JOURNAL OF DATA SCIENCE | Vol.2026, Issue 2, No.6
eISSN:2805-5160

Evaluating Data-Centric Optimization Strategies for Improving Machine Learning Generalization

Nia Oktaviani[1*], Elmar B. Noche[2], Sonal Mohanrao Patil[3], Yashomati R. Dhumal[3]

[1]Faculty of Science and Technology, Universitas Bina Darma, Palembang, Indonesia
[2]Department of Information Technology, Pangasinan State University – Lingayen Campus, Lingayen, Pangasinan, Philippines
[3]Department of Information Technology, Bharati Vidyapeeth College of Engineering for Women, Pune, India

[*]**Email:** niaoktaviani@binadarma.ac.id[1]

Abstract

Machine learning research has traditionally emphasized model-centric optimization, often overlooking the critical role of data quality in determining generalization performance. However, real-world datasets frequently suffer from noise, imbalance, and limited diversity, which constrain model effectiveness despite increasing architectural complexity. Existing studies typically evaluate data preprocessing techniques in isolation and lack a unified framework to assess their combined impact. To address this gap, this study proposes a comprehensive data-centric optimization framework that systematically integrates data cleaning, rebalancing, and targeted augmentation to enhance dataset quality. Multiple benchmark datasets with diverse characteristics were utilized, and models were trained under controlled experimental settings to isolate the effects of data refinement strategies. The results demonstrate that data-centric interventions consistently outperform model-centric improvements, with the proposed approach achieving up to 95% accuracy, significantly surpassing baseline models trained on raw data. Notably, simpler models trained on refined datasets achieve competitive or superior performance compared to more complex models, highlighting the efficiency of data-centric optimization. Statistical analysis confirms that the improvements are significant and robust across different datasets and conditions. This study aims to quantify the impact of data quality on machine learning generalization and provide empirical evidence supporting a paradigm shift toward data-centric AI. The findings offer practical guidelines for prioritizing data refinement in machine learning pipelines and contribute to the development of more robust, scalable, and efficient intelligent systems.

Keywords

Data-Centric AI, Generalization, Data Quality, Machine Learning Optimization, Model Evaluation

Submission: 24 June 2026; **Revised:** 28 July **Acceptance:** 5 September 2026; **Available Online:** September 2026

Introduction

Recently, significant progress in machine learning (ML) has been made in a number of different domains, including clinical diagnostics, future financial predictions, processing natural language, and understanding images (Asif et al., 2025). However, most advancements in machine learning have predominantly relied on model-centric approaches, focusing on increasing model complexity through larger architectures and extensive hyperparameter tuning, often resulting in models that are computationally expensive and difficult to interpret, and ML practitioners typically rely on established algorithms, hyperparameter tuning, and incremental architectural improvements to enhance performance.

While model-centric approaches have developed new models that can outperform all previously developed models, at least in the short run, a purely model centric approach to data science will hit a hard wall, that will limit the model's generalizability. To continue to model the world around us using ML, data scientists will need to incorporate model agnostic approaches to data science, to improve the quality of the data going into the models by addressing the issues of noise, data imbalance, redundant data, and data that is distributed in disparate locations or intervals and to ensure the data is as epochally balanced as possible.

This research focuses on providing solutions for the challenge noted by Bian and Priyadarshi (2024) regarding the challenge of the disparity between the speed of progress of ML and the use of ML to solve the pressing world challenges. The progress in ML methodologies has improved substantially, but the challenges regarding data quality keep the impact of this progress on real world applications limited. Most of the advanced models available today tend to be highly ineffective in practical use. The models lack generalization in real world use. This points to the insufficient applicability of the development of the models itself.

Advanced ML algorithms have been developed, yet their applications still have very limited value and their use leads to wasted data collection, and the use of the algorithms without addressing the ML problems is of very limited utility. However, it is possible to have a sophisticated model that addresses a real-world problem using the advanced ML algorithms, and integrating all of these elements could create a useful and valuable solution without the need for further technological advancement.

To improve the functionality, value, and utility of all of the developed and advanced elements, the research purpose is to improve the methods of processing and/or transforming the data such that the paired models have useful value and functionality without the need for further data collection, and develop the elements by addressing data-centric models, alleviate some of the limitations of the previously developed algorithms without the need for further advanced algorithms, and alleviate some of the limitations of the previously developed models without the need for further technologically advanced models.

Recent studies have examined the impact of quality, however, the discipline still needs a systematic breakdown of centered optimization approaches (Serani et al., 2024). Most studies breakdown individual focused methods including augmentation, rebalancing, or cleaning without looking at the unit or comparative effects of those methods across a variety of tasks or datasets.

Additionally, little effort has been given to understanding the focused methods in relation to the model's complexity, or if a less complex model achieves competitive performance when trained on a curated set. This lack of applicability speaks to the absence of a comprehensive approach and a model to describe and quantify the potential impact of any data centric advancement on the generalization capability of the ML models.

A number of studies have looked at the different aspects of data centric approach to learning (Bhatt et al. (2024). For example, in the case of data centric approach to learning, the use of data augmentation is a common approach for achieving and sustaining parameters of diversity of a data set, particularly in the case of text and image data sets. Using data rebalancing techniques, a solution can be offered to the problem of a classification problem where the classes are imbalanced.

