PREFACE TO THE EDITION

The rapid advancement of data science, machine learning, and artificial intelligence continues to transform scientific research, technological innovation, and decision-making across diverse domains. The **Eduschool International Journal of Data Science and Machine Learning (EIJDSML)** highlights significant developments in these fields, addressing the growing need for reliable, efficient, and trustworthy data-driven systems. The research contributions presented in this publication explore uncertainty quantification, cybersecurity, healthcare, privacy-preserving learning, meteorological analysis, and time-series anomaly detection, reflecting both emerging opportunities and important methodological challenges.

Reliable machine learning requires more than accurate predictions. Research on conformal prediction examines distribution-free uncertainty estimation, prediction coverage, and the challenges arising from changing data distributions. Similarly, the investigation of intrusion detection benchmarks reveals how dataset artifacts, validation strategies, and unseen attack behaviours can influence reported performance, emphasizing the importance of rigorous evaluation and realistic testing.

The application of intelligent technologies in healthcare and privacy-sensitive environments is another important research direction. The review of machine learning and deep learning for cardiac auscultation explores automated heart sound analysis while identifying challenges in dataset standardization, evaluation consistency, and clinical validation. Complementing this work, the study of federated and privacy-preserving machine learning examines decentralized training methods, including FedAvg, FedProx, and SCAFFOLD, alongside differential privacy, secure aggregation, and homomorphic encryption, highlighting the trade-offs between privacy, model performance, and communication efficiency.

Statistical methods and reliable evaluation practices are equally important in environmental analysis and temporal data modelling. The investigation of nocturnal wind direction over India applies non-parametric circular statistical techniques to identify temporal variations in wind direction and its association with wind speed. Meanwhile, the review of time-series anomaly detection critically examines benchmark weaknesses, the limitations of point-adjustment metrics, and the competitiveness of classical baseline methods, highlighting the need for fair comparisons and transparent reporting.

Collectively, these contributions demonstrate that meaningful progress in data science and machine learning depends not only on advanced algorithms but also on methodological rigor, reliable evaluation, statistical appropriateness, privacy protection, and practical applicability. We extend our sincere appreciation to all authors for their valuable scholarly contributions, the reviewers for their constructive feedback, and the Editorial Board for its continued guidance and commitment to academic excellence. We hope this publication serves as a valuable resource for researchers, academics, practitioners, and students, encouraging further research, interdisciplinary collaboration, and responsible innovation in data science and machine learning.

Dr. Ginne M James

Chief Editor

CONTENTS

SL. NO.	TITLE	AUTHOR	PAGE NO.
1	Conformal Prediction for Distribution-Free Uncertainty in Machine Learning Models	Mini T V	1-6
2	Divergent Failure Modes of Two Intrusion Detection Benchmarks under a Common Audit	Arya K S & Selvavinayaki	7-15
3	Enhancing Cardiac Auscultation with Machine Learning and Deep Learning: A Survey on Heart Sound Analysis	Akhil K K & R Devi	16-24
4	Federated and Privacy-Preserving Machine Learning: Methods and Trade-offs	Saritha E	25-30
5	Temporal Variability of Nocturnal Wind Direction over India: A Non-Parametric Circular Statistical Approach	Anju N B & Ansa Alphonsa Antony	31-43
6	Time-Series Anomaly Detection: Benchmark Flaws, Evaluation Metrics and Strong Baselines	Tintu George	44-50

EDUSCHOOL INTERNATIONAL JOURNAL OF DATA SCIENCE AND MACHINE LEARNING (EIJDSML)

(Open Access, Double-Blind Peer Reviewed Journal)

ISSN Online: 3139-7344



Conformal Prediction for Distribution-Free Uncertainty in Machine Learning Models

Mini T V

Associate Professor, Department of Computer Science, Sacred Heart College (Autonomous), Chalakudy, India

Article information

Received: 27th June 2026
Received in revised form: 30th July 2026
Accepted: 4th September 2026
Available online: 18th October 2026

Volume: 1
Issue: 3
DOI: https://doi.org/10.63090/EIJDSML/3139.7344.0011

Abstract

A point prediction says nothing about how far it might be wrong. Conformal prediction repairs that with very little machinery: hold out some labelled data, score how badly the model misses on it, and widen each new prediction by a quantile of those scores. The resulting set contains the true label with a chosen probability, for any model and any data distribution, provided the data are exchangeable. This review reads the literature with one question in mind, namely what exactly is promised and under which assumption. We go through the split recipe and the adaptive scores built for regression and classification, then the limits of the guarantee, which is marginal and not per individual. Most of the article concerns what happens when exchangeability fails: covariate and label shift, drifting streams, time series. We close with risk control, counterfactual inference and generative language models, and with a short list of checks a practitioner should run before trusting a reported coverage figure. No new experiment is reported; every number quoted is a theoretical bound taken from the cited source.

Keywords:- Conformal Prediction, Uncertainty Quantification, Prediction Sets, Coverage Guarantee, Exchangeability, Distribution Shift, Calibration

.

I. INTRODUCTION

A credit model that outputs 0.71 has told the loan officer very little. Is 0.71 a firm estimate or a shrug? Modern networks make the question harder, because their confidence scores are known to drift away from observed frequencies: Guo and colleagues measured the gap on image and text classifiers and found that deeper, wider models were worse calibrated than older ones [1]. Uncertainty has to be attached to predictions from outside.

Several families of methods try. Deep ensembles train the same network several times and read disagreement as doubt [2]. Monte Carlo dropout keeps dropout switched on at test time and treats the spread of outputs as an approximate posterior [3]. Both give useful signals. Neither comes with a finite-sample statement about how often the truth lands inside the reported interval, and a large comparison under dataset shift found that the quality of every tested method fell as the shift grew [4]. Surveys of the area list dozens of further variants [5], [6].

Conformal prediction takes another route. It does not model uncertainty at all. It measures it, on held-out data, and converts the measurement into a set with a coverage guarantee that holds for any predictor. The idea goes back to Vovk, Gammerman and Shafer [7], [8], and the cheap split version that most practitioners now use was introduced as the inductive confidence machine [9]. Two recent overviews cover the theory at length [10],

[11]. Our aim is narrower. We ask, for each method, what it assumes and what it actually promises, since that is where applied papers most often go wrong.

II. THE SPLIT CONFORMAL RECIPE

The recipe fits in four lines. Split the labelled data in two. Fit a model on the first part. On the second part, the calibration set of n points, compute a nonconformity score for every point; for regression the usual choice is the absolute residual. Then take the quantile of those scores at level ceil((n+1)(1-a))/n and call it q. For a new input the prediction set is every label whose score is at most q, which for the residual score is the interval from f(x) minus q to f(x) plus q. Fig 1. shows the flow.

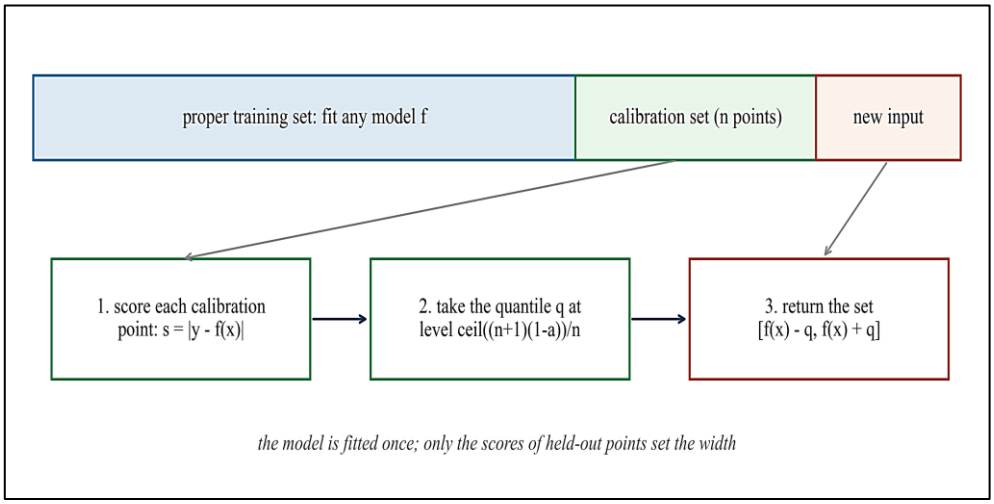

Fig. 1: Split conformal prediction for regression with the absolute residual score. Drawn by the authors from the description in [9], [12].

Why does this work? If the calibration points and the test point are exchangeable, meaning their joint distribution does not change when they are reordered, then the rank of the test score among the n+1 scores is uniform. So the test score falls below the chosen quantile with probability at least 1-a. Lei and colleagues give the statement for regression together with an upper bound: when the scores are almost surely distinct, coverage is also at most 1-a+1/(n+1), so the sets are not needlessly wide [12]. Nothing is assumed about the model. It can be a linear fit or a transformer, well specified or badly wrong. A poor model simply produces wide sets.

Splitting wastes data, and several schemes try to get it back. Cross-conformal prediction rotates the calibration fold as in cross-validation [13]. The jackknife+ uses leave-one-out residuals together with the leave-one-out predictions themselves, and its guarantee is a little weaker on paper, 1-2a instead of 1-a, although the authors report that coverage near the nominal level is what one typically sees [14]. Refitting n times is expensive, so Kim and colleagues showed that an ensemble built from bootstrap samples can supply the leave-one-out fits at no added training cost [15].

III. MAKING THE SETS ADAPTIVE

The plain residual score has an obvious flaw. Every interval has the same width. Easy inputs get intervals that are too wide and hard inputs get intervals that are too narrow, and the two errors cancel in the average. Conformalized quantile regression fixes the shape. One first fits lower and upper conditional quantiles with the pinball loss of Koenker and Bassett [16], then scores each calibration point by how far it falls outside that band, and finally shifts both ends by the conformal quantile [17]. The band keeps its input-dependent width and gains the coverage guarantee.

Classification has its own versions. The simplest score is one minus the probability the model gives to the true class. Sadinle and colleagues showed that thresholding the true class probabilities yields the smallest sets on average among rules with the same error level, and this score is the plug-in version of that rule [18]. Small on average is not the same as sensible for each input, though. Adaptive prediction sets add classes in order of decreasing probability until the cumulative mass passes a calibrated threshold, which gives larger sets on ambiguous inputs [19]. On problems with a thousand classes those sets can become unwieldy, and a regularised variant penalises classes far down the ranking to keep them short [20]. Table 1 puts the main scores side by side.

Table 1. Common nonconformity scores, as defined in the cited papers.

Task	Score	Shape of the set	Source
Regression	Absolute residual	Constant-width interval	[9], [12]
Regression	Distance outside a fitted quantile band	Input-dependent width	[17]
Classification	One minus probability of the true class	Smallest average size, if probabilities are right	[18]
Classification	Cumulative probability down to the true class	Larger on ambiguous inputs	[19]
Classification	Cumulative probability plus rank penalty	Shorter sets with many classes	[20]

One practical point is easy to miss. Calibration methods such as temperature scaling [1] or the recalibration of regression quantiles [21] change the scores the model emits, and they can make conformal sets smaller. They do not replace the conformal step. A recalibrated model still has no finite-sample coverage statement until its scores are conformalised.

IV. WHAT THE GUARANTEE DOES NOT SAY

The coverage statement is marginal. It averages over the calibration draw and over the test point. Ninety percent coverage can therefore mean ninety-nine percent for one group of patients and seventy for another. Can the method be strengthened so that coverage holds for every input? Not without assumptions. Vovk showed that exact conditional validity forces trivial sets whenever the input space is rich, a continuous one for example [22], and Barber and colleagues sharpened the result: even a relaxed, approximate form of conditional coverage cannot be had distribution-free without sets that are close to trivially wide, and they go on to ask which weaker requirements remain attainable [23].

