Empirical Analysis of Dataset Size Impact on Classification Performance in Precision Agriculture Using Machine Learning Models

Khadija Lechqar*, Mohammed Errais
Data Engineering and Intelligent Systems Laboratory, Faculty of Sciences Ain Chock, Hassan II university, Casablanca, Morocco
E-mail: khadija.lechqar@gmail.com
*Corresponding author

Keywords: Size of dataset, performance, precision agriculture

Received: January 24, 2025

This study empirically investigates the relationship between dataset size and classification performance in precision agriculture applications. Seven machine learning models (Decision Tree, Random Forest, Logistic Regression, SVM, Gaussian Naïve Bayes, KNN, and AdaBoost) were evaluated on seven agricultural datasets ranging from 100 to 4,000 samples. Performance was assessed using five metrics: accuracy, precision, recall, F1-score, and ROC-AUC. The methodology involved two phases: initial evaluation using complete datasets, followed by systematic analysis of subdivided datasets to examine performance variation with data volume. Statistical analysis using Pearson correlation coefficients revealed no significant correlation between dataset size and model performance ($r = 0.12$, $p > 0.05$). Results indicate that Random Forest and Decision Tree models achieved the highest average performance across datasets (88.48% and 85.37% accuracy, respectively). The findings suggest that dataset quality and problem characteristics have greater influence on classification performance than dataset size alone in precision agriculture applications.

Povzetek: Empirična študija na sedmih kmetijskih naborih pokaže, da velikost podatkov ni značilno povezana z uspešnostjo klasifikacije (pomembnejša sta kakovost in narava problema), pri čemer se najbolje izkažeta naključni gozd in odločitveno drevo.

1 Introduction

Precision agriculture represents a paradigm changing in agricultural practices, designed to address the growing global demand for food production through data-driven decision-making. Modern precision agriculture has adopted artificial intelligence as a core technology. Machine learning models are now increasingly used, starting from pre-harvest automation to irrigation scheduling, disease identification, weed control, yield forecasting, and pest management [1,2]. This technological changing has made farming operations more efficient while promoting sustainability [3].

Yet gathering agricultural data is problematic and directly affects how well machine learning systems work [4]. Farmers and researchers use various methods to collect information. Satellites and UAVs capture aerial images for remote monitoring [5,6]. Ground sensors placed near crops provide detailed, accurate readings [7]. Robots move through fields collecting data, while IoT devices monitor environmental conditions continuously. Linking all these data sources requires sophisticated communication networks [8]. The entire process costs significant money and time, demanding substantial investment in both equipment and expertise.

Datasets serve as fundamental building blocks for developing effective machine learning models [9]. Real-world agricultural data often suffers from quality issues due to environmental variables, sensor limitations, and data transmission challenges. Class imbalance, where minority and majority classes have unequal representation in classification problems, frequently leads to reduced model accuracy [10,11]. Previous research has suggested that larger datasets generally improve classification model performance [12], yet the specific relationship between dataset size and performance in precision agriculture contexts is underexplored.

However, one of the challenges of precision agriculture is the collection of data [4]. In fact, different devices are used for this purpose: satellites and UAVs for remote sensing process and aerial images[5][6], ground-based sensors which have the advantage of being close proximity to plants and providing good data accuracy[7], robots and other IoT devices. In addition, communication devices are also used to link these different components[8]. The cost of data collection is therefore high in terms of time and resources.

1.1 Research objectives and hypotheses

This study addresses the following research questions:

1. What is the relationship between dataset size and classification model performance in precision agriculture applications?

2. How do different machine learning algorithms respond to variations in dataset size?

3. What factors beyond dataset size influence model performance in agricultural classification tasks?

We hypothesize that while dataset size may influence performance, the relationship is not linear and that dataset quality characteristics play a more significant role in determining model effectiveness.

2 Related work

Several studies have examined the relationship between dataset size and machine learning model performance across different domains. This section provides a critical analysis of existing literature and identifies research gaps that this study addresses.