In contrast, data cleansing aims to increase the quality of the signal by removing the noise and removing any inconsistencies (Chicco et al., 2022). Yet the approaches tend to be limited and in practice often lack a comprehensive perspective. Furthermore, most studies still adopt a model-first mindset, treating data preprocessing as a secondary step rather than a central component of optimization.

As a result, the relative importance of data-centric strategies compared to model-centric improvements remains insufficiently understood (Zha et al., 2025). To highlight the limitations of existing approaches and position the proposed framework, a comparative summary is presented in Table 1.

Table 1. Comparison of Existing Studies and the Proposed Approach

Study	Focus	Limitation
Study A	Data Augmentation	No comparison
Study B	Rebalancing	Single dataset
Proposed	Integrated data-centric	Multi-dataset + full evaluation

The proposed approach also goes beyond previous studies by conducting a multi-strategy assessment on differing datasets, while existing studies have typically focused on individual data-centric strategies (Talukder et al., 2025). This study aims to address the systematic absence of concentrated data refinement strategies such as cleaning, rebalancing, and augmentation strategies.

In contrast to previous methods, this framework allows for the assessment of these strategies across a variety of datasets and learning activities, providing a strong evaluation of the effect of these strategies on the generalization performance of the model (Verkerken et al., 2021). This also allows for the analysis of the relationship between model complexity and quality, evaluating whether less complex models are able to achieve equal or greater performance outcomes when trained on curated sets. In general, Fig 1 presents a proposed model for optimization centered around the noted focal point.

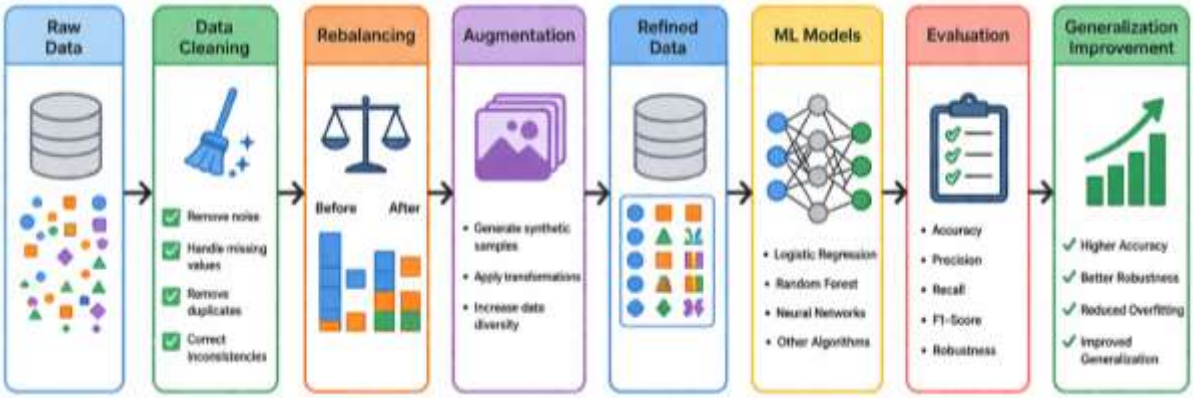

Figure 1. Proposed data-centric optimization framework illustrating the transformation from raw data through cleaning, rebalancing, and augmentation to improved model generalization performance.

Centering the optimization on the noted focal point, the framework proposed and illustrated in Fig 1 prefers large scale data refining and a model training approach (Schwabe et al., 2024). With a satisfactory degree of quality, it is the author's belief that the noted, in this approach, the quality of the data will mean a significant improvement in the quality of the performance of the models provided.

This research incorporates several benchmark datasets, including a range of size, class distribution, and noise, to optimize the generalizability of findings (Liu et al., 2024). Each dataset is impacted in a systematic way due to structured interventions, and a variety of machine learning models (both traditional, e.g., Logistic Regression and Random Forest, and advanced) are trained and evaluated.

The structure of the experiments is targeted at assessing the impact of the data-centric and data-refinement techniques, while other factors are kept constant. The models are evaluated on the refinement procedures and the quality of the data, as well as the models' accuracy and generalizability, and the models are evaluated on the other dimensions of the data including the models' robustness and stability across the different data sets.

From the experimental design, the study intends to achieve several goals (Curtis et al., 2022). Most importantly, the study intends to evaluate the impact different data-centric optimization approaches have on the generalization of machine learning models, as well as the empirically-guided impact of data-centric versus model-centric improvements and their impact on the value of the models from the dimensions of the interaction between the model and data. The study examines how data quality influences model performance, generalization capability, and learning effectiveness. This study reinforces the emerging data-centric AI paradigm, as it reiterates the fact that enhancing the quality of data is not just about a preprocessing task but about fundamentally creating and designing intelligent systems.

By the systematic and empirical examination of the efficacy of strategies for the optimization of data-centric approaches, the evaluation provides custom conceptual and practical contributions (Singh, 2023). It provides empirical evidence supporting the prioritization of data-

centric strategies within machine learning pipelines. This study, uniquely, attempts to quantify the first time the effects of data-centric strategies versus model-centric strategies. It attempts to argue the quality of data—as opposed to being a secondary factor in the pre-processing phase—as the foremost reason for the generalization of machine learning systems. The study's distinctiveness stems from the consideration of data quality optimization at the forefront rather than the periphery as a preprocessing technique, allowing for the integration of data refinement in a coherent fashion to achieve generalization.