Two weaker targets are reachable. Group-conditional coverage calibrates separately inside each predefined group [7], at the price of fewer calibration points per group. Training-conditional coverage asks that, for most calibration sets one might draw, the coverage on future points is close to nominal; split conformal has this property with a slack that shrinks as n grows [22]. In practice this is the reason a calibration set of fifty points is risky even though the marginal theorem holds for it. The realised coverage for that particular fifty can sit well away from the target.

V. WHEN EXCHANGEABILITY FAILS

Exchangeability is weaker than independence, but deployed systems still break it. Customers change. Sensors age. An earlier article in this journal catalogued the forms of drift a production model meets and the tests used to detect them [24]. What does each kind of drift do to a conformal set? Fig 2. arranges the answers as a ladder: as the assumption weakens, so does the promise.

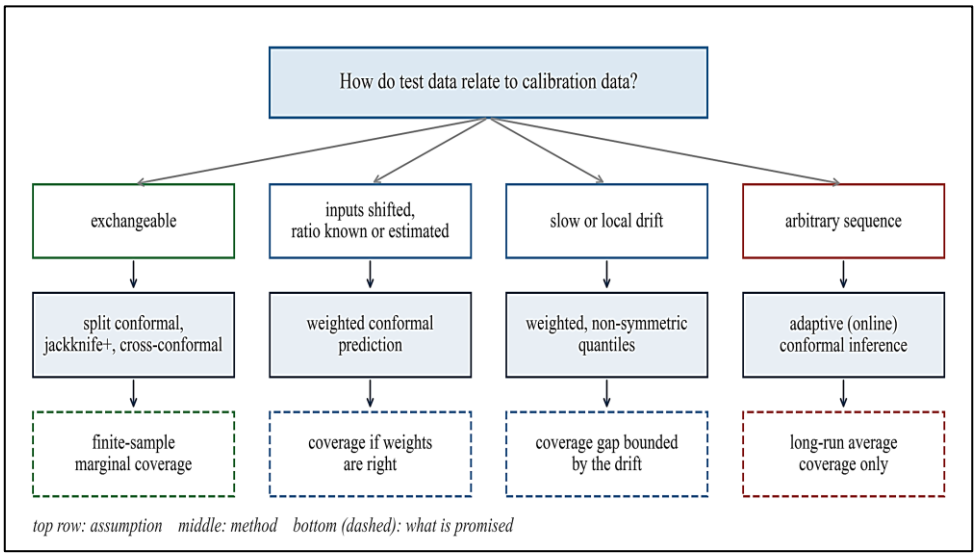

Fig. 2: Assumptions, methods and promises when test data depart from calibration data. Drawn by the authors from [25], [26], [27].

Take covariate shift first. Here the inputs move but the relation between input and label stays put. Tibshirani and colleagues reweight each calibration score by the likelihood ratio between test and training inputs and take a weighted quantile [25]. Coverage is restored exactly when the ratio is known. When it has to be estimated, the guarantee is only as good as the estimate. In high dimensions that is a real caveat. Label shift is the mirror case, with class proportions moving while class-conditional inputs stay fixed, and it has a similar weighted treatment [28].

Often nobody knows how the distribution moved. Barber and colleagues drop the symmetry requirement altogether: calibration points receive fixed weights, usually larger for recent data, and the loss of coverage is bounded by a weighted sum of total variation distances, each measuring how much the data would change if one calibration point were swapped with the test point [26]. The bound is honest about what it cannot do. If the world moved a lot, the gap is large, and the theorem says so.

Streams call for something else again. Adaptive conformal inference treats the miscoverage level as a quantity to be tuned online: after each observation it raises the level slightly if the last set covered and lowers it if the set missed [27]. The long-run fraction of misses then converges to the target for any sequence whatsoever, adversarial ones included. The price is that the statement is about a long-run average. It permits a run of bad sets followed by a run of uselessly wide ones. Zaffran and colleagues studied how the step size governs this trade-off on time series and proposed aggregating several step sizes with an online expert scheme [29].

Time series bring serial dependence even without drift. EnbPI builds intervals from bootstrap ensembles and out-of-sample residuals, avoids a held-out split, and proves approximate coverage under conditions on the error process [30]. For multi-step forecasts with recurrent networks, a split procedure over whole series with a correction across horizons gives finite-sample validity when the series themselves are exchangeable [31]. That last condition matters and is easy to overlook: it covers a panel of many patients, not one long sensor trace. Table 2 sets the settings against each other.

Table 2. Assumption and form of guarantee for methods that relax exchangeability, as stated by their authors.

Setting	Method	Guarantee	Source
Covariate shift	Likelihood-ratio weighted quantile	Marginal coverage, exact with known ratio	[25]
Label shift	Class-proportion weighted quantile	Marginal coverage under label shift	[28]
Unknown drift	Fixed weights, non-symmetric algorithm	Coverage gap bounded by total variation	[26]
Arbitrary stream	Online update of the miscoverage level	Long-run average coverage	[27], [29]
Single time series	Bootstrap ensemble residuals	Approximate, under error-process conditions	[30]
Panel of series	Split over whole series	Finite-sample, series exchangeable	[31]

VI. BEYOND COVERAGE

Coverage is one loss among many. A segmentation system may care about the fraction of a tumour that is missed, a retrieval system about recall. Risk-controlling prediction sets calibrate a threshold so that a chosen loss stays below a limit with high probability over the calibration draw [32]. Conformal risk control gives the expectation version for any bounded loss that shrinks monotonically as the set grows, and recovers ordinary conformal prediction when the loss is the miscoverage indicator [33].

Causal questions are a second extension. A counterfactual outcome is by definition never observed for the same unit, so the calibration scores come from one treatment arm and the targets from another. Lei and Candes recast this as covariate shift, with the propensity score as the weight, and obtain intervals for counterfactuals and for individual treatment effects [34]. Readers who want the surrounding framework of potential outcomes and identification will find it in a recent article of this journal [35]. The pairing is natural. An effect estimate without an interval invites exactly the overconfident decision that causal analysis is meant to prevent.

Generative language models pose the newest problem. There is no finite label set to threshold. Quach and colleagues sample candidate responses until a calibrated stopping rule fires and return the sampled set, with a guarantee that it contains at least one acceptable answer with the chosen probability [36]. This is early work. What counts as acceptable has to be defined by an automatic check, and the guarantee is relative to that check.

VII. PLACE AMONG OTHER UNCERTAINTY METHODS

It would be a mistake to read conformal prediction as a rival to ensembles or Bayesian approximations. It is a wrapper. A better base model, or a better uncertainty score from an ensemble [2], gives tighter sets after conformalisation, and the conformal step supplies the frequency statement those methods lack. The division of labour is clean: the base method decides the shape of the sets, the calibration step decides their size.

What conformal sets do not do is separate the sources of uncertainty. A wide set may come from noisy labels or from a region the model has never seen, and the distinction between aleatoric and epistemic uncertainty [6] is invisible in the set alone. Fig 3. shows how the works cited in this review divide across these themes.

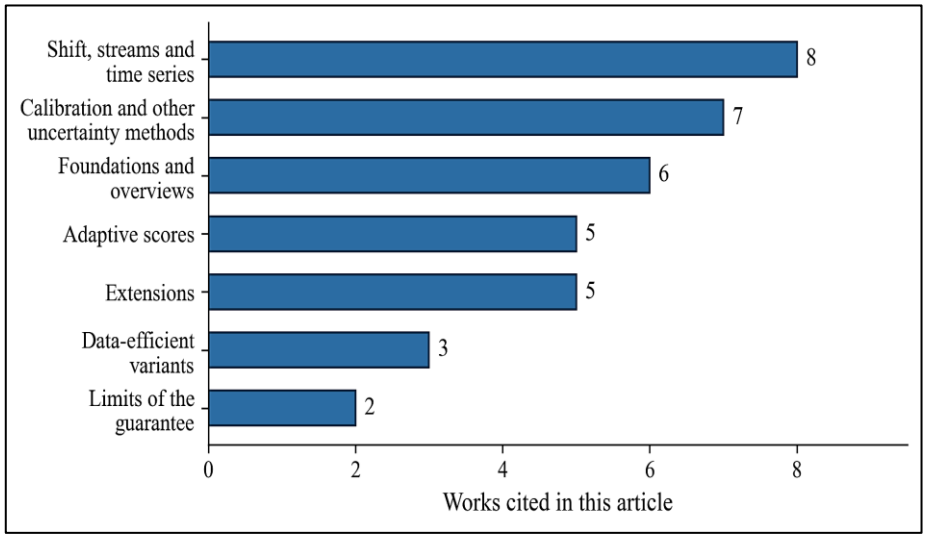

Fig. 3: The 36 works cited in this article, grouped by the role each plays in the review. The chart counts our own reference list and reports no experimental result.

VIII. CHECKS BEFORE TRUSTING A COVERAGE FIGURE

Reading applied papers with the theory in hand, we keep meeting the same omissions. First, was the calibration set truly held out? Tuning hyperparameters on it breaks exchangeability with the test point. Second, how large was it, given the training-conditional slack discussed in Section IV? Third, is coverage reported per subgroup as well as overall, given that a single marginal number hides exactly the failures that matter in medicine and lending? Fourth, is set size reported? Coverage alone can always be bought with wider sets.

A fifth check concerns time. If the data have a temporal order, a random split between calibration and test points mixes past and future and makes the exchangeable theorem look as if it applies. It does not. One of the methods in Section V should be used, and its weaker promise stated in the paper. We would add that drift monitoring and conformal calibration belong in the same pipeline, since a detected shift is the signal to recalibrate [24].

IX. CONCLUSION

Conformal prediction earns its popularity. It is cheap. It needs no change to the model, and its central theorem is short enough to check by hand. The care lies in the small print. The guarantee is marginal, it rests on exchangeability, and each relaxation of that assumption trades part of the promise away. None of this is hidden in the source papers; it tends to get lost on the way into applications.

Open problems remain on both sides. Useful approximate conditional coverage, in the relaxed senses that the impossibility results leave open [23], is still an active question. So is calibration under shift when the likelihood ratio must be learned. For generative models the field has barely started. Our practical advice is plain: report set size next to coverage, report both by subgroup, and say which assumption the reported guarantee stands on.

REFERENCES

[1] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in Proc. Int. Conf. Mach. Learn. (ICML), 2017, pp. 1321–1330.

[2] B. Lakshminarayanan, A. Pritzel, and C. Blundell, “Simple and scalable predictive uncertainty estimation using deep ensembles,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2017.

[3] Y. Gal and Z. Ghahramani, “Dropout as a Bayesian approximation: Representing model uncertainty in deep learning,” in Proc. Int. Conf. Mach. Learn. (ICML), 2016, pp. 1050–1059.