2.1 Literature review and gap analysis

Alshammari and Alshayeb [13] investigated software defect prediction, demonstrating that dataset size directly affects SVM model performance. Their results showed that datasets with fewer metrics enabled better SVM performance for defect prediction, although computational efficiency was not guaranteed for smaller datasets. However, this study was limited to software engineering applications and focused primarily on SVM models.

Imlawi and Alsharo [14] examined the impact of resampling techniques and dataset size on classification accuracy. They found that resampling methods were particularly beneficial for limited datasets (100-500 samples) and helped reduce biased estimation. While relevant to small dataset scenarios, their study did not address the specific challenges of agricultural data collection and processing.

Bailly et al. [15] focused on dataset size effects in prediction tasks using logistic regression and deep learning models. Their findings indicated that traditional machine learning algorithms were less influenced by dataset size but required feature interaction optimization for optimal performance, unlike deep learning models. This work provided insights into algorithm-specific responses to data volume but lacked domain-specific analysis for agriculture.

Althnian et al. [12] conducted an empirical study in the medical domain, concluding that model performance depends more on dataset distribution characteristics than size alone. They demonstrated that Adaptive Boosting (AB) and Naïve Bayes (NB) models showed greater robustness with limited medical data. However, their focus on medical applications limits direct applicability to agricultural contexts.

Lin et al. [16] studied multiclass classification performance as a function of dataset size and training/test split ratios, providing insights into data partitioning strategies but not addressing domain-specific agricultural challenges.

2.2 Summary of prior research

Table 1 summarizes key findings from previous studies on dataset size impact across different domains.

Table 1: Summary of related work

Study	Field	Dataset size range	Model used	Key findings
[13]	Software Engineering	100-2000	svm	Smaller datasets improved SVM performance
[14]	General Classification	100-500	Multiple	Resampling beneficial for small datasets
[15]	Biomedical	500-5000	LR, Deep Learning	Traditional ML less affected by size
[12]	Medical	1000-10000	Multiple	Distribution more important than size
[16]	General Classification	34-181	U-net	virtual sample generation method enhancing prediction accuracy for small datasets in high dimensions

2.3 Research gap identification

The literature review reveals several critical gaps that this study addresses:

1. Domain-specific analysis: No comprehensive study has examined dataset size impact specifically in precision agriculture contexts, where data collection challenges and environmental variables create unique constraints.

2. Systematic model comparison: Previous studies have not comprehensively compared multiple machine learning algorithms under consistent experimental conditions within agricultural applications.

3. Statistical validation: Many existing studies lack rigorous statistical analysis to validate claims about dataset size relationships.

4. Practical implications: Limited research has addressed the practical implications of dataset size constraints in resource-limited agricultural settings.

This study fills these gaps by providing a systematic empirical analysis of dataset size impact on classification

performance using multiple machine learning models specifically in precision agriculture applications.

3 Methodology

The main objective of this work, as mentioned above, is to study the effect of limited dataset on the performance of classification problems. It will be focused on tabular datasets in the agriculture field. As a first step, we selected seven datasets of different sizes and applied seven machine-learning models for classification. To take this study further, we worked on the two largest selected datasets and divided them into sub-datasets. The aim is to evaluate the variation of the performance compared with the size of dataset for the same classification problem.

3.1 Dataset selection and characteristics

We used seven datasets selected from precision agriculture field. The criteria of selection are:

1. Representativeness of common agricultural classification problems.
2. Availability of sufficient samples for subdivision analysis.
3. Diversity in feature types and problem complexity Balanced representation of different agricultural domains

The number of instances varies from 100 to 4000 and the number of features is between 4 and 9 features. Table 2 summarizes their structures.

Table 2: An overview of the datasets

Dataset	Field	Number of rows	Number of features	Type of data	balanced data
1	Fertilization	100	9	Numeric, Text	yes
2	Irrigation	201	4	Numeric	yes
3	Irrigation	501	6	Numeric	yes
4	Weather prediction	1461	8	Numeric, Text	no
5	Crop recommendation	2200	8	Numeric, Text	yes
6	Quality of water	3276	9	Numeric, Text	yes
7	Apple quality	4000	9	Numeric, Text	yes

3.2 Dataset subdivision strategy

To systematically examine dataset size impact, the two largest datasets (Dataset 6: 3,276 samples and Dataset 7: 4,000 samples) were randomly subdivided into smaller subsets while maintaining class distribution proportions (Table 3). The subdivision strategy was designed to create meaningful size variations for statistical analysis.