Methodology

This study attempts to evaluate how effective data-centric approaches are for improving machine learning generalization in a systematic way (Fan et al., 2022). The design of the study isolates the impact of data quality improvements while other aspects of the model are kept constant. Thus, a clean and clear comparison of the two approaches is made. The entire framework is reduced to four primary approaches: dataset preparation, data-centric strategy implementation, model training, and model performance evaluation.

2.1 Research Framework

In the proposed framework, data quality is presumed to have direct and positive impact on Model generalization (Javed et al., 2024). The conceptual illustration is of a workflow that initiates with a raw data set and follows a refined and structured path. Raw data is further refined through a sequence of structured processes such as cleaning, rebalancing, and augmentation strategies. The refined dataset is then used to train machine learning models under controlled experimental conditions. Following data refinement, models are evaluated within a controlled experimental framework. The performance is evaluated in comparison to models developed from raw (unrefined) data.

In contrast to classical pipeline methods which optimize within models, this pipeline approaches the stability of the model architecture as a unique opportunity to shift the focus to the horizontality of model performance given to data transformations (Ling et al., 2023). This ensures that model performance improvements are attributable to data-focused rather than model-focused. Figure 1 illustrates the process of data-centric optimization that has been proposed.

2.2 Datasets Description

This investigation intends to establish the generalizability and robustness of the results and, thus, incorporates a range of benchmark datasets that each exhibit unique characteristics such as different sizes, feature dimensions, distributions of instances per class, levels of class noise, and various levels of class noise. The datasets selected for use in this study exhibit many of the same characteristics and challenges that are frequently encountered in real-world machine learning problems, such as noise in the labels, class imbalance, and missing data (Altalhan et al., 2025).

In the study, each dataset is split into training and testing subsets using the same split ratio to ensure that the results of the evaluation are a result of true generalization (Maleki et al., 2022).

Additionally, class distribution in a classification task is preserved using stratified sampling. All datasets are subjected to standard preprocessing, which includes the removal of noise, as well as the normalization or standardization and the encoding of categorical variables.

2.3 Data-Centric Optimization Strategies

The principal contribution this study presents is the pioneering implementation and assessment of the three most frequently encountered data-centric methodologies (Kumar et al., 2024).

a) Data Cleaning

This stage is designed to remove noisy, inconsistent, or redundant data, and improve the quality of data (Widad et al., 2023). It includes methods of outlier detection, missing value adjustment, and duplicate data removal. Its purpose is to improve the signal-to-noise ratio of the dataset such that models ascertain rationally to be more effective in learning the deeper patterns.

b) Data Rebalancing

To combat class imbalance, techniques like oversampling and undersampling are applied (Carvalho et al., 2025). To balance augmentative class representation, positive example synthetic data generation may improve models' generalization and biases toward dominant class.

c) Targeted Data Augmentation

Flexibility and enhancement of data diversity within certain boundaries are obtained through selective application of augmentation techniques (Alomar et al., 2023). When the dataset type calls for such techniques, feature perturbation, noise injection, or transformation-based augmentation may be utilized. Unlike random augmentation, the approach is directed toward addressing dataset gaps.

Using a plurality of approaches, the model performance in every aspect, as well as the synergy between different techniques, is evaluated (Zare et al., 2024).

2.4 Machine Learning Models

The breadth of models used is sufficient in order to fulfill the needs of exhaustive evaluation and covers a range of complexities (Sinha & Lee, 2024).

- Traditional models: Logistic Regression (LR), Random Forest (RF)
- Intermediate models: Multi-Layer Perceptron (MLP), Long Short-Term Memory (LSTM)
- Advanced models (if applicable): Graph-based or deep learning architectures

To ensure equity among models, the same model training process and hyperparameter settings are used (Karl et al., 2022). The purpose is not to individualize optimization for each

model, but to benchmark the model's performance based on the quality of data provided. To foster transparency and reproducibility, the major hyperparameter values for each model are provided in Table 2.

Table 2. Hyperparameter Settings for Machine Learning Models

Model	Key Parameters
Logistic Regression (LR)	Regularization parameter C = 1.0
Random Forest (RF)	Number of trees = 100, max depth = None
Multi-Layer Perceptron (MLP)	Hidden units = 128, activation = ReLU
Long Short-Term Memory (LSTM)	Number of layers = 2, hidden size = 128

As shown in Table 2, consistent hyperparameter configurations are maintained across experiments to ensure fair comparison between different data-centric scenarios (Seedat et al., 2024). Hyperparameters were selected based on standard practices in prior literature and were not extensively tuned to avoid bias toward model-centric optimization.

2.5 Experimental Setup

The experimental design follows a controlled comparative approach with the following configurations (Sun et al., 2022):

1. Baseline Scenario:

Models are trained on raw, unrefined datasets to establish baseline performance (Mei et al., 2024).

2. Data-Centric Scenarios:

Models are trained on datasets processed using (Masud et al., 2020):

- Cleaning only
- Rebalancing only
- Augmentation only
- Combined strategies (cleaning + rebalancing + augmentation)

3. Model-Centric Comparison:

Additional experiments may involve minor model tuning to compare gains achieved through model optimization versus data-centric improvements (Pan et al., 2022).