[4] Y. Ovadia et al., “Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2019.
[5] M. Abdar et al., “A review of uncertainty quantification in deep learning: Techniques, applications and challenges,” Inf. Fusion, vol. 76, pp. 243–297, 2021, doi: 10.1016/j.inffus.2021.05.008.
[6] E. Hüllermeier and W. Waegeman, “Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods,” Mach. Learn., vol. 110, no. 3, pp. 457–506, 2021, doi: 10.1007/s10994-021-05946-3.
[7] V. Vovk, A. Gammerman, and G. Shafer, Algorithmic Learning in a Random World. New York, NY, USA: Springer, 2005, doi: 10.1007/b106715.
[8] G. Shafer and V. Vovk, “A tutorial on conformal prediction,” J. Mach. Learn. Res., vol. 9, no. 12, pp. 371–421, 2008.
[9] H. Papadopoulos, K. Proedrou, V. Vovk, and A. Gammerman, “Inductive confidence machines for regression,” in Machine Learning: ECML 2002, Lecture Notes in Computer Science. Berlin, Germany: Springer, 2002, pp. 345–356, doi: 10.1007/3-540-36755-1_29.
[10] A. N. Angelopoulos and S. Bates, “Conformal prediction: A gentle introduction,” Found. Trends Mach. Learn., vol. 16, no. 4, pp. 494–591, 2023, doi: 10.1561/2200000101.
[11] M. Fontana, G. Zeni, and S. Vantini, “Conformal prediction: A unified review of theory and new challenges,” Bernoulli, vol. 29, no. 1, pp. 1–23, 2023, doi: 10.3150/21-bej1447.
[12] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman, “Distribution-free predictive inference for regression,” J. Amer. Statist. Assoc., vol. 113, no. 523, pp. 1094–1111, 2018, doi: 10.1080/01621459.2017.1307116.
[13] V. Vovk, “Cross-conformal predictors,” Ann. Math. Artif. Intell., vol. 74, nos. 1–2, pp. 9–28, 2015, doi: 10.1007/s10472-013-9368-4.
[14] R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani, “Predictive inference with the jackknife+,” Ann. Statist., vol. 49, no. 1, pp. 486–507, 2021, doi: 10.1214/20-aos1965.
[15] B. Kim, C. Xu, and R. Barber, “Predictive inference is free with the jackknife+-after-bootstrap,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2020.
[16] R. Koenker and G. Bassett, “Regression quantiles,” Econometrica, vol. 46, no. 1, p. 33, 1978, doi: 10.2307/1913643.
[17] Y. Romano, E. Patterson, and E. Candès, “Conformalized quantile regression,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2019.
[18] M. Sadinle, J. Lei, and L. Wasserman, “Least ambiguous set-valued classifiers with bounded error levels,” J. Amer. Statist. Assoc., vol. 114, no. 525, pp. 223–234, 2019, doi: 10.1080/01621459.2017.1395341.
[19] Y. Romano, M. Sesia, and E. Candès, “Classification with valid and adaptive coverage,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2020.
[20] A. N. Angelopoulos, S. Bates, M. Jordan, and J. Malik, “Uncertainty sets for image classifiers using conformal prediction,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2021.
[21] V. Kuleshov, N. Fenner, and S. Ermon, “Accurate uncertainties for deep learning using calibrated regression,” in Proc. Int. Conf. Mach. Learn. (ICML), 2018, pp. 2796–2804.
[22] V. Vovk, “Conditional validity of inductive conformal predictors,” Mach. Learn., vol. 92, nos. 2–3, pp. 349–376, 2013, doi: 10.1007/s10994-013-5355-6.
[23] R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani, “The limits of distribution-free conditional predictive inference,” Inf. Inference, vol. 10, no. 2, pp. 455–482, 2021, doi: 10.1093/imaiai/iaaa017.
[24] V. Sakhtipriyanka, “Data-centric AI and MLOps: Drift detection and model monitoring,” Eduschool Int. J. Data Sci. Mach. Learn., vol. 1, no. 2, pp. 8–14, Jul. 2026, doi: 10.63090/EIJDSML/3139.7344.0007.
[25] R. J. Tibshirani, R. F. Barber, E. Candès, and A. Ramdas, “Conformal prediction under covariate shift,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2019.
[26] R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani, “Conformal prediction beyond exchangeability,” Ann. Statist., vol. 51, no. 2, pp. 816–845, 2023, doi: 10.1214/23-aos2276.
[27] I. Gibbs and E. Candès, “Adaptive conformal inference under distribution shift,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2021.
[28] A. Podkopaev and A. Ramdas, “Distribution-free uncertainty quantification for classification under label shift,” in Proc. Conf. Uncertainty Artif. Intell. (UAI), 2021, pp. 844–853.
[29] M. Zaffran, O. Feron, Y. Goude, J. Josse, and J. Dieuleveut, “Adaptive conformal predictions for time series,” in Proc. Int. Conf. Mach. Learn. (ICML), 2022, pp. 25834–25866.
[30] C. Xu and Y. Xie, “Conformal prediction interval for dynamic time-series,” in Proc. Int. Conf. Mach. Learn. (ICML), 2021, pp. 11559–11569.
[31] K. Stankeviciute, A. M. Alaa, and M. van der Schaar, “Conformal time-series forecasting,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2021.
[32] S. Bates, A. Angelopoulos, L. Lei, J. Malik, and M. Jordan, “Distribution-free, risk-controlling prediction sets,” J. ACM, vol. 68, no. 6, pp. 1–34, 2021, doi: 10.1145/3478535.
[33] A. N. Angelopoulos, S. Bates, A. Fisch, L. Lei, and T. Schuster, “Conformal risk control,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2024.
[34] L. Lei and E. J. Candès, “Conformal inference of counterfactuals and individual treatment effects,” J. Roy. Statist. Soc. B, vol. 83, no. 5, pp. 911–938, 2021, doi: 10.1111/rssb.12445.
[35] M. Rasmi, “Causal inference and machine learning for reliable decision-making,” Eduschool Int. J. Data Sci. Mach. Learn., vol. 1, no. 2, pp. 1–7, Jul. 2026, doi: 10.63090/EIJDSML/3139.7344.0006.
[36] V. Quach et al., “Conformal language modeling,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2024.

EDUSCHOOL INTERNATIONAL JOURNAL OF DATA SCIENCE AND MACHINE LEARNING (EIJDSML)

(Open Access, Double-Blind Peer Reviewed Journal)

ISSN Online: 3139-7344



Divergent Failure Modes of Two Intrusion Detection Benchmarks under a Common Audit

Arya K S[1*], Selvavinayaki[2]

[1] Research Scholar, Department of Computer Science, Nehru Arts and Science College, Coimbatore, India
[2] Associate Professor, Department Computer Science, Nehru Arts and Science College, Coimbatore, India

Article information

Received: 9th September 2026
Received in revised form: 20th September 2026
Accepted: 1st October 2026
Available online: 18th October 2026

Volume: 1
Issue: 3
DOI: https://doi.org/10.63090/EIJDSML/3139.7344.0012

Abstract

Two public corpora carry much of the empirical load in machine learning research on intrusion detection for wireless sensor networks and for industrial internet of things deployments. Accuracy on both is routinely reported near the ceiling. We audit both with one procedure, holding features, models, metrics and seeds fixed, and they fail in different ways. On Edge-IIoTset the binary task is settled before any model is fitted. Four string valued fields record how the capture was serialised rather than how the traffic behaved, because the absent value was written as one token on benign rows and a different token on attack rows, and a depth one decision stump on any of them reaches 1.0000 where the majority baseline is 0.8417. All seven learners in the bank reach that ceiling. Removing the four fields separates them again. On WSN-DS the same audit returns almost nothing, and that is the second result. Discarding 8,873 duplicate records moves the bank average equal error rate by 0.0000, and a node disjoint split that shares no sensor with training moves it by 0.0005, against a spread of 0.0158 across the models themselves. A third check finds what neither corpus level audit does. With one attack family withheld from training, detection of the withheld behaviour runs from 0.007 to 0.886. These checks are cheap, and which of them bites is not predictable in advance.

Keywords:- Intrusion Detection, Benchmark Datasets, Data Leakage, Evaluation Protocol, Wireless Sensor Networks, Industrial Internet of Things

I. INTRODUCTION

Two corpora carry most of the empirical weight in machine learning work on intrusion detection for constrained networks. WSN-DS [1] holds 374,661 records from simulations of the LEACH routing protocol, covering four denial of service behaviours and normal traffic. Edge-IIoTset [2] is different in kind, being telemetry from a layered testbed of industrial and consumer devices, with fourteen attack types recorded alongside benign traffic. Both are public. Both load in one call. That convenience is part of why they appear so often [3], [4], [5], [6], [7], [8].

Reported accuracy on them is high. It is also reached by learners of very different capacity, and that deserves a second look. When a corpus returns the same answer to a decision stump and to a deep network, the number is describing the corpus and not the learner. The concern is old. Leakage has been documented across scientific fields [9], [10], [11], and intrusion detection carries its own long record of benchmark criticism [12], [13], [14], [15], [16]. What that literature does not tell a reader is which fault to expect where.

We ran one audit over both corpora and got two different answers. On Edge-IIoTset there is an encoding artefact in the distributed file, where four string valued fields carry the identity of the capture session rather than the behaviour of the traffic, because the absent value was serialised one way for benign rows and another way for attack rows. Any integer coding of those fields recovers the source file. On WSN-DS we expected the second familiar fault and did not find it. That corpus is normally split at record level, which puts records of the same simulated sensor on both sides and leaves duplicate rows straddling the split, so a grouped partition ought to cost something. It costs almost nothing here.

Four contributions follow. We locate the Edge-IIoTset artefact at the level of individual field values, and measure what it is worth to a bank of seven learners. We re-evaluate WSN-DS under four partitions of the same records, with features, models, metrics and seeds held fixed, and report the negative result that comes back. One attack family at a time is then withheld from training, which uncovers a weakness neither of the first two checks reaches. Last comes a short reporting protocol. Its value is that the outcome of each check cannot be guessed from the corpus alone.

II. RELATED WORK

A. Data leakage

Kaufman et al. define leakage as the introduction of information about the target that would not legitimately be available when a prediction is made [9]. Kapoor and Narayanan gathered reports of it across many fields and called it a principal driver of results that fail to reproduce [10]. Lones lists it among the recurring pitfalls of applied machine learning [11]. One feature is common to all three accounts. The fault sits in the data or in the procedure, never in the learner, so a stronger model does not repair it. Usually it makes the symptom worse.

B. Critiques of intrusion detection benchmarks

Scepticism about these benchmarks is older than either corpus studied here. McHugh questioned how the DARPA evaluations were built [13]. Tavallaee et al. showed that KDD Cup 99 carries large numbers of duplicate records, which pulls a learner towards the frequent attacks [14]. The list is long. Sommer and Paxson set out why anomaly detection is harder in operation than in the laboratory [12], and Ring et al. surveyed the corpora available at the time [15]. Gharib et al. went further and proposed criteria that a corpus ought to satisfy [16].

A more recent line of work is closer to what we do. Errors in the generation and labelling of CICIDS2017 have been traced [17], [18], measured [19] and re-analysed [20], and Flood et al. catalogue design faults that recur across benchmark corpora [21]. Catillo and colleagues argue across three papers that results obtained on public corpora do not transfer, and that published gains are smaller than they look [22], [23], [24]. Apruzzese et al. show how far a result moves when the evaluation is varied [25], while heterogeneity of features and formats obstructs comparison in its own right [26], [27]. Arp et al. gather the lessons into recommendations [28]. Neither corpus studied here appears in that work, though both are treated as standard resources [29].

C. Partition dependence in other domains

How a score depends on the split is documented well outside security. Saeb et al. showed that record level cross validation inflates clinical results relative to a subject level split [30], and Hammerla and Plotz made the same point for activity recognition [31]. General treatments exist for structured data [32], [33] and for time series [34]. Varma and Simon documented the bias that appears when one set of data both guides model selection and estimates error [35], and Allix et al. showed what ignoring the timeline costs in malware classification [36]. The warning is not new. WSN-DS has exactly the structure those papers describe, carrying a sensor identifier and a round counter, and is normally split as though it carried neither.

III. CORPORA, PARTITIONS AND MODELS

A. Corpora

WSN-DS [1] was generated with NS-2 from simulations of the LEACH clustering protocol. The distributed file holds 374,661 records, of which 8,873 are exact duplicates of another row, leaving 365,788 distinct records. Two columns are not behavioural. One is the sensor identifier, with 11,120 distinct values, and the other is the round counter, with 196. We drop both from the feature matrix and keep them as grouping variables, which leaves 16 behavioural features and five classes. Table 1 gives the counts.