Table 3: Description of the two sub-datasets

Dataset	Field	Number of rows
6	Quality of water	3276
6-1		500
6-2		1000
6-3		2000
7	Apple quality	4000
7-1		500
7-2		1000
7-3		2000

3.3 Data preprocessing

We applied several preprocessing steps to clean our data and make it suitable for machine learning models:
Handling missing values: When data points were missing, we used different strategies depending on the variable type. For numerical data, we filled gaps with the average value. For categorical data, we used the most common category.
Scaling features: Since our numerical features had different ranges, we standardized them using z-score normalization. This prevents features with larger values from dominating the models.
Converting categories: Some models only work with numbers. So, we converted categorical variables into numerical format using one-hot encoding.
Managing outliers: We found outliers using the interquartile range method. Rather than removing these data points (which would shrink our already limited datasets), we used winsorization to cap extreme values.

3.4 Machine-learning models

For each dataset, we used seven machine-leaning models for classification. They are namely:
K-Nearest Neighbors KNN: is a model of supervised learning. The predicted value is based on the values of the K nearest neighbors based **using** Euclidean, mahalanobis, or Manhattan [1].
Decision Trees DT: are designed for supervised data mining. They have the structure of tree; each node presents a feature. It has the advantage of solving problems with complex data [3].
Random Forest RF: It is an ensemble machine learning approach. It consist on combining the results of multiple decision trees to reach a single result[2].
Gaussian Naïve Bayes GNB: is used for classification purposes. It is based on probabilistic approach and supposes that features have a normal distribution[4][17].
Logistic regression LR is a linear model giving the relationship between input and output variables. In fact, it predicts the probability that the input belongs to a class or not[5].

Adaptive Boosting AdaBoost is an ensemble model based on weighted instances of data in function of previous classifications. The final aim is to improve, gradually, the accuracy [6].

Support Vector Machine SVM is a supervised machine learning model. It consists on finding a line or a hyperplane according to the number of features in the input. The goal is to classify the data points in input [12].

3.5 Performance metrics

In this work, the performance of models is evaluated by using the known five metrics: accuracy precision recall, f1, roc auc. They give an idea about positive and negative

correctly or incorrectly classification and so the performance of the applied models:

Accuracy: Overall correctness of predictions [18].

Precision: Proportion of true positive predictions among positive predictions.

Recall: Proportion of true positive instances correctly identified

F1-Score: Harmonic mean of precision and recall

ROC-AUC: Area under the receiver operating characteristic curve

3.5 Experimental setup

Our experiments ran on:

Computer: Windows 10 machine with Intel Core i5 2.6 GHz processor and 8GB RAM

Programming Tools: Python 3.8.5, scikit-learn 0.24.2, pandas 1.3.3, numpy 1.21.2

Testing method: 5-fold cross-validation to get reliable performance estimates

Reproducibility: Set random seed to 42 so results can be replicated

3.6 Statistical analysis

We used several statistical methods to determine whether dataset size really affects performance:

Pearson correlation coefficients to measure linear relationships

ANOVA tests to compare performance across different size groups

Mann-Whitney U tests for cases where data wasn't normally distributed

4 Results and discussion

4.1 Performance analysis of complete datasets

The experimental results demonstrate varying performance patterns across different models and datasets, providing insights into the relationship between dataset size and classification effectiveness.

4.1.1 Accuracy analysis

Accuracy measurements across all models and datasets reveal significant variation in performance that does not correlate linearly with dataset size (Table 4). The average accuracy ranges from 62.99% (Dataset 6) to 98.92% (Dataset 2), indicating that dataset characteristics beyond size substantially influence model performance.