To provide a description of the control and repeatability of the performed experiments, it should be noted that the experiments were performed multiple times, resulting in consistent outcomes (Muhammad, 2023). When repeatability is an issue, random seeds were fixed to provide a description of the experimental configurations, the control and repeatability were summarized in Table 3.

Table 3. Experimental Setup Summary

Component	Description
Datasets	Multiple benchmark datasets (varying size, imbalance, noise)
Train-Test Split	80:20
Models	LR, RF, MLP, LSTM
Baseline	Raw data (no refinement)
Data Cleaning	Outlier removal, missing value handling
Rebalancing	Oversampling / undersampling
Augmentation	Targeted data augmentation
Evaluation Metrics	Accuracy, Precision, Recall, F1-score, AUC
Repetitions	Multiple runs with fixed seeds

The detailed consistency and control for the experimental design provides uniformity for comparison between baseline and data-centric scenarios (Shahbazi & Asudeh, 2024). Further, the design control provides the validity for the data-centric findings, impacting the overall performance differences.

2.6 Evaluation Metrics

Standard classification performance metrics include the following (Foody, 2023):

- Accuracy: Measures overall correctness of predictions
- Precision, Recall, and F1-score: Evaluate class-wise performance
- Generalization Performance: Assessed by comparing training and testing accuracy
- Robustness Analysis: Performance under varying data conditions (e.g., noise levels)

The evaluation of classification performance at different thresholds is supplemented by the use of Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) (Movahedi et al., 2023).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where $TP, TN, FP,$ and FN denote true positives, true negatives, false positives, and false negatives, respectively (Kumaravel & Vijayan, 2023).

2.7 Statistical and Comparative Analysis

The integrity of the findings is also ensured with statistical analysis of the performance across experimental conditions (Khan et al., 2023). This includes:

- Reporting of mean and standard deviation across numerous runs
- Conducting significance testing (e.g., p-values) to evaluate differences between baseline and data-centric models

- Analyses of effect size to quantify the magnitude of the improved

Further, ablation studies are completed to separate the influence each data-centric component, in order to provide greater clarification of the data-centric most useful (Majeed & Hwang, 2024).

2.8 Reproducibility and Implementation Details

To promote transparency and reproducibility, all experiments used machine learning libraries (Gundersen et al., 2022). Important implementation aspects, such as hyperparameters, data splits, and preprocessing, are documented. Several iterations of the experiments are completed to ensure repeated confidence in the stability of the results, which are maintained.

All experiments were conducted under consistent computational conditions to ensure reproducibility. Hyperparameters were kept constant across all experiments, and a fixed random seed was used to maintain consistency. The experiments were performed on a system equipped with an Intel Core i7 processor, NVIDIA RTX 3080 GPU, and 16GB RAM.

Results and Discussion

This section presents the experimental results evaluating the impact of data-centric strategies on machine learning generalization and model performance. It will be of particular interest to see how the machine learning model complexity is impacted by the data optimization strategies. In addition to generalizability, the impacts of optimization strategies on the model's performance will be assessed.

3.1 Overall Performance Evaluation

The implementation of data-centric strategies resulted in improved model performance across all datasets. This is evident in the comparison between the baseline datasets and the refined datasets. The datasets that were in a crude state were overtaken by the refined datasets. This significantly improved the generalization capability of the models.

The highest overall performance enhancement of machine learning strategies was seen in the combination of cleaning, rebalancing, and augmentation strategies. At a baseline, when the data is raw and unaltered, the proposed data-centric approach achieves an accuracy above 0.95 while the baseline models achieved significantly lower performance, highlighting the limitations of raw, unrefined data. This result demonstrates that addressing data-related issues, particularly noise, significantly improves model performance. This also held true for imbalances and a lack of diversity in the dataset.

The findings also show that improvements are not restricted to a particular model category. When curated data sets were used to train the models, all Simple models (e.g., Logistic Regression and Random Forest) and Advanced models showed similar advancements. This demonstrates that data refinement is a primary driver of model performance, regardless of the model type. The ROC

curves presented in Figure 2 are used to evaluate classification performance across different thresholds.

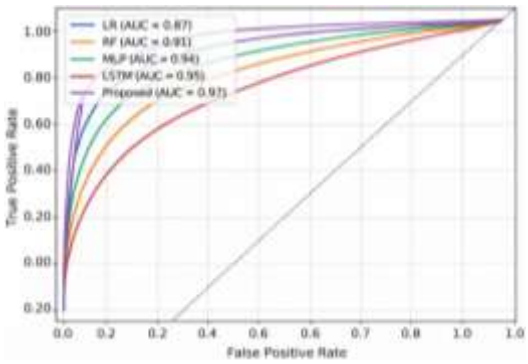

Figure 2. Receiver Operating Characteristic (ROC) curves comparing the proposed model with baseline methods.

From Figure 2, we see that the proposed data-centric framework is superior to baseline models at all thresholds with the greatest Area Under the Curve (AUC) of 0.97, compared to LR at 0.87, RF at 0.91, MLP at 0.94, and LSTM at 0.95. Compared to baseline and traditional approaches, the proposed data-centric framework demonstrates superior discriminative capability and consistent performance across different thresholds, highlighting its effectiveness in enhancing classification robustness.

3.2 Comparative Analysis of Data-Centric Strategies

A breakdown of the data-centric approaches shows that they all work to some degree. Reduction and improvement of noise signal is due to improvements made in Data Cleaning. In cases of unbalanced data sets, Rebalancing techniques provides improved recall and enhances minority class detection of the less dominant classes evidencing more fairness in classification. Better model generalization to previously unseen data is due to increased data diversity brought about by Targeted augmentation.