Edge-IIoTset [2] ships with a subset prepared for machine learning, and that subset holds 157,800 rows. We follow the preprocessing distributed with the corpus, which removes address, timestamp and payload columns, then drop duplicate and incomplete rows. Both steps run before any split. What remains is 152,196 records over 38 features, carrying a binary label and a 15 way label, of which 24,101 are benign and 128,095 are attack. So answering attack every time already scores 0.8417 on the held out fifth. That is the baseline for Section 4.

Table 1. The two corpora after the preprocessing of Section A.

Property	WSN-DS	Edge-IIoTset
Records as distributed	374,661	157,800
Duplicate records removed	8,873	5,604
Records analysed	365,788	152,196
Behavioural features	16	38
Classes	5	15
Grouping variable	sensor id, 11,120 values	none
Ordering variable	round, 196 values	none
Majority class share	90.8 per cent	84.2 per cent

Note. WSN-DS class counts after de-duplication are Normal 332,040, Blackhole 10,049, Grayhole 13,909, Flooding 3,157, TDMA 6,633. The Edge-IIoTset majority share is the attack share of the held out fifth used throughout Section 4.

B. Partitioning protocols

The partition is the object of study here. Table 2 states all four in full. R+ is the common practice, a stratified random split over records with duplicates left in place. R repeats it after duplicates are removed. That isolates what the duplicates alone are worth. N is node disjoint, meaning no sensor identifier appears on both sides, so the detector has to generalise to a device it never saw. T is ordered in time, training on the earlier eighty per cent of rounds.

Table 2. The four partitions. Features, models, metrics and seeds are identical across all four.

Code	Split	Duplicates	What it asks
R+	stratified random over records	retained	can the model recall records like these
R	stratified random over records	removed	the same question without repeated rows
N	grouped by sensor identifier	removed	does it hold on an unseen sensor
T	rounds up to the 80th percentile train	removed	does it hold later in the simulation

C. Model bank, metrics and seeds

The model bank is wide in capacity rather than tuned. It runs from a single decision tree through random forests [37] and gradient boosted trees to a small neural network, and on Edge-IIoTset it also carries Gaussian naive Bayes, a linear support vector machine and k nearest neighbours. No hyperparameter search is performed. Winning on either corpus is not the point. A leak announces itself when models this far apart in capacity return the same number, and a genuine task announces itself when they do not.

Distance and margin based learners receive features standardised on the training partition alone. The scaler never sees a test row. Three metrics are reported. They are binary accuracy, the equal error rate and macro averaged F1 over the multiclass label, the last of these because plain accuracy hides the minority attacks [38]. Results are averaged over 5 seeds on WSN-DS and 3 on Edge-IIoTset, and overlapping standard deviations are not read as a difference [39], [40], [41]. The temporal partition is the exception. It has one realisation by construction, so only the model seed varies. The multiclass task on Edge-IIoTset runs on five learners. One against the rest over that many classes is prohibitively slow for the linear support vector machine and for k nearest neighbours.

IV. PLACEHOLDER ENCODING IN EDGE-IIOTSET

A. Single feature separability

We fitted a depth one decision tree to each feature on its own and scored it on the held out fifth. 4 come back perfect: dns.qry.name.len, mqtt.conack.flags, mqtt.protoname and mqtt.topic. Each scores 1.0000. The majority baseline is 0.8417. The best of the rest, mqtt.hdrflags, reaches 0.8734, and Figure 1(b) shows the ten strongest fields with the four marked. So one comparison on one field, learned from a fifth of the data, separates benign from attack traffic without a single error.

B. Mechanism of the artefact

The four fields are string typed in the distributed file, and the absent value is not written the same way on both sides of the label. On benign rows it appears as the token 0. On attack rows it appears as 0.0. Table 3 gives the value level contingency for mqtt.topic and Figure 1(a) plots the same counts, where the value 0 covers 22,855 benign records and no attack records, while 0.0 covers 128,095 attack records and no benign ones. Every distinct value of all four fields is pure in the label, as Table 4 records.

The explanation is prosaic. Benign and attack captures were processed as separate frames and concatenated afterwards, and the two frames inferred different types for the same empty field, so a floating point zero landed

on one side and an integer zero on the other. Nothing about the traffic is encoded. What is encoded is the file the row came from. Any label encoder, whether factorisation, ordinal coding or one hot coding, hands that straight to the model. Kaufman et al. describe the shape of the fault [9], and Flood et al. found related ones in other benchmark corpora [21].

Table 3. Value level contingency for mqtt.topic in the Edge-IIoTset machine learning subset.

Value of mqtt.topic	Benign records	Attack records
0	22,855	0
0.0	0	128,095
Temperature_and_Humidity	1,246	0

Table 4. The four fields that separate the label exactly, with the accuracy of a depth one stump on each.

Field	Distinct values	Values pure in the label	Stump accuracy
dns.qry.name.len	8	8	1.0000
mqtt.conack.flags	3	3	1.0000
mqtt.protoname	3	3	1.0000
mqtt.topic	3	3	1.0000
majority baseline			0.8417

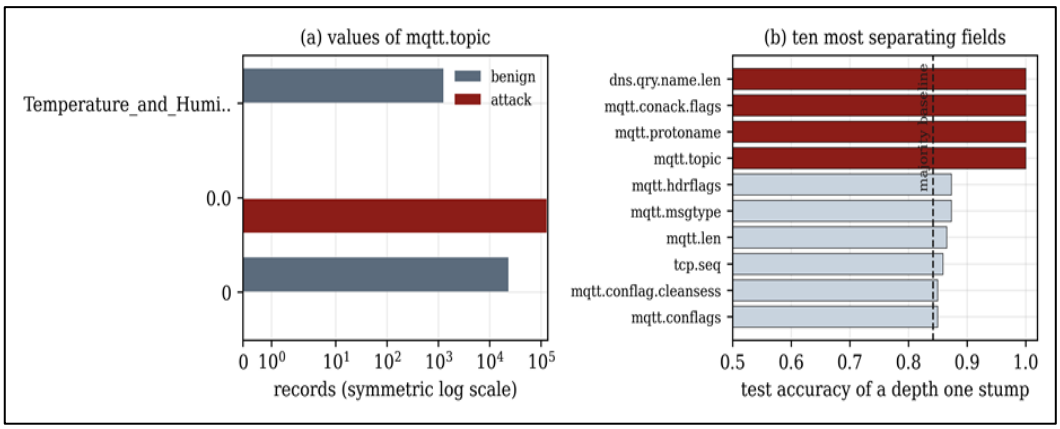

Fig. 1: (a) Records per value of mqtt.topic, benign against attack, on a symmetric log scale. (b) Test accuracy of a depth one stump on each of the ten strongest fields; the four provenance fields are shaded darker and the dashed line is the majority baseline.

C. Effect on the model bank

Table 5 and Figure 2 give the bank with the four fields kept and with them removed. Kept, 7 of the 7 learners exceed 0.9999 binary accuracy, and the spread across the bank is 0.0000. Gaussian naive Bayes is among them at 1.0000. A model assuming conditional independence and one Gaussian per feature does not reach a perfect score on a hard problem. It reaches one when a single coordinate already carries the answer.

Remove the four fields and 34 features remain. The picture changes. Gaussian naive Bayes drops to 0.8839, the weakest member of the bank is now Gaussian naive Bayes at 0.8839, and the spread widens to 0.1092. Tree ensembles stay high, which is expected and not in dispute. The ordering is the useful part. On the full feature set the bank is degenerate, in that capacity buys nothing at all; on the reduced set the models separate and the ranking carries information again.

The 15 way task behaves differently, which bounds the claim and is worth stating plainly. Macro F1 on the full feature set averages 0.819 over the 5 models evaluated on it, well short of any ceiling, and dropping the four fields moves it by at most 0.043. The four fields separate benign from attack. They do not separate one attack type from another. Published multiclass results on this corpus are therefore not compromised the way the binary ones are, and any reading of Section 4 should stop at the binary task.

Table 5. Binary accuracy on Edge-IIoTset with the four fields kept and removed. Mean and standard deviation over 3 seeds.

Model	All fields	Four fields removed	Change
Decision tree	1.0000 (0.0000)	0.9927 (0.0005)	-0.0073
Random forest	1.0000 (0.0000)	0.9927 (0.0005)	-0.0073
LightGBM	1.0000 (0.0000)	0.9931 (0.0005)	-0.0069
Gaussian naive Bayes	1.0000 (0.0000)	0.8839 (0.0029)	-0.1161
Linear SVM	1.0000 (0.0000)	0.9293 (0.0006)	-0.0707
k nearest neighbours	1.0000 (0.0000)	0.9872 (0.0003)	-0.0128
Neural network	1.0000 (0.0000)	0.9863 (0.0007)	-0.0137

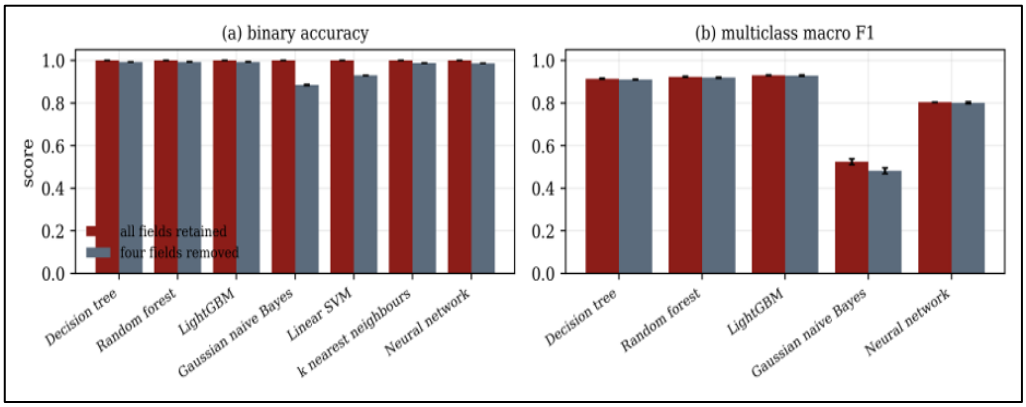

Fig. 2: The Edge-IIoTset model bank with all fields kept and with the four provenance fields removed. (a) binary accuracy, (b) macro F1 over the 15 way label for the five models evaluated on it. Error bars are one standard deviation over seeds.

V. PARTITION INSENSITIVITY IN WSN-DS

WSN-DS carries every structural feature that usually causes trouble. It has a sensor identifier and a round counter. 8,873 of its rows are exact duplicates of another row, and under a record level split 9,143 sensor identifiers land on both sides at once. On the reasoning of Section 4, and on the precedent of Saeb et al. [30] and Hammerla and Plotz [31], a grouped partition ought to cost something here. Table 6 and Figure 3 report the four partitions of Table 2 over the same records. The numbers barely move.

Discarding the duplicates moves the bank average equal error rate by 0.0000. A node disjoint split, which shares no sensor at all with training, adds a further 0.0005, and macro F1 over the five way label hardly stirs either, reading 0.929 under R+, 0.928 under R and 0.930 under N. Now compare the models. Under R+ they run from 0.0107 to 0.0265, a range of 0.0158, about 31 times the partition effect. The protocol is not what sets the number here.

The reason appears to be granularity. WSN-DS spreads its records over 11,120 sensors, about 33 on average, so no one identity carries enough of the corpus to be worth memorising, and the behaviour that defines an attack recurs across sensors rather than sitting with one. The risk is real. The effect is nil. We report it because the reasoning that predicts leakage from a shared identifier was correct in the clinical and activity recognition settings [30], [31], and a reader should know it does not always transfer.

