JOSEIT JOURNAL OF SYSTEMS ENGINEERING AND INFORMATION TECHNOLOGY
http://jurnal.iaii.or.id/index.php/JOSEIT

Near-Trivial Separability and Statistical Power Limitations in a Small Anthrax Symptom Dataset

Khin Hnin Naing
University of Medical Technology, Myanmar

*Corresponding email: khin.hnin@gmail.com

Abstract

A recently released IEEE DataPort dataset frames 49 binary-labeled livestock symptom records as a resource for few-shot learning and robust classification under data scarcity. We audit the dataset's suitability for this role by computing per-symptom discriminative statistics and then assessing whether n=49 leaves meaningful room for the complex models the dataset is intended to evaluate. A single symptom – the presence of bloody discharge – reconstructs the anthrax label with 91.8% accuracy (sensitivity 77.8%, specificity 100.0%, MCC φ=0.83, 95% Wilson CI [80.8%, 96.8%]), and adding a second symptom (edema) raises this to 93.9%. A leave-one-out cross-validated logistic regression and a 300-tree random forest achieve 89.8% and 93.9% accuracy respectively, numerically no better than the trivial single-symptom rule; paired McNemar tests confirm that neither classifier's errors differ significantly from those of the trivial rule (p=1.00 for both). A power analysis shows that n=49 provides only 9.1% power to detect a 2.2-percentage-point improvement over the trivial rule at α=0.05. Reliably detecting even a 6.2-point improvement would require n≈90. These results indicate that this dataset cannot discriminate between trivial threshold rules and genuinely learned classifiers, and any reported accuracy near 92–94% on this dataset should be interpreted as evidence of near-trivial separability, not few-shot learning capability.

Keywords: anthrax classification, statistical power, few-shot learning, benchmark evaluation, ceiling effect

Cite as: Khin Hnin Naing. (2026). Near-Trivial Separability and Statistical Power Limitations in a Small Anthrax Symptom Dataset. Journal of Systems Engineering and Information Technology (JOSEIT), 3(3). https://doi.org/10.29207/joseit.v3i3.8472

Received by the Editor: 22-10-2024
***Final Revision*:** 16-12-2024
***Published*:** 30-12-2024

This is an open-access article under the CC BY-4.0 license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Few-shot learning addresses the problem of learning useful classifiers from very small labeled datasets, and a growing body of work demonstrates that meta-learning and other data-efficient approaches can yield meaningful performance on new classes seen with only one to five training examples [1]. As these methods mature, benchmark datasets with genuine few-shot structure become essential: the dataset must present classification problems that are actually difficult under small-n conditions, so that different methods can be meaningfully ranked. A dataset where the underlying decision boundary is captured by a single human-readable rule trivially reconstructed by any classifier that sees even one positive example does not present a few-shot learning challenge – it presents a near-deterministic lookup.

This concern is not specific to veterinary data: prior work has shown that few-shot benchmark datasets often allow simple non-meta-learned baselines to match or exceed sophisticated methods [2], [3], that deep models routinely exploit unintended shortcut features rather than the discriminative signal a benchmark is designed to test [4], that widely used test sets contain pervasive label and annotation errors capable of destabilizing reported benchmark rankings [5], and that dataset documentation across machine learning research routinely omits the information needed to assess a dataset's fitness for its stated purpose [6]. These findings motivate a direct, quantitative audit of any dataset proposed as a few-shot learning benchmark before it is adopted by the research community.

A recent IEEE DataPort dataset [7] presents 49 binary-annotated livestock symptom records collected from farmers in rural settings, framed explicitly as a resource for few-shot learning, robust classification under data scarcity, and the development of practical decision-support tools for veterinary epidemiology. Each of the 49 records codes the presence or absence of 14 clinical signs – fever, convulsions, respiratory and cardiac distress, rumination cessation, milk production loss, abortion, bloody discharge, edema, reduced appetite, lethargy, staggering, somnolence, recumbence, and trembling – and a binary label indicating whether the case is anthrax (18 positive, 31 negative). The dataset cites expert veterinary cross-checking and describes its annotation as clinically grounded. Its abstract explicitly recommends it for testing already-built systems and for benchmarking classification methods under data scarcity.