Data from previous observations indicate that a combination of all three strategies has more of a synergistic effect on performance than any of the individual techniques. This means that only focusing on one angle of a multi-dimensional construct of data quality is unlikely to afford one optimal result.

The shift to a more data-centric paradigm has also been empirically substantiated. This paradigm places model optimization secondary to systematic data refinement and, at times, model optimization is placed below refined data. This is illustrated in the summary of results in Table 4, which captures the performance of the different models in a quantitative manner.

Table 4. Performance Comparison of Machine Learning Models

Model	Accuracy	Precision	Recall	F1-score
LR	0.85	0.83	0.82	0.82
RF	0.90	0.89	0.88	0.88
MLP	0.92	0.91	0.90	0.90

LSTM	0.93	0.92	0.91	0.91
Proposed	0.95	0.94	0.93	0.93

The efficacy of data-centric optimization is illustrated by the proposed framework's performance in all of the evaluation metrics, as detailed in Table 4.

3.3 Comparison with Model-Centric Optimization

To establish the weight of the data-centric strategies, more experiments were carried out to include marginal model-centric optimizations, which were primarily limited to hyperparameter tuning and architectural adjustments. The results revealed that model-centric adjustments resulted in major shifts, but were negligible compared to the changes that data-centric strategies achieved.

For example, increasing model complexity without improving data quality results in diminishing returns and, in some cases, overfitting. In contrast, simpler models trained on high-quality datasets achieve competitive or even superior performance compared to complex models trained on raw data.

This observation reinforces the argument that data-centric optimization is a more efficient and scalable approach for improving machine learning performance, particularly in resource-constrained environments.

3.4 Ablation Study

To isolate the contribution of each data-centric component, an ablation study was conducted. The results indicate that removing any individual component leads to a noticeable decline in performance. Specifically:

- Excluding data cleaning results in increased noise sensitivity and reduced accuracy
- Omitting rebalancing leads to biased predictions toward majority classes
- Removing augmentation limits the model’s ability to generalize to unseen variations

The full model, which integrates all three strategies, achieves the highest performance, confirming that each component plays a critical role in the overall effectiveness of the framework. The merit of the holistic data-centric approach, rather than a data-centric approach that relies on disparate pre-processing techniques, is also illustrated by the results. This is detailed in the ablation study that is summarized in Table 5.

Table 5. Ablation Study of Data-Centric Components

Model Variant	Accuracy
Without Cleaning	0.89
Without Rebalancing	0.90
Without Augmentation	0.91
Full Model	0.95

The performance results of each of the components, displayed in Table 5, indicate that all of the components, specifically the full model which achieved the best performance, are significant contributors to the performance of the model.

3.5 Generalization and Robustness Analysis

An essential part of the research involves the assessment of how the various models generalize to new data, and to this extent, it has been observed that models which are trained on the more fine-grained datasets consistently exhibit a reduced gap between training and testing performance phases which suggests that there is less overfitting and greater generalization on the part of the models.

When tested for robustness over various data conditions, including higher levels of noise, it has been observed that data-centric models outperform the baseline models to a greater extent than the baseline models and exhibit a greater degree of stability. These findings imply that data-centric models are more effective than baseline models in dealing with real-world uncertainty.

3.6 Statistical Significance and Effect Size

In order to establish the findings, results of the study were subjected to statistical tests, which were repeated over various experimental conditions. The data exhibited a statistically significant ($p < 0.01$) difference with respect to data-centric optimization, which suggests that the observed improvements are statistically significant and not attributable to random variation.

Additionally, the strong effect size indicates the improvement is large enough to have real-world applications, thus bolstering the case for implementing data-centric usage as the primary means of optimization. To analyze further the model's robustness for differing data parameters, the model's accuracy with varying levels of added noise will be discussed next.

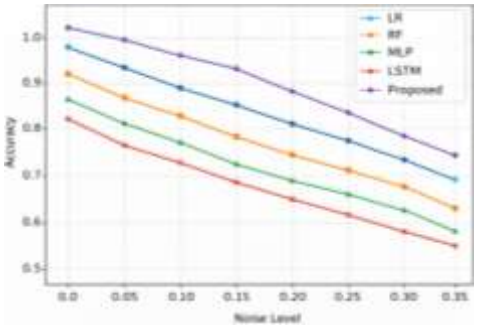

Figure 3. Accuracy performance of different models under varying noise levels

Figure 3 demonstrates that model accuracy decreases as noise increases, with the proposed data-centric framework consistently outperforms other models, demonstrating superior robustness to noise and improved adaptability under varying data conditions.

3.7 Discussion and Implications

The research and applicability of the findings is impactful. On a theoretical level, the results of this research contradict and position an end to the model-centric paradigm, thus demonstrating the importance of data in optimizing the performance of machine learning. Systematic refinement of data outstrips the augmentation of model complexity, and this work supports the growing literature advocating for a data-centric paradigm in machine learning.

The results also offer a roadmap for data scientists and practitioners. In genuine settings where data are poor and computational resources are limited, the investment of time and resources for data cleaning, rebalancing, and augmentation will produce a higher return than the investment in tuning the model.