Statistical analysis reveals no significant correlation between dataset size and accuracy (Pearson $r = 0.12$, $p = 0.78$). Random Forest achieved the highest average performance (88.48%), followed by Decision Tree (85.37%) and Gaussian Naïve Bayes (83.79%).

4.2.2 Precision analysis

Precision metrics demonstrate similar patterns to accuracy, with performance variations that do not align with dataset size expectations (Table 5).

Decision Tree models achieved the highest average precision (86.95%), closely followed by Random Forest (86.45%). The notably poor performance on Dataset 6 across all models suggests inherent data quality issues rather than size-related limitations

4.1.3 Comprehensive performance analysis

Tables 6, 7, and 8 present the complete results for recall, F1-score, and ROC-AUC metrics, respectively. These results consistently support the finding that dataset size alone does not determine classification performance.

4.1 Analysis of dataset subdivision results

The systematic analysis of subdivided datasets provides direct evidence regarding the relationship between dataset size and performance within identical problem contexts (Table 9).

The subdivision analysis reveals non-monotonic relationships between dataset size and performance. For Dataset 6 subdivisions, accuracy varied irregularly: 60.42% (500 samples), 61.49% (1000 samples), and 59.75% (2000 samples). Similarly, Dataset 7 subdivisions showed: 80.07% (500 samples), 78.43% (1000 samples), and 81.75% (2000 samples). Statistical analysis confirms no significant correlation between subset size and performance (Pearson $r = -0.08$, $p = 0.85$ for Dataset 6; $r = 0.34$, $p = 0.73$ for Dataset 7).

Table 4: Accuracy results for complete datasets (%).

Datatset	Model							
	RF	AdaBoost	LR	SVC	GNB	DT	KNN	Average
1	90.00	60.00	100.00	85.00	85.00	90.00	85.00	85.00
2	99.00	99.00	99.49	99.00	98.50	99.00	98.50	98.92
3	83.87	81.04	77.83	81.25	81.04	83.61	78.05	80.95
4	89.41	52.55	81.91	57.18	88.05	84.98	72.01	75.15
5	99.54	19.54	82.04	85.22	98.63	97.95	88.18	81.58
6	68.44	63.10	62.95	62.95	64.17	62.19	57.16	62.99
7	89.12	78.87	70.75	89.25	71.12	79.87	88.00	80.99
Average	88.48	64.87	82.14	79.98	83.79	85.37	80.99	

Table 5: Precision results for complete datasets.

Datatset	Model							
	RF	AdaBoost	LR	SVC	GNB	DT	KNN	Average
1	94.16	54.09	100.00	85.00	81.25	95.00	93.33	86.11
2	99.35	99.35	99.35	100.00	98.75	99.35	99.33	99.35
3	68.98	86.35	74.26	76.52	80.06	85.83	72.35	77.76
4	88.08	74.16	78.27	57.18	86.95	87.88	71.23	77.67
5	99.57	14.56	83.07	85.42	98.77	98.05	89.11	81.22
6	67.35	60.11	39.63	39.63	61.57	62.52	54.73	55.07
7	87.68	79.00	69.34	87.71	70.14	80.06	86.66	80.08
Average	86.45	66.80	77.70	75.92	82.49	86.95	80.96	

Table 6: Racall metric for complete datasets.

Datatset	Model							
	RF	AdaBoost	LR	SVC	GNB	DT	KNN	Average
1	90.00	60.00	100.00	85.00	85.00	90.00	85.00	85.00
2	99.33	99.33	100.00	98.66	99.33	99.33	98.66	99.23
3	71.42	67.60	74.16	79.75	76.06	73.83	74.61	73.91
4	89.41	52.55	81.91	62.79	88.05	84.98	72.01	75.95
5	99.54	19.54	82.04	85.22	98.63	97.95	88.18	81.58
6	68.44	63.10	62.95	62.95	64.17	62.19	57.16	62.99
7	90.58	79.38	72.51	90.83	71.57	77.60	89.31	81.68
Average	86.96	63.07	81.94	80.74	83.26	83.70	80.70	

Table 7: F1 -score metric for complete datasets.