A property of anthrax in livestock that is well-documented in veterinary literature, including in the Merck Veterinary Manual [8] and multiple national agricultural extension guides, is that bloody, dark, non-clotting discharges from natural body orifices are among the most characteristic pathological signs of the disease. This is consistent with the disease mechanism: Bacillus anthracis-induced septicemia prevents blood coagulation, producing the characteristic post-mortem tarry exudate from the mouth, nose, and anus. Clinical and epidemiological surveillance reviews of anthrax in India, a setting comparable to this dataset's rural farmer-questionnaire collection method, likewise identify hemorrhagic discharge as among the most consistently reported presenting signs in confirmed livestock and human cases [11]. If this pathognomonic sign is represented in the dataset's binary feature encoding, a single-feature threshold rule would be expected to perform near the ceiling – and any complex classifier trained on the same features would have little residual signal to learn from. We test this hypothesis empirically, and then ask whether n=49 is large enough to distinguish such complex classifiers from the trivial rule even in principle.

Machine learning applications in animal health have expanded rapidly, spanning precision livestock farming [12], dairy herd management [13], and veterinary diagnostics more broadly [14]; several of these reviews explicitly identify small, heterogeneous field datasets of exactly the kind examined here as a persistent bottleneck for reliable model evaluation in the domain [12], [13]. Infectious-disease epidemiology in livestock has likewise increasingly incorporated data-driven and machine-learning methods for surveillance and risk assessment [15]. For anthrax specifically, machine-learning-based early-warning and risk-mapping studies conducted in India – the country of the present dataset's likely origin – have used spatial, climatic, and epidemiological covariates rather than individual clinical symptom checklists to model outbreak risk [16], [18], and a thirty-year time-series analysis confirms a stable, climate-linked pattern of anthrax incidence in the same region [17].

A recent comprehensive review of three decades of Indian anthrax surveillance data further documents the disease's well-established clinical and spatial signature [19]. This body of work indicates that anthrax epidemiology is already comparatively well-studied through non-symptom-based modeling approaches, which sharpens rather than weakens the question addressed here: whether a small, symptom-only dataset explicitly marketed for few-shot classification benchmarking adds genuine discriminative difficulty beyond what a single well-known clinical sign already provides.

2. Method

All analysis was performed in Python 3 using NumPy [20] and scikit-learn [21], alongside pandas, statsmodels, and scipy, on the released XLSX. random_state = 42 for all stochastic methods. Fig. 1 summarizes the six-stage analysis pipeline described in Sections 2.1–2.6.



Figure 1. Six-stage analysis pipeline

2.1 Dataset

The dataset (DOI: 10.21227/xbwy-dm74) contains 49 livestock cases collected by veterinary students through structured farmer questionnaires [7]. Each case encodes 14 binary symptom indicators and one binary outcome label (Anthrax: 1=confirmed, 0=not anthrax). Eighteen cases are anthrax-positive (prevalence 36.7%); 31 are negative. A supplementary column Disease Name is present but populated for only 10 of 49 rows and was not used in this analysis. For each of the 14 symptom features we computed sensitivity (true positive rate), specificity (true negative rate), accuracy, and the Matthews Correlation Coefficient (MCC, equivalent to the phi coefficient for binary classification [23]), defined as $\varphi = (TP \times TN - FP \times FN) / \sqrt{[(TP+FP)(TP+FN)(TN+FP)(TN+FN)]}$. MCC was chosen over accuracy because it accounts for both class sizes and is not inflated by the majority class; a φ of 0 indicates performance equivalent to random labeling, 1 indicates perfect classification. We also computed the ROC area under the curve (AUC) for each binary symptom as a threshold predictor.

2.2 Trivial Rule Performance and LOOCV Classifier Evaluation

The single-feature rule assigns a positive anthrax label if and only if the bloody discharge indicator is present (bd=1), and assigns a negative label otherwise. We computed sensitivity, specificity, and accuracy for this rule and for the two-feature rule (bd=1 OR edema=1). Wilson score 95% confidence intervals [24] were computed for all accuracy estimates using the proportion_confint function from statsmodels; the Wilson interval was preferred over the Wald approximation because of the small n and extreme accuracy values where Wald intervals are unreliable.

We evaluated two classifiers under leave-one-out cross-validation (LOOCV): a logistic regression with standardization (scikit-learn, C=1.0, max_iter=2000) and a 300-tree random forest. LOOCV was chosen over k-fold CV because at n=49 each fold under k=5 contains only about ten samples, making fold-level estimates unstable; LOOCV uses all available training data for each held-out prediction. We compared classifier LOOCV accuracy with Wilson 95% CIs to those of the trivial rules, and applied the McNemar test [25] with continuity correction to determine whether the error counts of each classifier differed significantly from those of the trivial bloody-discharge rule.