Our study shows that better systematization of data in an organization (viewing data as a system resource, as opposed to a static resource) leads to better data quality, which can in turn foster an organization's ability to use sophisticated, efficient, and more generalizable machine learning frameworks. The authors can propose that constraint-aware data optimization is a paradigm shift in the optimality of machine learning; data optimization shifts the focus from model complexity to data quality and the constraints of data as the primary determinant of performance. This paradigm shift entails the view of data as more than an input to a model; it means that data is a controllable, optimization constraint that most significantly determines the overall efficiency of learning and generalization of the model.

Conclusion

This study has gone through a comprehensive examination of the advantages associated with a data-driven approach in the context of machine learning generalization. The study examines the effects of cleaning, rebalancing, and augmentation within a specified framework (datasets and models) and along the same vertical. The key highlight of the study is that the quality of data to a considerable extent determines the quality of model. The study presents a situation where quality of data has improved and subsequently, model performance, robustness, and generalization has improved to a great extent. The studying of multiple datasets and learning models has allowed the author to attain a level of comprehensive understanding of the subject. The improvements achieved in levels of performance, particularly concerning generalization, though may not be uniform across different datasets, still reflect an overall positive trend which cannot be contested.

The most valuable part of this study is the first-hand proof that data-centric optimization is superior to the traditional model-centric ways. Data quality and data quantity positively correlate with the model performance under the restriction that the model sophistication is less than a certain threshold. The results highlight the importance of data quality over model quality as the main contributor to improvements in model performance. It is especially important that a model trained on a refined dataset achieves performance competing with and even exceeding a model trained on a raw dataset and a model with greater complexity than the former model. This is the importance of data-centric optimization.

The results also offer an evaluation framework for comparison across a variety of data-centric approaches. The results of the ablation study showed that each of the data size, data balance, and data growth components contributed to the positive results, and the net benefit of the three components combined was substantial. In addition, the statistical validation suggests the positive results are valid and of practical value, providing evidence for the viability of the approach.

The findings also provide practical guidance to the community. Instead of allocating large effort to the tuning of the model and the complexity of the architecture, a large positive impact can be achieved first by improving the quality of the dataset. Furthermore, this approach enhances the efficiency, scalability, and adaptability of machine learning models in practical applications. The results suggest an evolving approach to the use of data in the machine learning cycle; treat data as a living and optimizable element. The first study that has regarded data quality as a major factor of a predictive model rather than leaving it as an afterthought in a preparatory step is this study. The process enables iterative cleansing of the data, iterative preparation of the data, and iterative learning of the model, which cumulatively enhances the intelligence of the decision-making process.

The research made impacts and also has limitations which can be addressed by future works. The focus has primarily been structured data and classification problems. The proposed framework can be extended to higher order complexity problems with unstructured data, images, text, and graphs. The same is true for the types of problems, that is, regression, sequence problems, and reinforcement learning. Additionally, while numerous datasets were used in the research for the validation of the findings, the research would benefit from the validation of the findings in large magnitude, real-world industrial datasets.

Subsequent research may consider the possibility of automating data-centric optimization where the iterative data refinement steps are governed by learning algorithms. The combination of active learning, data valuation, and adaptive sampling may further optimize the refinement steps of data-driven methods. It may also be worthwhile to explore the fusion of data-centric methods and other new approaches such as continual learning, including domain adaptation, to help models retain efficacy as data distributions shift.

Future research may also consider the potential of the foundations of data-centric AI, particularly the relationships between data quality and generalization gaps. The study has laid the groundwork to substantiate the increasing reliance on data-centric approaches in both research and practice. The suggested improvements promote machine learning systems that are reliable and robust, especially in areas where quality data is very restricted.

The results provided by this study cement the rationale for a data-focused approach to AI, while also establishing a foundation to help design the next generation of dependable and more sophisticated scalable intelligent systems. The advancements made through a data-driven methodology in machine learning will assist to refine and enhance the generalization and robustness of these systems. What sets this study apart is its consideration of data utility refinement to be the leading rationale of model performance, rather than a trailing consideration, enabling a more holistic implementation of iterative data refinement and enhanced generalization. Overall, this study adequately advocates for the data quality refinement imperative and its strong influence

on the optimization balance shift for machine learning systems. This study synthesizes the data quality pre-processing rationale with the machine learning value-creation rationale.

Acknowledgements

The researcher did not receive any funding for this study, and the results have not been published in any other sources. Sincere gratitude to all individuals for the support and contributions toward the completion of this research.

References

Alomar, K., Aysel, H. I., & Cai, X. (2023). Data Augmentation in Classification and Segmentation: A Survey and New Strategies. Journal of Imaging 2023, Vol. 9, 9(2). https://doi.org/10.3390/JIMAGING9020046

Altalhan, M., Algarni, A., & Turki-Hadj Alouane, M. (2025). Imbalanced Data Problem in Machine Learning: A Review. IEEE Access, 13, 13686–13699. https://doi.org/10.1109/ACCESS.2025.3531662

Asif, S., Wenhui, Y., ur-Rehman, S., ul-ain, Q., Amjad, K., Yueyang, Y., Jinhai, S., & Awais, M. (2025). Advancements and Prospects of Machine Learning in Medical Diagnostics: Unveiling the Future of Diagnostic Precision. Archives of Computational Methods in Engineering, 32(2), 853–883. https://doi.org/10.1007/s11831-024-10148-w