T is the one that moves. Its macro F1 is 0.795, well below the rest. Read that carefully. The attack rate in its test partition is 0.264 against 0.092 under R+, because attacks concentrate in the later rounds, so T changes the task as well as the split. We draw no conclusion from the drop beyond a caution about ordered splits here.

Table 6. WSN-DS under the four partitions of Table 2. Equal error rate on the binary task and macro F1 on the five way task, averaged over seeds. Standard deviations appear in Figure 3.

Model	EER R+	EER R	EER N	EER T	F1 R+	F1 R	F1 N	F1 T
Decision tree	0.0136	0.0143	0.0149	0.0064	0.968	0.966	0.966	0.969
Random forest	0.0108	0.0105	0.0109	0.0005	0.978	0.977	0.978	0.979
LightGBM	0.0107	0.0104	0.0109	0.0013	0.979	0.978	0.979	0.926
Gaussian naive Bayes	0.0265	0.0271	0.0271	0.0358	0.758	0.757	0.759	0.537
Neural network	0.0127	0.0122	0.0133	0.0174	0.964	0.963	0.967	0.562
bank average	0.0149	0.0149	0.0154	0.0123	0.929	0.928	0.930	0.795

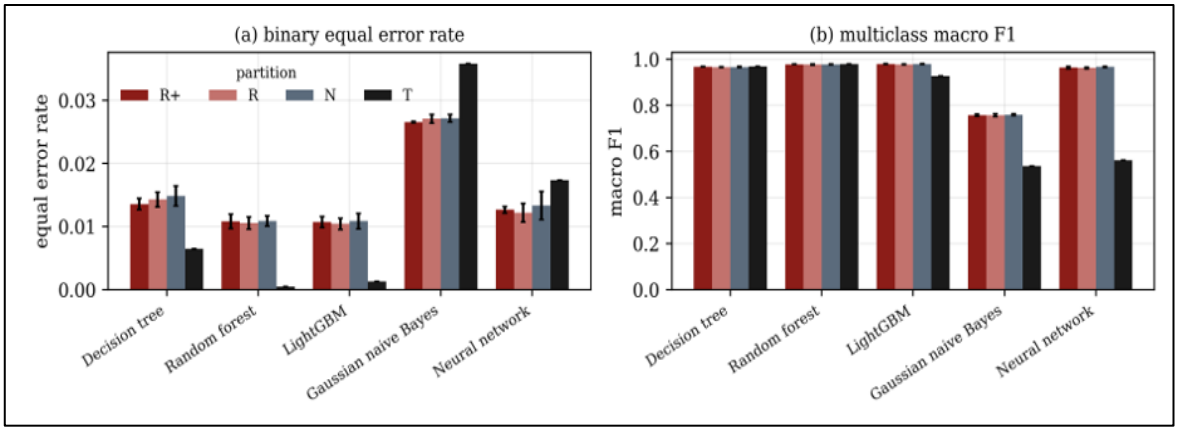

Fig. 3: WSN-DS under the four partitions of Table 2. (a) equal error rate on the binary task, (b) macro F1 on the five way task. Error bars are one standard deviation over seeds.

VI. DETECTION OF WITHHELD ATTACK FAMILIES

A deployed detector meets attacks that were never in its training set, which is the open set condition [42]. We tested it directly. One family at a time is removed from training and left in the test partition, so the detector has to flag a behaviour it has never seen, and the withholding is applied under both the random and the node disjoint partition.

Table 7 and Figure 4 give the result, and it is not the result Section 5 would lead one to expect. Detection of the withheld family varies enormously across families and barely at all across partitions. Blackhole is caught at 0.886 under the random partition and 0.873 under the node disjoint one. Flooding is caught at 0.007 and 0.007. Across all four families the largest movement between the two partitions is 0.034, while the range across families is 0.879. False positive rates on normal traffic stay under 0.002 everywhere, so the variation sits in recall and not in a shifted decision threshold.

Two readings follow, and the second bounds Section 5. A single headline number on this corpus says nothing about behaviour under a novel attack, because that number is dominated by the families the model has already seen, and a bank that looks strong in Table 6 recovers 0.007 of a withheld flooding attack. The partition is not what drives it either, which is consistent with Section 5. Neither the grouped split nor the removal of duplicates reaches this failure, and no aggregate over the five classes reveals it, so per family recall under a withheld family earns a place in the reporting on its own.

Table 7. Detection rate on an attack family withheld from training, averaged over 3 models and 3 seeds, with the false positive rate on normal traffic.

Withheld family	Random: detected	Random: false positives	Node disjoint: detected	Node disjoint: false positives
Blackhole	0.886	0.002	0.873	0.002
Grayhole	0.467	0.001	0.470	0.001
Flooding	0.007	0.001	0.007	0.001
TDMA	0.475	0.001	0.442	0.001

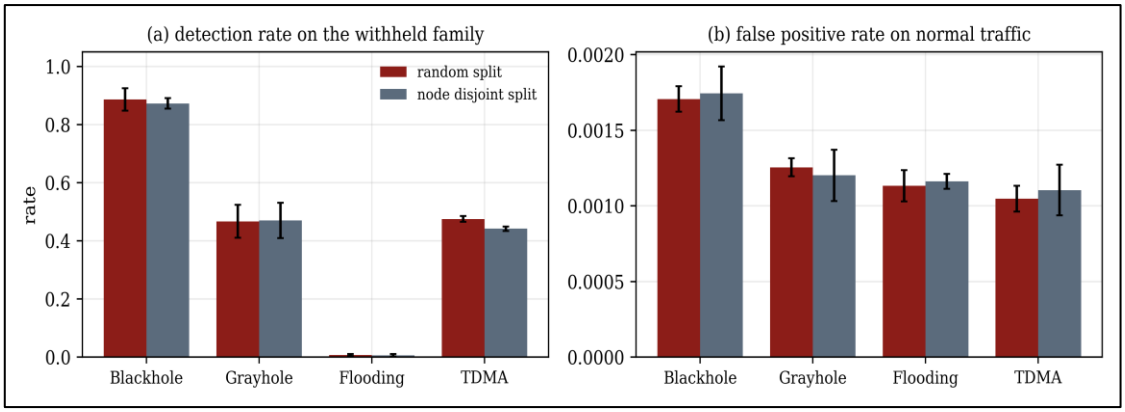

Fig. 4: Withheld attack families under the random and the node disjoint partition. (a) detection rate on the withheld family, (b) false positive rate on normal traffic. Bars are averaged over the three models.

VII. A REPORTING PROTOCOL

Cheap checks would have caught all three findings before submission. Table 8 lists them, and together they cost less than one training run of a moderate network. Most are not novel; versions appear in Arp et al. [28] and in the pitfalls Lones catalogues [11]. What Sections 4 to 6 add is a reason to run the whole list. Not the one item that looks relevant. A single feature baseline settled Edge-IIoTset in seconds and told us nothing about WSN-DS, while the grouped split did the reverse and returned a clean negative. Withholding a family found what neither saw. Which check bites is a property of the corpus.

Table 8. A reporting protocol for results on these corpora.

	Check	What it catches
1	Report the best single feature baseline	a column that already carries the label
2	Report a grouped or temporal split beside the random one	structure shared across the split
3	State whether duplicate records were removed, and when	copies spread across train and test
4	Report the majority class baseline	accuracy that beats nothing
5	Include one low capacity model in the bank	a task that is easier than claimed
6	Report per class recall, not only an average	minority attacks hidden by aggregation
7	State where preprocessing sits relative to the split	statistics fitted on test rows
8	Report spread over seeds and avoid ranking on overlap	differences that are not differences

VIII. THREATS TO VALIDITY

Two corpora are examined, not a survey. The encoding artefact of Section 4 is a property of one distributed file. It should not be generalised to other releases of Edge-IIoTset, nor to the raw captures the subset was built from, and whether it reaches the full corpus is untested here. Nor does anything above say the corpora should be abandoned. WSN-DS is the only public corpus of its kind for LEACH routing attacks, and Edge-IIoTset covers a breadth of device types that few alternatives match.

The bank is fixed and untuned. No number above is a claim about the best achievable score. One further caution concerns the identifiers, which in WSN-DS come from a simulation, so a node disjoint split over them is a weaker test than one over physical hardware, and the null result of Section 5 should be read as applying to simulated identities in the first instance.

XI. CONCLUSION

One audit, two corpora, two different answers. On Edge-IIoTset the binary task is decided by four fields that encode how the capture was serialised, and a depth one stump on any of them scores 1.0000 where the majority baseline is 0.8417. All seven learners reach that ceiling and they separate once the fields are dropped. On WSN-DS the same reasoning predicts a partition effect and the measurement does not find one, with duplicates worth 0.0000 and a node disjoint split worth 0.0005 against a model spread of 0.0158. Withholding one attack family then exposes a third failure that neither check reaches, with detection of the withheld behaviour running from 0.007 to 0.886.

The practical consequence is narrow. A binary result on Edge-IIoTset should be reported with the four fields removed, or not reported. A result on WSN-DS should carry a grouped split beside it, and here that split comes back clean, which is worth printing rather than a reason to skip it. What the three results share is that the outcome was not predictable from the corpus description.

REFERENCES

[1] Almomani, B. Al-Kasasbeh, and M. AL-Akhras, "WSN-DS: A Dataset for Intrusion Detection Systems in Wireless Sensor Networks," Journal of Sensors, vol. 2016, pp. 1-16, 2016, doi: 10.1155/2016/4731953.

[2] M. A. Ferrag, O. Friha, D. Hamouda, L. Maglaras, and H. Janicke, "Edge-IIoTset: A New Comprehensive Realistic Cyber Security Dataset of IoT and IIoT Applications for Centralized and Federated Learning," IEEE Access, vol. 10, pp. 40281-40306, 2022, doi: 10.1109/access.2022.3165809.

[3] S. Salmi and L. Oughdir, "Performance evaluation of deep learning techniques for DoS attacks detection in wireless sensor network," Journal of Big Data, vol. 10, no. 1, 2023, doi: 10.1186/s40537-023-00692-w.

[4] Rana, S. Prajapat, P. Kumar, and K. Kumar, "Performance Evaluation of Machine Learning Models for Intrusion Detection in Wireless Sensor Networks: A Case Study Using the WSN DS Dataset," in Lecture Notes in Networks and Systems, pp. 173-180, 2024, doi: 10.1007/978-981-99-8129-8_15.

[5] M. Douiba, S. Benkirane, A. Guezzaz, and M. Azrour, "An improved anomaly detection model for IoT security using decision tree and gradient boosting," The Journal of Supercomputing, vol. 79, no. 3, pp. 3392-3411, 2022, doi: 10.1007/s11227-022-04783-y.