Datatset	Model							
	RF	AdaBoost	LR	SVC	GNB	DT	KNN	Average
1	88.83	53.42	100.00	81.94	82.00	89.16	85.14	82.92
2	99.33	99.33	100.00	98.66	99.33	99.33	98.66	99.23
3	71.42	67.60	74.16	79.75	76.06	73.83	74.61	73.91
4	88.30	52.42	79.75	59.02	86.42	86.18	71.50	74.79
5	99.53	15.20	81.91	84.71	98.60	97.94	87.98	80.83
6	65.87	58.98	48.64	48.64	59.58	62.34	55.42	57.06
7	89.11	78.68	70.89	89.25	70.94	97.11	87.96	83.42
Average	86.05	60.80	79.33	77.42	81.84	86.55	80.18	

Table 8: ROC-AUC metric for complete datasets

Datatset	Model							
	RF	AdaBoost	LR	SVC	GNB	DT	KNN	Average
1	100.00	82.28	100.00	98.27	100.00	94.44	97.22	96.03
2	100.00	98.66	100.00	100	100.00	98.66	99.90	99.60
3	89.02	93.33	92.85	91.28	92.56	81.79	85.30	89.44
4	93.57	89.88	88.69	87.92	92.50	84.43	84.69	88.81
5	99.99	69.39	98.91	99.62	99.98	98.88	98.28	95.00
6	61.72	55.03	50.00	50.00	55.71	59.89	51.32	54.81
7	89.15	78.88	70.78	89.27	71.13	79.83	88.02	81.00
Average	90.49	81.06	85.89	88.05	87.41	85.41	86.39	

Table 9: Accuracy for the two sub-datasets

Datatset	Model							
	RF	AdaBoost	LR	SVC	GNB	DT	KNN	Average
6	68.44	63.10	62.95	62.95	64.17	62.19	57.16	62,99
6_1	65.85	60.36	58.73	58.73	63.82	58.13	57.31	60,42
6_2	63.41	61.58	63.41	61.58	66.46	58.53	55.48	61,49
6_3	66.15	55.79	60.67	60.06	60.67	60.97	53.96	59,75
7	89.12	78.87	70.75	89.25	71.12	79.87	88.00	80,99
7_1	88.50	76.50	70.00	88.83	71.33	77.50	87.83	80,07
7_2	84.50	74.00	71.00	87.00	73.00	79.00	80.50	78,43
7_3	87.50	77.25	79.25	87.75	75.00	79.50	86.00	81,75

4.2 Statistical validation of findings

Comprehensive statistical analysis supports the conclusion that dataset size does not significantly impact classification performance in the studied agricultural applications:

1. Correlation analysis: Pearson correlation coefficients between dataset size and average performance across all metrics showed no significant relationships (all p-values > 0.05).

2. ANOVA results: Analysis of variance comparing performance across size groups revealed no significant differences ($F = 1.23$, $p = 0.34$ for accuracy).

3. Mann-Whitney U tests: Non-parametric comparisons between small (≤500 samples) and large (≥2000 samples) datasets showed no significant performance differences ($U = 42$, $p = 0.67$).

4.4 Discussion of key findings

4.4.1 Dataset quality vs. dataset size

The results strongly suggest that dataset quality characteristics have a more substantial impact on model performance than dataset size. Dataset 6, despite being among the largest (3,276 samples), consistently showed the poorest performance across all models and metrics. This pattern indicates potential data quality issues such as:

1. High noise levels in feature measurements.
2.Inadequate feature representation of the underlying problem
3. Class overlap or ambiguous classification boundaries
4. Measurement errors or inconsistent data collection procedures.

Conversely, Dataset 2, with only 201 samples, achieved the highest average performance (98.92% accuracy), demonstrating that well-curated, high-quality small datasets can outperform larger, lower quality datasets.

4.4.2 Model-specific performance patterns

Random Forest and Decision Tree models consistently achieved superior performance across datasets, suggesting that tree-based algorithms are particularly well-suited for agricultural classification tasks. This may be attributed to:

Interpretability: Tree-based models provide clear decision rules that align with agricultural domain knowledge.