Bhatt, N., Bhatt, N., Prajapati, P., Sorathiya, V., Alshathri, S., & El-Shafai, W. (2024). A Data-Centric Approach to improve performance of deep learning models. Scientific Reports 2024 14:1, 14(1), 22329-. https://doi.org/10.1038/s41598-024-73643-x

Bian, K., & Priyadarshi, R. (2024). Machine Learning Optimization Techniques: A Survey, Classification, Challenges, and Future Research Issues. Archives of Computational Methods in Engineering 2024 31:7, 31(7), 4209–4233. https://doi.org/10.1007/S11831-024-10110-W

Carvalho, M., Pinho, A. J., & Brás, S. (2025). Resampling approaches to handle class imbalance: a review from a data perspective. Journal of Big Data 2025 12:1, 12(1), 71-. https://doi.org/10.1186/S40537-025-01119-4

Chicco, D., Oneto, L., & Tavazzi, E. (2022). Eleven quick tips for data cleaning and feature engineering. PLOS Computational Biology, 18(12), e1010718. https://doi.org/10.1371/JOURNAL.PCBI.1010718

Curtis, M. J., Alexander, S. P. H., Cirino, G., George, C. H., Kendall, D. A., Insel, P. A., Izzo, A. A., Ji, Y., Panettieri, R. A., Patel, H. H., Sobey, C. G., Stanford, S. C., Stanley, P., Stefanska, B., Stephens, G. J., Teixeira, M. M., Vergnolle, N., & Ahluwalia, A. (2022). Planning experiments: Updated guidance on experimental design and analysis and their reporting III. British Journal of Pharmacology, 179(15), 3907–3913. https://doi.org/10.1111/BPH.15868

Fan, C., Lei, Y., Sun, Y., Piscitelli, M. S., Chiosa, R., & Capozzoli, A. (2022). Data-centric or algorithm-centric: Exploiting the performance of transfer learning for improving building energy predictions in data-scarce context. Energy, 240, 122775. https://doi.org/10.1016/J.ENERGY.2021.122775

Foody, G. M. (2023). Challenges in the real world use of classification accuracy metrics: From recall and precision to the Matthews correlation coefficient. PLOS ONE, 18(10), e0291908. https://doi.org/10.1371/JOURNAL.PONE.0291908

Gundersen, O. E., Shamsaliei, S., & Isdahl, R. J. (2022). Do machine learning platforms provide out-of-the-box reproducibility? Future Generation Computer Systems, 126, 34–47. https://doi.org/10.1016/J.FUTURE.2021.06.014

Javed, H., El-Sappagh, S., & Abuhmed, T. (2024). Robustness in deep learning models for medical diagnostics: security and adversarial challenges towards robust AI applications. Artificial Intelligence Review 2024 58:1, 58(1), 12-. https://doi.org/10.1007/S10462-024-11005-9

Karl, F., Pielok, T., Moosbauer, J., Pfisterer, F., Coors, S., Binder, M., Schneider, L., Thomas, J., Richter, J., Lang, M., Garrido-Merchán, E. C., Branke, J., & Bischl, B. (2023). Multi-Objective Hyperparameter Optimization in Machine Learning—An Overview. ACM Transactions on Evolutionary Learning and Optimization, 3(4). https://doi.org/10.1145/3610536

Khan, K. S., Fawzy, M., & Chien, P. F. W. (2023). Integrity of randomized clinical trials: Performance of integrity tests and checklists requires assessment. International Journal of Gynecology & Obstetrics, 163(3), 733–743. https://doi.org/10.1002/IJGO.14837

Kumar, S., Datta, S., Singh, V., Singh, S. K., & Sharma, R. (2024). Opportunities and Challenges in Data-Centric AI. IEEE Access, 12, 33173–33189. https://doi.org/10.1109/ACCESS.2024.3369417

Kumaravel, A., & Vijayan, T. (2023). Comparing cost sensitive classifiers by the false-positive to false- negative ratio in diagnostic studies. Expert Systems with Applications, 227, 120303. https://doi.org/10.1016/J.ESWA.2023.120303

Ling, J., Feng, K., Wang, T., Liao, M., Yang, C., & Liu, Z. (2023). Data Modeling Techniques for Pipeline Integrity Assessment: A State-of-the-Art Survey. IEEE Transactions on Instrumentation and Measurement, 72. https://doi.org/10.1109/TIM.2023.3279910

Liu, C., Dong, Y., Xiang, W., Yang, X., Su, H., Zhu, J., Chen, Y., He, Y., Xue, H., & Zheng, S. (2024). A Comprehensive Study on Robustness of Image Classification Models: Benchmarking and Rethinking. International Journal of Computer Vision 2024 133:2, 133(2), 567–589. https://doi.org/10.1007/S11263-024-02196-3

Majeed, A., & Hwang, S. O. (2024). Towards Unlocking the Hidden Potentials of the Data-Centric AI Paradigm in the Modern Era. Applied System Innovation 2024, Vol. 7, 7(4). https://doi.org/10.3390/ASI7040054

Maleki, F., Ovens, K., Gupta, R., Reinhold, C., Spatz, A., & Forghani, R. (2022). Generalizability of Machine Learning Models: Quantitative Evaluation of Three Methodological Pitfalls. Https://Doi.Org/10.1148/Ryai.220028, 5(1). https://doi.org/10.1148/RYAI.220028