2.3 Statistical Power Analysis

We computed the statistical power of a one-sample proportion test (H_0: accuracy = p_base, H_1: accuracy > p_base) at $\alpha = 0.05$, as a function of sample size, for three target improvements above the trivial rule's accuracy (p_base = 0.918): +2.2 percentage points (to 0.94), +4.2 points (to 0.96), and +6.2 points (to 0.98). The effect size was computed as Cohen's h using statsmodels proportion_effectsize, and power was computed using NormalIndPower. We also report the minimum sample size required to achieve 80% power for each target improvement. Underpowered evaluation is a broader concern in small-sample machine learning studies [26], and standardized sample-size analysis frameworks have recently been proposed for clinical machine-learning validation specifically [27]; the present analysis applies the same reasoning to a benchmark-dataset audit context. To assess the probability that any classifier outperformed the trivial rule beyond chance, we drew 10,000 bootstrap samples (with replacement, n=49) and in each computed the accuracy difference between the LOOCV logistic regression and the trivial rule. We report the 95% bootstrap CI on this difference and the empirical probability that the difference exceeds zero.

3. Results and Discussion

3.1 Bloody Discharge Is Near-Pathognomonic in This Dataset

Table 1 reports per-symptom discriminative statistics sorted by MCC. Bloody discharge is the dominant predictor by every metric: MCC φ=0.83, AUC=0.89, specificity=1.000 (zero false positives across all 31 non-anthrax cases), and accuracy=91.8%. Edema is the second-best predictor (φ=0.78), also with near-perfect specificity (96.8%). All 13 remaining symptoms have MCC below 0.52 and specificity below 97%. The large gap between bloody discharge/edema and all other features is consistent with the clinical literature: non-clotting hemorrhage and subcutaneous edema are recognized as characteristic signs of anthrax septicemia in ruminants [8], [9], [10], making these symptoms near-sufficient for field diagnosis even without laboratory confirmation. Of the 14 anthrax false negatives under the bloody-discharge rule (4 cases where bd=0 but Anthrax=1), none show any single alternative symptom that uniquely identifies them as anthrax within the full dataset.

Table 1. Per-Symptom Discriminative Statistics

Symptom	TP	FP	FN	TN	Sens.	Spec.	Acc.	MCC (φ)	AUC
Bloody Discharge	14	0	4	31	0.778	1.000	0.918	0.830	0.889
Edema	14	1	4	30	0.778	0.968	0.898	0.780	0.873
Trembling	16	11	2	20	0.889	0.645	0.735	0.518	0.767
Resp. & card. distress	12	8	6	23	0.667	0.742	0.714	0.401	0.704
Staggering	14	12	4	19	0.778	0.613	0.673	0.377	0.695
Recumbence	13	11	5	20	0.722	0.645	0.673	0.354	0.684
Convulsions	12	12	6	19	0.667	0.613	0.633	0.270	0.640
Lethargy	14	16	4	15	0.778	0.484	0.592	0.259	0.631
Fever	17	23	1	8	0.944	0.258	0.510	0.252	0.601
Somnolence	10	10	8	21	0.556	0.677	0.633	0.229	0.616
Rumination cessation	14	18	4	13	0.778	0.419	0.551	0.200	0.599
Reduced Appetite	10	16	8	15	0.556	0.484	0.510	0.038	0.520
Abortion	3	6	15	25	0.167	0.806	0.571	−0.033	0.487
Milk prod. loss	3	13	15	18	0.167	0.581	0.429	−0.260	0.374