[6] M. M. Alani, "An explainable efficient flow-based Industrial IoT intrusion detection system," Computers and Electrical Engineering, vol. 108, p. 108732, 2023, doi: 10.1016/j.compeleceng.2023.108732.
[7] S. Soliman, W. Oudah, and A. Aljuhani, "Deep learning-based intrusion detection approach for securing industrial Internet of Things," Alexandria Engineering Journal, vol. 81, pp. 371-383, 2023, doi: 10.1016/j.aej.2023.09.023.
[8] M. A. Ferrag et al., "Revolutionizing Cyber Threat Detection With Large Language Models: A Privacy-Preserving BERT-Based Lightweight Model for IoT/IIoT Devices," IEEE Access, vol. 12, pp. 23733-23750, 2024, doi: 10.1109/access.2024.3363469.
[9] S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, "Leakage in data mining," ACM Transactions on Knowledge Discovery from Data, vol. 6, no. 4, pp. 1-21, 2012, doi: 10.1145/2382577.2382579.
[10] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," Patterns, vol. 4, no. 9, p. 100804, 2023, doi: 10.1016/j.patter.2023.100804.
[11] M. A. Lones, "Avoiding common machine learning pitfalls," Patterns, vol. 5, no. 10, p. 101046, 2024, doi: 10.1016/j.patter.2024.101046.
[12] R. Sommer and V. Paxson, "Outside the Closed World: On Using Machine Learning for Network Intrusion Detection," in 2010 IEEE Symposium on Security and Privacy, pp. 305-316, 2010, doi: 10.1109/sp.2010.25.
[13] J. McHugh, "Testing Intrusion detection systems," ACM Transactions on Information and System Security, vol. 3, no. 4, pp. 262-294, 2000, doi: 10.1145/382912.382923.
[14] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, pp. 1-6, 2009, doi: 10.1109/cisda.2009.5356528.
[15] M. Ring, S. Wunderlich, D. Scheuring, D. Landes, and A. Hotho, "A survey of network-based intrusion detection data sets," Computers & Security, vol. 86, pp. 147-167, 2019, doi: 10.1016/j.cose.2019.06.005.
[16] Gharib, I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "An Evaluation Framework for Intrusion Detection Dataset," in 2016 International Conference on Information Science and Security (ICISS), pp. 1-6, 2016, doi: 10.1109/icissec.2016.7885840.
[17] Sharafaldin, A. Habibi Lashkari, and A. A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization," in Proceedings of the 4th International Conference on Information Systems Security and Privacy, pp. 108-116, 2018, doi: 10.5220/0006639801080116.
[18] G. Engelen, V. Rimmer, and W. Joosen, "Troubleshooting an Intrusion Detection Dataset: the CICIDS2017 Case Study," in 2021 IEEE Security and Privacy Workshops (SPW), pp. 7-12, 2021, doi: 10.1109/spw53761.2021.00009.
[19] L. Liu, G. Engelen, T. Lynar, D. Essam, and W. Joosen, "Error Prevalence in NIDS datasets: A Case Study on CIC-IDS-2017 and CSE-CIC-IDS-2018," in 2022 IEEE Conference on Communications and Network Security (CNS), pp. 254-262, 2022, doi: 10.1109/cns56114.2022.9947235.
[20] Rosay, E. Cheval, F. Carlier, and P. Leroux, "Network Intrusion Detection: A Comprehensive Analysis of CIC-IDS2017," in Proceedings of the 8th International Conference on Information Systems Security and Privacy, pp. 25-36, 2022, doi: 10.5220/0010774000003120.
[21] R. Flood, G. Engelen, D. Aspinall, and L. Desmet, "Bad Design Smells in Benchmark NIDS Datasets," in 2024 IEEE 9th European Symposium on Security and Privacy (EuroS&P), pp. 658-675, 2024, doi: 10.1109/eurosp60621.2024.00042.
[22] M. Catillo, A. Del Vecchio, A. Pecchia, and U. Villano, "A Critique on the Use of Machine Learning on Public Datasets for Intrusion Detection," in Communications in Computer and Information Science, pp. 253-266, 2021, doi: 10.1007/978-3-030-85347-1_19.
[23] M. Catillo, A. Del Vecchio, A. Pecchia, and U. Villano, "Transferability of machine learning models learned from public intrusion detection datasets: the CICIDS2017 case study," Software Quality Journal, vol. 30, no. 4, pp. 955-981, 2022, doi: 10.1007/s11219-022-09587-0.
[24] M. Catillo, A. Pecchia, and U. Villano, "Machine Learning on Public Intrusion Datasets: Academic Hype or Concrete Advances in NIDS?," in 2023 53rd Annual IEEE/IFIP International Conference on Dependable Systems and Networks - Supplemental Volume (DSN-S), pp. 132-136, 2023, doi: 10.1109/dsn-s58398.2023.00038.
[25] G. Apruzzese, L. Pajola, and M. Conti, "The Cross-Evaluation of Machine Learning-Based Network Intrusion Detection Systems," IEEE Transactions on Network and Service Management, vol. 19, no. 4, pp. 5152-5169, 2022, doi: 10.1109/tnsm.2022.3157344.
[26] M. Sarhan, S. Layeghy, and M. Portmann, "Towards a Standard Feature Set for Network Intrusion Detection System Datasets," Mobile Networks and Applications, vol. 27, no. 1, pp. 357-370, 2021, doi: 10.1007/s11036-021-01843-0.
[27] T. M. Booij, I. Chiscop, E. Meeuwissen, N. Moustafa, and F. T. H. D. Hartog, "ToN_IoT: The Role of Heterogeneity and the Need for Standardization of Features and Attack Types in IoT Network Intrusion Data Sets," IEEE Internet of Things Journal, vol. 9, no. 1, pp. 485-496, 2022, doi: 10.1109/jiot.2021.3085194.
[28] D. Arp et al., "Lessons Learned on Machine Learning for Computer Security," IEEE Security & Privacy, vol. 21, no. 5, pp. 72-77, 2023, doi: 10.1109/msec.2023.3287207.
[29] Z. Ahmad, A. Shahid Khan, C. Wai Shiang, J. Abdullah, and F. Ahmad, "Network intrusion detection system: A systematic study of machine learning and deep learning approaches," Transactions on Emerging Telecommunications Technologies, vol. 32, no. 1, 2020, doi: 10.1002/ett.4150.

[30] S. Saeb, L. Lonini, A. Jayaraman, D. C. Mohr, and K. P. Kording, "The need to approximate the use-case in clinical machine learning," GigaScience, vol. 6, no. 5, 2017, doi: 10.1093/gigascience/gix019.
[31] N. Y. Hammerla and T. Plötz, "Let's (not) stick together," in Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 1041-1051, 2015, doi: 10.1145/2750858.2807551.
[32] D. R. Roberts et al., "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure," Ecography, vol. 40, no. 8, pp. 913-929, 2017, doi: 10.1111/ecog.02881.
[33] L. A. Yates, Z. Aandahl, S. A. Richards, and B. W. Brook, "Cross validation for model selection: A review with examples from ecology," Ecological Monographs, vol. 93, no. 1, 2023, doi: 10.1002/ecm.1557.
[34] C. Bergmeir and J. M. Benítez, "On the use of cross-validation for time series predictor evaluation," Information Sciences, vol. 191, pp. 192-213, 2012, doi: 10.1016/j.ins.2011.12.028.
[35] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," BMC Bioinformatics, vol. 7, no. 1, 2006, doi: 10.1186/1471-2105-7-91.
[36] Allix, T. F. Bissyandé, J. Klein, and Y. Le Traon, "Are Your Training Datasets Yet Relevant?," in Lecture Notes in Computer Science, pp. 51-67, 2015, doi: 10.1007/978-3-319-15618-7_5.
[37] Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001, doi: 10.1023/a:1010933404324.
[38] 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, no. 1, 2020, doi: 10.1186/s12864-019-6413-7.
[39] T. G. Dietterich, "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms," Neural Computation, vol. 10, no. 7, pp. 1895-1923, 1998, doi: 10.1162/089976698300017197.
[40] C. Nadeau and Y. Bengio, "Inference for the Generalization Error," Machine Learning, vol. 52, no. 3, pp. 239-281, 2003, doi: 10.1023/a:1024068626366.
[41] R. R. Bouckaert and E. Frank, "Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms," in Lecture Notes in Computer Science, pp. 3-12, 2004, doi: 10.1007/978-3-540-24775-3_3.
[42] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult, "Toward Open Set Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 7, pp. 1757-1772, 2013, doi: 10.1109/tpami.2012.256.

EDUSCHOOL INTERNATIONAL JOURNAL OF DATA SCIENCE AND MACHINE LEARNING (EIJDSML)

(Open Access, Double-Blind Peer Reviewed Journal)

ISSN Online: 3139-7344



Enhancing Cardiac Auscultation with Machine Learning and Deep Learning: A Survey on Heart Sound Analysis

Akhil K K[1*], R Devi[2]

[1] Research Scholar, Department of Computer Science, School of Computer Science, Vels Institute of Science Technology & Advanced Studies, Pallavaram, India
[2] Professor & Head, Department of Applied Computing & Emerging Technologies, Vels Institute of Science Technology & Advanced Studies, Pallavaram, India

Article information

Received: 10th September 2026
Received in revised form: 1st October 2026
Accepted: 10th October 2026
Available online: 18th October 2026

Volume: 1
Issue: 3
DOI: https://doi.org/10.63090/EIJDSML/3139.7344.0013

Abstract

Cardiovascular disease kills more people than anything else. Early detection changes outcomes, and heart sounds are one of the cheapest places to look for the first signs. Phonocardiogram (PCG) recordings carry information about valve function that a trained ear can pick up, which is why auscultation has survived two centuries of newer imaging and why it is still the first test a patient meets. But listening is slow. It needs experience, and two clinicians can disagree about the same murmur. Machine learning (ML) and deep learning (DL) offer a way to automate that step and make it reproducible. This review asks what the literature has actually achieved, and what it has not. We searched IEEE Xplore, SpringerLink, PubMed, ScienceDirect and Google Scholar for work published between 2015 and 2024. Section 2 gives the search strings, the screening steps and the inclusion rules. Thirty-seven primary studies survived. Each was read in full, and its dataset, signal processing chain, model and reported performance were recorded in a comparison table that covers thirteen of them in detail. Reported accuracy in this field is high and rising. It is also, taken at face value, misleading. The studies reviewed here use different corpora, different class definitions, different segmentation strategies and different validation protocols, so 99.80% in one paper and 73.6% in another are not measuring the same thing. Standardised datasets are missing. So are agreed metrics. Paediatric recordings are thin on the ground, recording conditions vary, and almost nothing here has been tested in the clinic where it would have to work. The review pulls recent advances together, then sets the reported numbers against the methodological limits that decide how much those numbers are worth, which surveys in this area tend to skip.

Keywords:- Cardiovascular diseases, Machine Learning, Deep learning, Heart sound classification, Phonocardiogram

I. INTRODUCTION

Heart disease remains the largest single cause of death worldwide. Much of it announces itself in sound. When a valve leaks or narrows, blood moving through it turns turbulent, and that turbulence reaches the chest wall as a murmur [1] [6]. Clinicians have used murmurs as a first indicator of valve trouble for as long as the stethoscope has existed [3] [8] [38]. Grading them is another matter. Quality, timing and intensity are all judged by ear, and that judgement is itself treated as a diagnostic feature in cardiovascular assessment [2] [7].

Auscultation survives because it is cheap, fast and needs no more equipment than a stethoscope. It is also hard. Heart sounds sit low in frequency, they are quiet, and they compete with breathing, gut noise and whatever else the patient's body is doing at the time [4]. Accuracy tracks the listener's experience closely. The examiner's skill with the instrument, the room, and the character of the sound itself all shape what ends up written in the notes [9] [10].

Those limits are what pushed AI into the problem. A murmur can point to valvular disease, to a congenital defect, or to heart failure, and telling those apart by ear is a skill that takes years to build and stays uneven even then. ML and DL models can be trained to make the same call from a raw recording. They make it in seconds [5].

None of this is free. These methods depend on how the audio was captured, on what the signal processing chain does to it before the model ever sees it, and on having enough well-labelled recordings to learn from. Picking a model that separates normal from abnormal reliably is a further problem, and not a solved one.

This review works through recent progress in heart sound analysis and asks how far ML and DL have moved past the limits of manual auscultation. Signal processing, datasets and classification methods each get a section. The remaining challenges are named. So are the directions that look most likely to help, and the limits on what the reported numbers can support.

Section 2 covers the search. Section 3 is the body of the review, and Section 4 collects the challenges and where they point. Section 5 closes.

II. SEARCH STRATEGY

A. Research questions

Four questions shaped the search and, later, the way the findings are organised.

- What are the systems used for capturing heart sounds?
- What are the signal processing techniques used for analyzing heart sounds?
- Which datasets are commonly used for training machine learning models in heart disease and murmur prediction?
- What classification techniques are applied to heart sound data for detecting murmurs and predicting heart diseases?