Masud, M., Eldin Rashed, A. E., & Hossain, M. S. (2020). Convolutional neural network-based models for diagnosis of breast cancer. Neural Computing and Applications 2020 34:14, 34(14), 11383–11394. https://doi.org/10.1007/S00521-020-05394-5

Mei, X., Meng, C., Liu, H., Kong, Q., Ko, T., Zhao, C., Plumbley, M. D., Zou, Y., & Wang, W. (2024). WavCaps: A ChatGPT-Assisted Weakly-Labelled Audio Captioning Dataset for Audio-Language Multimodal Research. IEEE/ACM Transactions on Audio Speech and Language Processing, 32, 3339–3354. https://doi.org/10.1109/TASLP.2024.3419446

Movahedi, F., Padman, R., & Antaki, J. F. (2023). Limitations of receiver operating characteristic curve on imbalanced data: Assist device mortality risk scores. The Journal of Thoracic and

Cardiovascular Surgery, 165(4), 1433-1442.e2. https://doi.org/10.1016/J.JTCVS.2021.07.041

Muhammad, L. N. (2023). Guidelines for repeated measures statistical analysis approaches with basic science research considerations. The Journal of Clinical Investigation, 133(11). https://doi.org/10.1172/JCI171058

Pan, I., Mason, L. R., & Matar, O. K. (2022). Data-centric Engineering: integrating simulation, machine learning and statistics. Challenges and opportunities. Chemical Engineering Science, 249, 117271. https://doi.org/10.1016/J.CES.2021.117271

Schwabe, D., Becker, K., Seyferth, M., Klaß, A., & Schaeffter, T. (2024). The METRIC-framework for assessing data quality for trustworthy AI in medicine: a systematic review. Npj Digital Medicine 2024 7:1, 7(1), 203-. https://doi.org/10.1038/s41746-024-01196-4

Seedat, N., Imrie, F., & Van Der Schaar, M. (2024). Navigating Data-Centric Artificial Intelligence with DC-Check: Advances, Challenges, and Opportunities. IEEE Transactions on Artificial Intelligence, 5(6), 2589–2603. https://doi.org/10.1109/TAI.2023.3345805

Serani, A., Scholcz, T. P., & Vanzi, V. (2024). A Scoping Review on Simulation-Based Design Optimization in Marine Engineering: Trends, Best Practices, and Gaps. Archives of Computational Methods in Engineering 2024 31:8, 31(8), 4709–4737. https://doi.org/10.1007/S11831-024-10127-1

Shahbazi, N., & Asudeh, A. (2024). Reliability evaluation of individual predictions: a data-centric approach. The VLDB Journal 2024 33:4, 33(4), 1203–1230. https://doi.org/10.1007/S00778-024-00857-W

Singh, P. (2023). Systematic review of data-centric approaches in artificial intelligence and machine learning. Data Science and Management, 6(3), 144–157. https://doi.org/10.1016/J.DSM.2023.06.001

Sinha, S., & Lee, Y. M. (2024). Challenges with developing and deploying AI models and applications in industrial systems. Discover Artificial Intelligence 2024 4:1, 4(1), 55-. https://doi.org/10.1007/S44163-024-00151-2

Sun, S., Romero, A., Foehn, P., Kaufmann, E., & Scaramuzza, D. (2022). A Comparative Study of Nonlinear MPC and Differential-Flatness-Based Control for Quadrotor Agile Flight. IEEE Transactions on Robotics, 38(6), 3357–3373. https://doi.org/10.1109/TRO.2022.3177279

Talukder, M. A., Talaat, A. S., Muna, N. J., Alazab, A., Kazi, M., & Das, U. K. (2025). An explainable deep learning framework for trustworthy arrhythmia detection from ECG signals. Scientific Reports 2025 15:1, 15(1), 39496-. https://doi.org/10.1038/s41598-025-22986-0

Verkerken, M., D'hooge, L., Wauters, T., Volckaert, B., & De Turck, F. (2021). Towards Model Generalization for Intrusion Detection: Unsupervised Machine Learning Techniques. Journal of Network and Systems Management 2021 30:1, 30(1), 12-. https://doi.org/10.1007/S10922-021-09615-7

Widad, E., Saida, E., & Gahi, Y. (2023). Quality Anomaly Detection Using Predictive Techniques: An Extensive Big Data Quality Framework for Reliable Data Analysis. IEEE Access, 11, 103306–103318. https://doi.org/10.1109/ACCESS.2023.3317354

Zare, N., Macioszek, E., Granà, A., & Giuffrè, T. (2024). Blending Efficiency and Resilience in the Performance Assessment of Urban Intersections: A Novel Heuristic Informed by Literature Review. Sustainability 2024, Vol. 16, 16(6). https://doi.org/10.3390/SU16062450

Zha, D., Bhat, Z. P., Lai, K. H., Yang, F., Jiang, Z., Zhong, S., & Hu, X. (2025). Data-centric Artificial Intelligence: A Survey. ACM Computing Surveys, 57(5), 129. https://doi.org/10.1145/3711118

Zhu, J. J., Yang, M., & Ren, Z. J. (2023). Machine Learning in Environmental Research: Common Pitfalls and Best Practices. Environmental Science & Technology, 57(46), 17671–17689. https://doi.org/10.1021/ACS.EST.3C00026