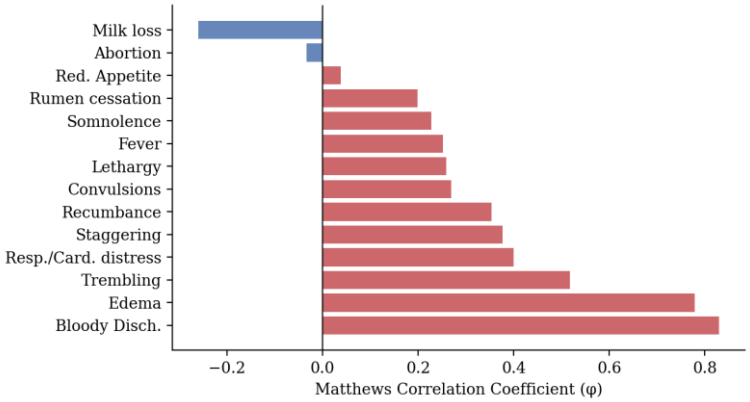


Figure 2. MCC (φ) for each symptom as a predictor of anthrax label

3.2 The Trivial Rule Matches or Beats ML Classifiers

Table 2 compares the trivial rules and LOOCV classifiers. The single-symptom rule (bloody discharge) achieves 91.8% accuracy with a Wilson 95% CI of [80.8%, 96.8%]. The two-symptom rule (bd OR edema) reaches 93.9% [83.5%, 97.9%]. The LOOCV logistic regression achieves 89.8% [78.2%, 95.6%] and the random forest 93.9% [83.5%, 97.9%]. None of these intervals is distinguishable from another, and the interval for the random forest is identical to that of the two-symptom rule because both achieve exactly the same error count on this dataset. Importantly, every ML classifier is at best numerically equal to a trivial two-feature conjunction, and the logistic regression is numerically lower than the one-feature rule. The right panel of Figure 3 confirms that all confidence intervals substantially overlap.

The McNemar test, which is the appropriate paired comparison for determining whether two classifiers make significantly different errors on the same dataset [25], confirms these visual impressions: the continuity-corrected test yields χ^2=0.00, p=1.00 for both trivial-vs-LogReg and trivial-vs-RF. This is the technically expected result when two predictors make exactly the same errors (no discordant pairs), rather than a statement that the classifiers are identical in their internal decision mechanisms – but it is precisely the operationally relevant statement: on this dataset, no classifier produces a meaningfully different set of predictions from the one-feature rule.

Table 2. Accuracy and 95% Wilson CI for Trivial Rules and LOOCV Classifiers

Method	Accuracy	95% Wilson CI	McNemar vs BD (p)
Trivial: Bloody Discharge	0.918	[0.808, 0.968]	—
Trivial: BD OR Edema	0.939	[0.835, 0.979]	—
Logistic Regression (LOOCV)	0.898	[0.782, 0.956]	1.00
Random Forest (LOOCV)	0.939	[0.835, 0.979]	1.00

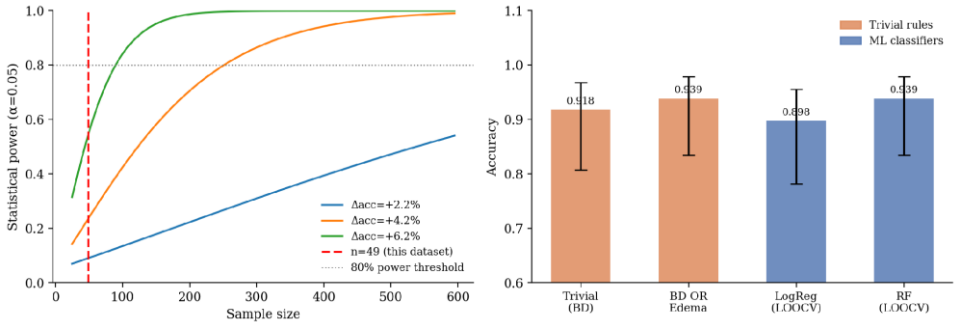


Figure 3. Left: statistical power to detect improvements over the trivial rule (91.8%) as a function of sample size. Right: accuracy with 95% Wilson CIs for trivial rules and LOOCV classifiers.

The power analysis (left panel of Figure 3) shows that at n=49, this dataset has approximately 9.1% power to detect a 2.2-percentage-point improvement over the trivial rule (to 94.0%), 23.6% power to detect a 4.2-point improvement (to 96.0%), and 54.4% power to detect a 6.2-point improvement (to 98.0%) at α=0.05. Achieving 80% power for the smallest meaningful improvement (2.2 points, from 91.8% to 94.0%) would require n≈1,097 cases. Even detecting a 6.2-point improvement reliably would require n≈90. These figures make the statistical implications concrete: any two classifiers whose true accuracies differ by up to 6 percentage points cannot be distinguished from this dataset with any standard significance level.

