2023).

Quality of data is a primary consideration; low-quality data can lead to inaccurate or misleading model predictions (Dhinakaran et al., 2023). Quality is determined by factors such as accuracy, completeness, consistency, and relevancy of the data to the problem at hand (Gaayathri et al., 2023). Additionally, the quantity of data is equally important (Harendharan & Boussi Rahmouni, 2023). Insufficient data can hinder a model's ability to learn effectively, leading to poor performance, while an abundance of data can enhance its learning capabilities and predictive accuracy (Jain et al., 2023).

Diversity in the dataset ensures that the model is exposed to a wide range of scenarios, reducing the risk of bias and enhancing its ability to function effectively across varied conditions (Jeba et al., 2023). A lack of diversity can result in models that perform well in certain environments but fail in others, particularly those that are underrepresented in the training data (Kanyimama, 2023). Representation, closely related to diversity, addresses the need for the dataset to encompass a broad spectrum of characteristics, particularly in fields like healthcare, where demographic factors such as age, gender, and ethnicity can significantly influence outcomes (Lodha et al., 2023).

The paper explores the challenges in dataset selection, such as the availability of high-quality, diverse, and representative data (Magare et al., 2020). It discusses the trade-offs that often need to be made between these factors due to constraints in data availability or collection. The ethical considerations in dataset collection and the need to avoid invasive or biased data-gathering methods are also examined (Marar et al., 2023).