The four questions do double duty. They set the boundary of what was searched for, and in Section 3.5 they come back as the frame for what the corpus actually says.

B. Data Collection Process

The search covered IEEE Xplore, PubMed, ScienceDirect, SpringerLink and Google Scholar, filtered to 2015-2024. Four query strings were used: ("heart sound" AND classification AND "machine learning"); ("heart murmur" AND detection AND ("artificial intelligence" OR "deep learning")); ("heart disease" AND prediction AND dataset); and ("machine learning" AND "heart disease"). Titles and abstracts came first. Full texts followed for anything that survived, and the reference lists of the retained papers were then checked by hand for work the strings had missed.

C. Screening and Selection

Thirty-seven primary studies came through screening. Each was checked against the inclusion rules, with weight given to whether the method was described clearly enough to be judged, which dataset was used, and what was actually measured. A few further references were kept for the benchmark datasets and for background, not as primary studies. Three of the retained items are preprints ([16], [18] and [34]). They document systems or corpora the field already uses, which is why they are here, but their results have not been through peer review. One more thing about scope. Per-database yield counts are not reported, and this is a structured narrative review rather than a PRISMA-registered systematic review.

D. Inclusion & Exclusion Criteria

Inclusion criteria:

- Articles published from 2015 to 2024 on heart disease and murmur prediction using ML and DL.
- Studies with empirical records or theoretical models focused on heart sound analysis and machine learning applications.
- Papers addressing key challenges like signal processing, classification accuracy, and real-time prediction.
- Peer-reviewed, open-access research papers, plus the preprints named in Section 2.3.
- Exclusion Criteria:

- Non-peer-reviewed studies, other than the preprints named in Section 2.3, and papers unrelated to heart disease or murmur prediction.
- Papers lacking sufficient methodological detail or those focusing on unrelated topics.

III. ADVANCEMENTS IN HEART SOUND ANALYSIS: ML, DL AND EMERGING TRENDS

Demand for cardiac screening is rising faster than the supply of people trained to do it. That gap is the case for automating the listening step. Murmur sounds carry enough information to flag disease, and ML and DL methods have become the standard way of pulling that information out of a recording [11]. The analysis has sharpened. Fig. 1 shows the workflow most of this work follows, from capture through preprocessing and feature extraction to a classifier, and the subsections below take it stage by stage.

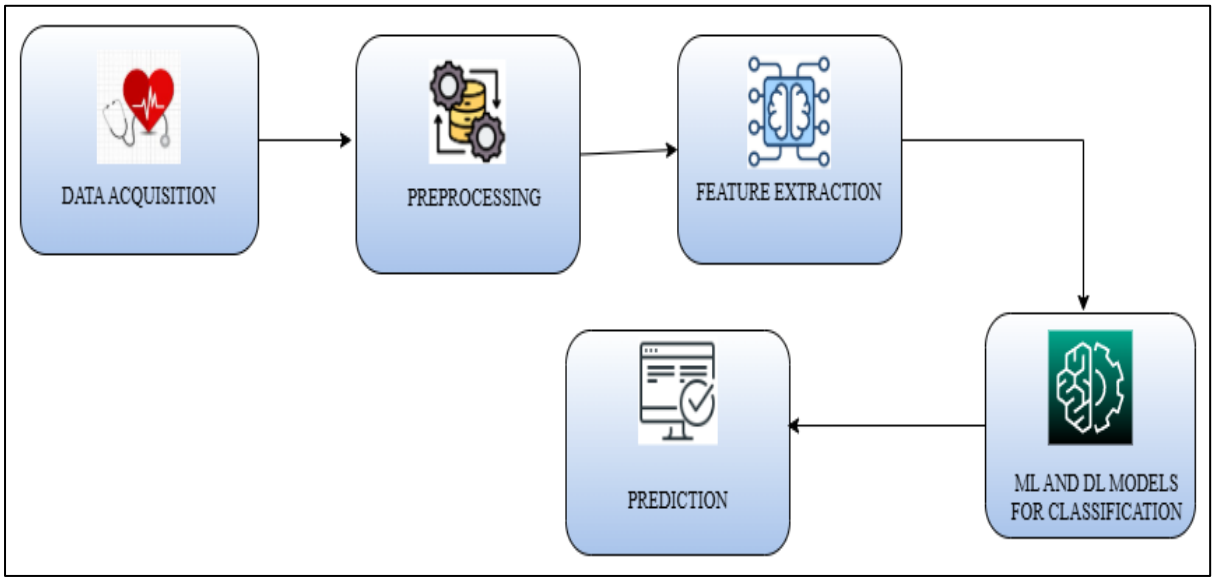

Fig 1: Heart Sound Analysis using ML and DL Models

A. DATA ACQUISITION

1. Approaches for Capturing Heart Sounds

Capture hardware varies more than you would expect. The cheapest systems start from a stethoscope and a Bluetooth link. In [33] a stethoscope bell is fixed to the microphone of a dismantled Bluetooth headset, and [32] builds a Bluetooth sensing unit that passes recordings on for analysis. The system in [13] instead puts ECG and PCG amplification on a custom CMOS chip and shows both signals on a smartphone. The rig in [3] is heavier. Electrical detectors, analogue conditioning circuitry, a data acquisition card, a host computer and software for visualising the signal all sit in the chain.

The design in [15] is an electronic stethoscope assembled from parts that could be bought separately: a microphone, a preamplifier, a Butterworth filter, an AVR ATmega-32 microcontroller, a power amplifier and an LCD. Nothing exotic. A different approach appears in [4], where the sensor is worn as a finger ring and collects heart sounds alongside photoplethysmography (PPG).

Digital stethoscopes are the usual source of PCG recordings [16] [17]. Study [18] went after model reliability rather than raw accuracy, pairing Monte Carlo dropout for uncertainty estimation with temperature scaling for calibration. In [19], neuromorphic sensors that mimic the cochlea were mapped onto FPGAs so the processing could keep up with the signal in real time. Hardware is only half of it. The BUET multi-disease heart sound dataset [34] and the CirCor DigiScope dataset [35] both widened the range of recordings the field can train on, and public corpora of that kind have done as much for progress here as any sensor design.

2. Datasets Utilized

Fig. 2 counts how often each dataset appears across the reviewed studies. PhysioNet/CinC Challenge 2016 dominates with nine uses. CirCor DigiScope and the 2022 Challenge account for three each, MITHSDB for two, and BMD-HS, the paediatric corpus, Yaseen [36], Kaggle Heartbeat Sounds and PASCAL for one apiece. Two cautions apply to that tally. The categories overlap, because CirCor DigiScope is the corpus released for the 2022 Challenge and MITHSDB is one of the databases folded into the 2016 set, so the bars are not mutually exclusive. The Yaseen count also reads low against Table 1, where two of the tabulated studies use it.

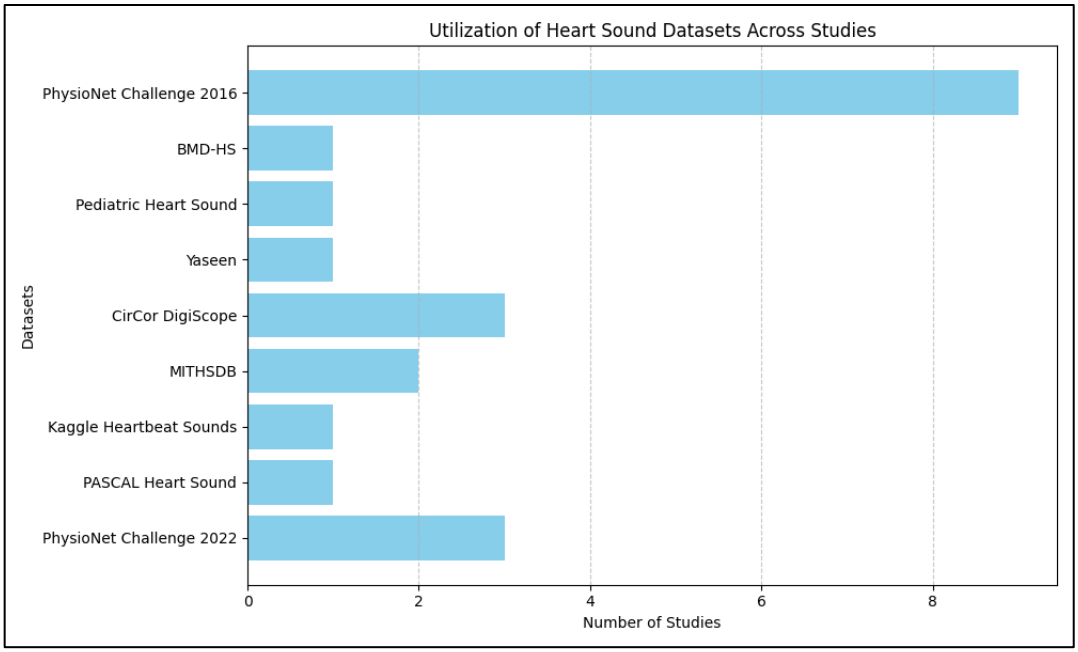

Fig 2: Dataset use across the reviewed studies

B. Preprocessing

Preprocessing decides how much signal survives to the model. Several chains have been built for it, and the differences between them matter more than their length suggests. The work in [1] [13] runs ECG and PCG through dual-channel analogue front ends, with amplifiers and a high-pass sigma-delta modulator doing the conditioning. Noise drops. The two streams stay synchronised, which is what makes a combined reading worth trusting.

A Bluetooth-linked acquisition unit appears in [32] [33]. Denoising there is handled by the Hilbert-Huang transform, while a double-threshold segmentation algorithm marks the physiological events, which leaves a cleaner trace with the cardiac events already located in it.

Shannon entropy has been used to find heart sound boundaries and photoplethysmography peaks [4] [8]. The intelligent stethoscope in [21] locks its ECG channel to the heart sound channel, and that synchronisation is what lets the two be read together rather than side by side. Small engineering choices carry weight here.

C. Feature Extraction

Feature extraction sets the ceiling. Whatever the features fail to encode, no downstream model recovers, which makes this stage worth more attention than the architecture choices that follow it. The Hilbert-Huang transform, mentioned above for denoising, doubles as a feature extractor, with double-threshold segmentation supporting the physiological analysis [32].

Continuous and discrete wavelet transforms carry the load in [16] and [7], where the problem is picking murmurs out of a noisy band. In [22] a denoising autoencoder learns deep features that feed a one-dimensional CNN; those features held up better under interference than Mel-frequency cepstral coefficients (MFCCs), and the improved network reached 99.01% accuracy. Learned features can beat engineered ones.

Decomposition methods do similar work elsewhere. Harmonic-percussive source separation, grey-level co-occurrence matrices and wavelet transforms all appear in [23] and [24]. Studies [25] and [26] convert PCG signals into two-dimensional images, chaogram and scalogram respectively, so that image models can be brought to bear. RNNs and hidden semi-Markov models split the signal into S1, S2, systole and diastole in [27]. The Temporal-Framing Adaptive Network in [29] does the same job without any prior assumption about how long each state lasts. Trained on only 50 recordings from the 2016 Challenge and tested on 2,099 more from twelve other databases, it reached an average F1 of 96.72%, ahead of both the logistic-regression HSMM and a bidirectional gated recurrent network. MFCCs go in directly as classifier inputs in [28], while [20] computes an improved MFCC variant and passes it to a deep residual network, reporting 94.43% accuracy. The scheme in [39] is the odd one out: short-term features computed over five-second windows are merged with long-term features taken from the whole recording, and across the twenty-two algorithms tested the merge lifted accuracy by 3 to 10%, with the best ensemble reaching 92.7%.