Feature interaction handling: Ability to capture complex interactions between environmental and agricultural variables.

Robustness to noise: Tree-based methods are relatively resilient to noisy features common in agricultural datasets. AdaBoost showed the most variable performance,

suggesting sensitivity to dataset characteristics and potential overfitting issues with small or noisy datasets.

4.4.3 Implication for precision agriculture

Our findings suggest some important practical considerations for farmers and agricultural technology developers:

Investment: In case of limited resources, there is a need to focus on improving data quality rather than just collecting more data. Better sensors and careful data collection practices will likely give more improvement than simply gathering larger amounts of lower-quality information.

Choosing the right algorithm: Tree-based methods like Random Forest and Decision Trees seem particularly well-suited for agricultural problems. They're not only accurate but also provide interpretable results that farmers can understand and trust.

Data collection strategy: Instead of prioritizing volume, concentrate on sensor accuracy and thorough data preprocessing. Clean, well-processed data from fewer sources often outperforms noisy data from many sources.

4.5 How our results compare to previous research

Our findings both support and extend conclusions from earlier studies:

Agreement with medical domain research: Althnian et al. [12] found that how data is distributed matters more than dataset size in medical applications. Our agricultural study reaches the same conclusion, showing this principle applies across different domains.

Challenging common assumptions: Our agricultural-specific analysis shows the relationship between quantity of data and better performance isn't always straightforward. Sometimes quality trumps quantity.

Building on cross-domain evidence: Studies in software engineering [13] and biomedical fields [15] have reported similar patterns. This suggests that the size-performance relationship might depend heavily on the specific domain and problem characteristics.

5 Limitations and future work

Several limitations should be mentioned in this study: 1. **Dataset Scope**: Analysis was limited to seven agricultural datasets; broader dataset inclusion could strengthen generalizability.

2. **Feature engineering**: Limited exploration of advanced feature engineering techniques that might influence the size-performance relationship.

3. **Temporal aspects:** Static analysis did not consider temporal variations in agricultural data that might affect model performance.

4. **Cross-domain validation**: Results are specific to precision agriculture and may not generalize to other domains.

Future research directions include:

1. Investigation of deep learning model responses to dataset size variations
2. Analysis of temporal dataset characteristics in agricultural applications
3. Development of data quality metrics specific to agricultural machine learning
4. Cross-domain validation of size-performance relationships

5 Conclusion

Precision agriculture aims to manage agricultural cycle, improving resources and crop yield. Machine learning had an important role in this implementation since it analyzes collected data from different sources (sensors, satellites, etc), and gives predictions for different agricultural activities. However, the collection of data is not a simple task. So, this work focused on the impact of the size of dataset on the performance of seven machine learning models, namely: Decision Tree, Random Forest, Logistic Regression, SVM, Gaussian Naïve Bayes, KNN and AdaBoost. The empirical study shows that there is no relation between the size of the dataset and the performance of machine learning models. The performance is related to the type of problems and the dataset itself.

References

[1] T. Ayoub Shaikh, T. Rasool, and F. Rasheed Lone, “Towards leveraging the role of machine learning and artificial intelligence in precision agriculture and smart farming,” *Comput. Electron. Agric.*, vol. 198, no. June 2021, p. 107119, 2022, doi: 10.1016/j.compag.2022.107119.

[2] S. Condran, M. Bewong, M. Z. Islam, L. Maphosa, and L. Zheng, “Machine Learning in Precision Agriculture: A Survey on Trends, Applications and Evaluations over Two Decades,” *IEEE Access*, vol. 10, no. June, pp. 73786–73803, 2022, doi: 10.1109/ACCESS.2022.3188649.

[3] A. Sen, R. Roy, and S. R. Dash, “Smart Farming Using Machine Learning and IoT,” *Agric. Informatics Autom. Using IoT Mach. Learn.*, vol. 3, no. March, pp. 13–34, 2021, doi: 10.1002/9781119769231.ch2.

[4] E. M. B. M. Karunathilake, A. T. Le, S. Heo, Y. S. Chung, and S. Mansoor, “The Path to Smart Farming: Innovations and Opportunities in Precision Agriculture,” *Agric.*, vol. 13, no. 8, pp. 1–26, 2023, doi: 10.3390/agriculture13081593.