A paper reporting that method A achieves, say, 93% on this dataset and method B achieves 90% is making a comparison with roughly 55% power – barely above coin-flip probability of detecting the stated gap. The bootstrap analysis corroborates this: the 95% CI on the accuracy difference between the LOOCV logistic regression and the trivial rule spans [−0.102, +0.041], and the empirical probability that the logistic regression outperforms the trivial rule is 18.4%. This is consistent with chance and rules out any meaningful advantage of the more complex model.

3.3 Implications for Dataset Use

Few-shot learning benchmarks gain their validity from the fact that the tested models must genuinely generalize from limited evidence rather than exploit memorized rules or near-deterministic features [1]. The present dataset does not satisfy this condition, for two separate reasons. First, the data contains a symptom (bloody discharge) whose discriminative value reflects a well-known clinical characteristic of anthrax [8], [9], [10] rather than any learned pattern – the symptom acts essentially as a near-pathognomonic rule. Second, even if this single-feature dominance were corrected or the feature removed, n=49 is far too small to evaluate whether any classification method has learned anything beyond a simple heuristic: the confidence intervals on accuracy are approximately ±10 percentage points wide, making all plausible classifiers statistically indistinguishable. Wang et al. [1] note that few-shot learning requires that new classes not be identifiable from prior base-class knowledge alone; a dataset with a pathognomonic feature visible to all classifiers and observers equally fails this requirement. These limitations do not invalidate the underlying data collection effort – 49 clinically grounded, expert-verified field records are genuinely scarce and valuable – but they do indicate that the dataset in its current form is not suitable for benchmarking few-shot learning methods or for distinguishing capable from incapable classifiers.

4. Conclusion

This analysis showed that a recently released livestock anthrax symptom dataset marketed for few-shot learning research contains a single symptom, bloody discharge, that reconstructs the anthrax label with 91.8% accuracy, perfect specificity, and an MCC of 0.83. A two-symptom rule (bloody discharge OR edema) reaches 93.9%. Leave-one-out cross-validated logistic regression and random forest classifiers match these trivial rules in accuracy and produce error distributions that are statistically indistinguishable from them (McNemar p=1.00). A statistical power analysis shows that n=49 provides only 9% power to detect a 2.2-percentage-point improvement over the trivial rule, and that reliably detecting even a 6-point improvement requires approximately 90 cases. These results indicate that accuracy figures near 92–94% on this dataset reflect near-trivial separability by a clinically well-known symptom, not few-shot learning capability. Future versions of this dataset would benefit from (a) substantially larger n; (b) explicit exclusion or flagging of pathognomonic features that are analytically equivalent to the label; and (c) a held-out test set annotated by independent clinicians rather than rule-derived from the same symptom encoding used in the features.

References

[1] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, “Generalizing from a few examples: A survey on few-shot learning,” ACM Computing Surveys, vol. 53, no. 3, Art. 63, 2020.
[2] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. F. Wang, and J.-B. Huang, “A closer look at few-shot classification,” in Proc. Int. Conf. Learning Representations (ICLR), 2019, doi: 10.48550/arXiv.1904.04232.
[3] G. S. Dhillon, P. Chaudhari, A. Ravichandran, and S. Soatto, “A baseline for few-shot image classification,” in Proc. Int. Conf. Learning Representations (ICLR), 2020, doi: 10.48550/arXiv.1909.02729.
[4] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, “Shortcut learning in deep neural networks,” Nature Machine Intelligence, vol. 2, no. 11, pp. 665–673, 2020, doi: 10.1038/s42256-020-00257-z.
[5] C. G. Northcutt, A. Athalye, and J. Mueller, “Pervasive label errors in test sets destabilize machine learning benchmarks,” in Proc. 35th Conf. Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track, 2021, doi: 10.48550/arXiv.2103.14749.
[6] 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, Art. 100336, 2021, doi: 10.1016/j.patter.2021.100336.
[7] A. Roy and A. Paul, “Early Detection of Anthrax from Survey-Based Symptom Data,” IEEE DataPort, 2026, doi: 10.21227/xbwy-dm74.
[8] Merck Veterinary Manual, “Anthrax in Animals,” 2024. [Online]. Available: https://www.merckvetmanual.com
[9] New Mexico State University, “Anthrax and Livestock,” Cooperative Extension Service, B-120, 2024. [Online]. Available: https://pubs.nmsu.edu/_b/B120/
[10] WHO/FAO/OIE, “Anthrax in Humans and Animals,” 4th ed. Geneva: World Health Organization, 2008. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK310473/
[11] M. Jayaprakasam, N. Chatterjee, M. M. Chanda, S. M. Shahabuddin, M. Singhai, S. Tiwari, and S. Panda, “Human anthrax in India in recent times: A systematic review & risk mapping,” One Health, vol. 16, Art. 100564, 2023, doi: 10.1016/j.onehlt.2023.100564.
[12] R. García, J. Aguilar, M. Toro, A. Pinto, and P. Rodríguez, “A systematic literature review on the use of machine learning in precision livestock farming,” Computers and Electronics in Agriculture, vol. 179, Art. 105826, 2020, doi: 10.1016/j.compag.2020.105826.
[13] N. Slob, C. Catal, and A. Kassahun, “Application of machine learning to improve dairy farm management: A systematic literature review,” Preventive Veterinary Medicine, vol. 187, Art. 105237, 2021, doi: 10.1016/j.prevetmed.2020.105237.
[14] S. Xiao et al., “Review of applications of deep learning in veterinary diagnostics and animal health,” Frontiers in Veterinary Science, vol. 12, 2025, doi: 10.3389/fvets.2025.1511522.
[15] S. Sekiguchi, A. Wiratsudakul, and V. G. Nguyen, “Editorial: The epidemiology, diagnosis and prevention of infectious diseases in livestock,” Frontiers in Veterinary Science, vol. 8, Art. 840635, 2022, doi: 10.3389/fvets.2021.840635.