[5] D. Radočaj, M. Jurišić, and M. Gašparović, “The Role of Remote Sensing Data and Methods in a Modern Approach to Fertilization in Precision Agriculture,” *Remote Sens.*, vol. 14, no. 3, 2022, doi: 10.3390/rs14030778.

[6] P. K. Singh and A. Sharma, “An intelligent WSN-

UAV-based IoT framework for precision agriculture application," *Comput. Electr. Eng.*, vol. 100, no. July 2021, p. 107912, 2022, doi: 10.1016/j.compeleceng.2022.107912.

[7] H. Bagha, A. Yavari, and D. Georgakopoulos, "Hybrid Sensing Platform for IoT-Based Precision Agriculture," *Futur. Internet*, vol. 14, no. 8, 2022, doi: 10.3390/fi14080233.

[8] C. R. Kagan, D. P. Arnold, D. J. Cappelleri, C. M. Keske, and K. T. Turner, "Special report: The Internet of Things for Precision Agriculture (IoT4Ag)," *Comput. Electron. Agric.*, vol. 196, no. January, 2022, doi: 10.1016/j.compag.2022.106742.

[9] A. Paullada, I. D. Raji, E. M. Bender, E. Denton, and A. Hanna, "Data and its (dis)contents: A survey of dataset development and use in machine learning research," *Patterns*, vol. 2, no. 11, p. 100336, 2021, doi: 10.1016/j.patter.2021.100336.

[10] M. Zheng, F. Wang, X. Hu, Y. Miao, H. Cao, and M. Tang, "A Method for Analyzing the Performance Impact of Imbalanced Binary Data on Machine Learning Models," *Axioms*, vol. 11, no. 11, 2022, doi: 10.3390/axioms11110607.

[11] P. Wibowo and C. Fatichah, "An in-depth performance analysis of the oversampling techniques for high-class imbalanced dataset," *Regist. J. Ilm. Teknol. Sist. Inf.*, vol. 7, no. 1, pp. 63–71, 2021, doi: 10.26594/register.v7i1.2206.

[12] A. Althnian *et al.*, "Impact of dataset size on classification performance: An empirical evaluation in the medical domain," *Appl. Sci.*, vol. 11, no. 2, pp. 1–18, 2021, doi: 10.3390/app11020796.

[13] M. A. Alshammari and M. Alshayeb, "The effect of the dataset size on the accuracy of software defect prediction models: An empirical study," *Intel. Artif.*, vol. 24, no. 68, pp. 72–88, 2021, doi: 10.4114/intartif.vol24iss68pp72-88.

[14] J. Imlawi and M. Alsharo, "Evaluating classification accuracy: The impact of resampling and dataset size," *Int. J. Bus. Inf. Syst.*, vol. 24, no. 1, pp. 91–101, 2017, doi: 10.1504/IJBIS.2017.080947.

[15] A. Bailly *et al.*, "Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models," *Comput. Methods Programs Biomed.*, vol. 213, p. 106504, 2022, doi: 10.1016/j.cmpb.2021.106504.

[16] L. S. Lin, Y. S. Lin, D. C. Li, and Y. H. Liu, "Improved learning performance for small datasets in high dimensions by new dual-net model for non-linear interpolation virtual sample generation," *Decis. Support Syst.*, vol. 172, no. April, p. 113996, 2023, doi: 10.1016/j.dss.2023.113996.

[17] K. Lechqar and M. Errais, "Crop Recommendation in the Context of Precision Agriculture," in *Advances on Intelligent Computing and Data Science*, 2023, pp. 523--532, doi: https://doi.org/10.1007/978-3-031-36258-3_46.

[18] S. García, A. Fernández, J. Luengo, and F. Herrera, "A study of statistical techniques and performance measures for genetics-based machine learning: Accuracy and interpretability," *Soft Comput.*, vol. 13, no. 10, pp. 959–977, 2009, doi: 10.1007/s00500-008-0392-y.