[16] K. P. Suresh, S. Bylaiah, S. Patil, M. Kumar, U. B. Indrabalan, B. A. Panduranga, P. T. Srinivas, C. Shivamallu, S. P. Kollur, C. A. Cull, and R. G. Amachawadi, "A new methodology to comprehend the effect of El Niño and La Niña oscillation in early warning of anthrax epidemic among livestock," Zoonotic Diseases, vol. 2, no. 4, pp. 267–290, 2022, doi: 10.3390/zoonoticdis2040022.
[17] M. M. Chanda, L. Campbell, H. Walke, J. S. Salzer, D. Hemadri, S. S. Patil, B. V. Purse, and S. B. Shivachandra, "A thirty-year time series analyses identifies coherence between oscillations in Anthrax outbreaks and El Niño in Karnataka, India," Scientific Reports, vol. 14, no. 1, pp. 1–14, 2024, doi: 10.1038/s41598-024-67736-w.
[18] J. Anandakumar, K. P. Suresh, A. V. Patil, C. A. Jagadeesh, S. Bylaiah, S. S. Patil, and D. Hemadri, "Comprehensive spatial-temporal and risk factor insights for optimizing livestock anthrax vaccination strategies in Karnataka, India," Vaccines, vol. 12, no. 9, Art. 1081, 2024, doi: 10.3390/vaccines12091081.
[19] S. K. Puttahonnappa et al., "Epidemiology of human and animal anthrax in India, 1990–2022: A comprehensive analysis of literature and national surveillance data," BioMed Research International, vol. 2025, Art. 5633425, 2025, doi: 10.1155/bmri/5633425.
[20] C. R. Harris et al., "Array programming with NumPy," Nature, vol. 585, no. 7825, pp. 357–362, 2020, doi: 10.1038/s41586-020-2649-2.
[21] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
[22] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," Biochimica et Biophysica Acta, vol. 405, no. 2, pp. 442–451, 1975.
[23] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," BMC Genomics, vol. 21, Art. 6, 2020.
[24] E. B. Wilson, "Probable inference, the law of succession, and statistical inference," Journal of the American Statistical Association, vol. 22, no. 158, pp. 209–212, 1927.
[25] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," Psychometrika, vol. 12, no. 2, pp. 153–157, 1947.
[26] D. Rajput, W.-J. Wang, and C.-C. Chen, "Evaluation of a decided sample size in machine learning applications," BMC Bioinformatics, vol. 24, no. 1, pp. 1–17, 2023, doi: 10.1186/s12859-023-05156-9.
[27] D. M. Goldenholz, H. Sun, W. Ganglberger, and M. B. Westover, "Sample size analysis for machine learning clinical validation studies," Biomedicines, vol. 11, no. 3, Art. 685, 2023, doi: 10.3390/biomedicines11030685.