D. Classification Using Ml Models

CNNs do most of the work in this literature. The architecture in [21] separates heart sound signals by reading the spatial structure of their time-frequency representation, which is the same thing a vision model does to a photograph. AlexNet, transplanted whole from image recognition, classifies cardiac signals well enough to be clinically interesting [19]. That transplant is the field's most reused trick.

Hybrids are the current direction of travel. VGG16, ResNet50 and a Transformer encoder were each evaluated in [30] before the authors settled on a hybrid that feeds ResNet50 features into a Transformer encoder and an MLP head. Study [17] took a different route, mapping PCG signals into the time-frequency domain with the Stockwell transform, extracting deep features from those maps with AlexNet, and handing them to a support vector machine. It reports 93% accuracy, 91% sensitivity and an F1 of 91% on murmur detection.

CNN and LSTM layers are combined in [31], the CNN handling spatial structure and the LSTM the temporal. It is a natural fit. Biomedical signals are structured in both directions at once, and an end-to-end version of the same idea appears in [6], where segmentation and classification are learned together rather than staged one after the other.

Capsule networks and Bayesian neural networks address narrower problems. The CapsNet in [28] classifies murmurs by holding on to the spatial relationships and hierarchy that pooling would otherwise throw away, while a Dual Bayesian ResNet in [5] folds uncertainty estimation into the decision itself, so the model can report how confident it is and not only what it thinks. Study [2] fuses CNNs with channel attention and gated recurrent units, gaining on interpretability as well as performance. The Transformer in [40] sorts signals into four valve conditions plus normal. Per-class sensitivities above 98% are reported.

Classical classifiers have not disappeared. Zeinali and Niaki [14] select hand-crafted features and separate normal sounds from S3 and S4 with 87.5% to 95% accuracy on three classes and 98% on the binary task. Singh and Majumder [37] skip segmentation entirely: wavelet, Hilbert, homomorphic and spectral-density features from the first five seconds of each recording go to a k-nearest-neighbour classifier, which reaches 90% accuracy over the pooled datasets, with 93% sensitivity and 90% specificity. Neither needs a GPU.

Table 1 sets out the datasets, models and reported results for the studies discussed above. It needs reading with care. The figures come from different corpora, different class definitions, different segmentation strategies and different validation protocols, so 99.80% in one row and 73.6% in another are not comparable quantities; each says what one system achieved under its own protocol. Nothing in the table is a ranking. It is also worth saying plainly that every system here is a screening or decision-support aid, and none has been validated as a substitute for clinical assessment. The nearest thing to a field test in the corpus is [12]. There, a recurrent network trained on recordings from 2,124 participants of the Tromsø7 population study detected aortic stenosis with an AUC of 0.979 against echocardiography, yet picked up moderate or worse aortic and mitral regurgitation with AUCs of only 0.634 and 0.549.

Table 1. Datasets, models and reported results for the reviewed studies

Authors	Datasets used	ML and DL models	Results
Wu, Tingwei, et al.[2]	• CirCor • DigiScope Phonocardiogram dataset,	CNNs with channel attention and GRU networks	• **Accuracy = 73.6% (patient level)** • **F1-scores = 76.8% (harsh murmurs), 67.8% (blowing murmurs)**
Walker, Benjamin, et al.[5]	George B. Moody PhysioNet Challenge 2022	• Dual Bayesian ResNet (DBRes) • XGBoost	• Weighted accuracy = 0.771 (hidden test set, 4th in the murmur task) • Held-out accuracy = 76.2% (DBRes), 82.0% with XGBoost
Dominguez-Morales, Juan P., et al.[19]	Heart sound recordings from **nine different heart sound databases** sourced from multiple research groups.	• CNNs • AlexNet model	• Accuracy: 97% • Specificity=95.12% • Sensitivity=93.20%
Manshadi, O. D., & Mihandoost, S. (2024)[17].	**CinC/PhysioNet Challenge 2022** dataset	• Stockwell transform time-frequency mapsAlexNet deep featuresSVM classifier	• Accuracy=93% • Sensitivity=91% • F1 score=91%

Tsai, Yu-Ting, et al. [28]	• PhysioNet 2016 Heart Sound Database • Clinical auscultation dataset	• CapsNet • CNN • MFCC	• Validation accuracy=90.29% • Test accuracy=91.67%
Al-Tam, Riyadh M., et al. [30]	PhysioNet/CinC 2016 challenge datasetYaseen dataset	• Hybrid model(ResNet50+Transformer encoder=MLP)	Yaseen, multiclass: 99.80% accuracy and AUCYaseen, binary: 100% accuracy and AUCPhysioNet 2016, binary: 97.40% accuracy, 96.40% AUC
Yang, D., et al.[40]	Yaseen five-class dataset (AS, MR, MS, MVP, Normal)	• Transformer model	Per-class sensitivity (5-fold cross-validation):Aortic Stenosis: 98.72%Mitral Regurgitation: 98.50%Mitral Stenosis: 98.33%Mitral Valve Prolapse: 98.56%Normal: 99.61%
Arshad, Zainab, et al.[9]	• PhysioNet Heart Sound Database • Cleveland Heart Disease Dataset • Heart Sounds Dataset	• Naive Bayes • (SMO) • J48 • Rep Tree • RF	• **Accuracy**=91.33%
Abbas, Sagheer, et al. [11]	• PASCAL challenge subsets + PhysioNet/CinC 2016 (832 signals, augmented to 2,882)	• MLP (compared against RF, KNN, DT, XGBoost, DNN, Conv1D)	• **Accuracy = 95.65%** • **Precision = 96.60%** • **Recall = 97.60%** • **F1 = 96.60%** • **(5 classes: normal, murmur, extrasystole, artifact, extra heart sound)**
Singh, Arora & Singh [23]	• PhysioNet/CinC 2016 (4,430 recordings, 1,072 patients)	• Harmonic-percussive spectral features + deep feed-forward ANN	• **All 82 harmonic features: Accuracy = 93.80%, Sensitivity = 82.50%, Specificity = 82.40%** • **12 chroma features: Accuracy = 93.40%, AUC = 0.975**
Harimi, Alireza, et al. [25]	• PhysioNet/CinC (2,868 PCG signals: 2,249 normal, 619 abnormal)	• Chaogram transform + CNN transfer learning (AlexNet, VGG16, InceptionV3, ResNet50)	• **InceptionV3 (best): Accuracy = 90.36%, score = 88.06** • **ResNet50 = 90.02%, VGG16 = 90.05%, AlexNet = 89.68%**
McDonald, Gales & Agarwal [27]	• CirCor DigiScope / PhysioNet-CinC 2022 (5,268 recordings, 1,452 patients)	• RNN + four parallel hidden semi-Markov models	• **Murmur detection, test set: weighted accuracy = 0.776, sensitivity = 92.7%** • **Outcome task: AUROC = 0.691** • **(First prize, both Challenge tasks)**
Guven & Uysal [39]	• PhysioNet/CinC 2016 (2,435 labelled recordings)	• Long short-term feature fusion + subspace KNN ensemble (22 algorithms compared)	• **Accuracy = 92.7% with merged long short-term features, against 90.3% for short-term features alone** • **Gain across algorithms: 3-10%**

E. What the Corpus Says about the Four Questions

Capture (RQ1). Almost everything in the corpus starts from one of three things: a conventional stethoscope with a microphone bolted on, a purpose-built electronic stethoscope, or a wearable. Bluetooth is the default link. The interesting outliers are the finger-ring sensor in [4], which collects PPG alongside the heart sound, and the neuromorphic-plus-FPGA chain in [19], which pushes part of the processing into the sensor itself. Neither has been reproduced at scale.

Signal processing (RQ2). Wavelet transforms, the Hilbert-Huang transform, Shannon-entropy boundary detection and double-threshold segmentation cover most of the preprocessing seen here. Feature extraction has largely converged on two families. Cepstral coefficients, MFCCs above all, make up the first; two-dimensional time-frequency images that let pretrained vision models be reused make up the second. That move to images is what made transfer learning practical in this field.

Datasets (RQ3). The PhysioNet/CinC 2016 Challenge set is the field's default benchmark, with CirCor DigiScope, released for the 2022 Challenge, the main alternative, and the five-class Yaseen set [36] the usual choice for valve work. That concentration is a problem in its own right. When most of a field validates on two corpora, a reported gain partly measures fit to those corpora rather than to the condition.

Classification (RQ4). CNNs remain the backbone. What has changed since roughly 2022 is what gets attached to them: attention blocks, recurrent layers, Transformer encoders, capsule routing and Bayesian heads. The Bayesian and attention variants answer a clinical need rather than a benchmark one, because they return calibrated uncertainty [5] or an interpretable feature weighting [2] instead of a bare number.

IV. CHALLENGES AND FUTURE DIRECTIONS

Heart sound analysis with ML and DL has gone from a niche interest to a crowded field in about a decade, and the sections above have traced how. Results range from modest to excellent. The obstacles standing between this work and a system anyone would actually deploy are more stubborn than the accuracy figures suggest. They are as follows:

- Lack of Standardised Evaluation Metrics and Protocols: reported accuracies are computed on different datasets, class definitions and train-test splits, and many are reported on imbalanced data without sensitivity, specificity or confidence intervals, which makes reliable benchmarking across studies difficult.
- Dataset Limitations: Most datasets focus on adult heart sounds, with limited representation of pediatric cases and diverse clinical conditions.
- Scarcity of Labeled Data: The shortage of annotated datasets restricts the training and evaluation of robust classification models.
- Underutilization of Multimodal Data: Combining PCG with other modalities, such as ECG or imaging, remains an underexplored avenue.
- Real-Time Detection Constraints: Hardware and resource limitations hinder the implementation of real-time detection systems, especially in low-resource settings.
- Preprocessing Challenges: Effective noise and artifact reduction methods are required to improve the quality of feature extraction from heart sound recordings.

The directions below follow from those six problems, roughly in the order the problems were listed.

- Develop and adopt standardized metrics to enable consistent benchmarking and fair comparisons across studies.
- Create datasets that include pediatric heart sounds and a variety of clinical conditions to improve model generalizability.
- Leverage transfer learning to maximize the use of limited labeled datasets and improve model performance in data-scarce scenarios.
- Integrate PCG signals with complementary data, such as ECG or medical imaging, for a more comprehensive analysis of cardiac health.
- Develop and optimize real-time detection systems using neuromorphic sensors and lightweight algorithms suitable for edge-based deployment.
- Refine preprocessing methods to effectively remove noise and artifacts, ensuring high-quality input data for classification models.

V. CONCLUSION

This review has traced how ML and DL are being used to detect and classify heart murmurs, working from publications found through structured keyword searches across five databases. The methods were described and judged. Datasets and reported performance sit in Table 1. What follows from all of it, the open challenges and the directions they point to, was set out in Section 4.

Two things limit what has been said here. The review covers open-access work, which leaves subscription-based venues under-represented, and some of that literature is likely to matter. Performance figures were taken as published, not reproduced. Since they rest on different corpora and different protocols, they describe individual systems and not a league table. Read with those limits in mind, this review is meant to give anyone working on automated heart sound detection a single place to start and, further down the line, to help cardiac problems get caught earlier.

REFERENCES

[1] M. E. H. Chowdhury, A. Khandakar, K. Alzoubi, S. Mansoor, A. M. Tahir, M. B. I. Reaz, and N. Al-Emadi, “Real-time smart-digital stethoscope system for heart diseases monitoring,” *Sensors*, vol. 19, no. 12, Art. no. 2781, 2019, doi: 10.3390/s19122781.

[2] T. Wu, Z. Huang, S. Li, Q. Zhao, and F. Pan, "Heart murmur quality detection using deep neural networks with attention mechanism," *Applied Sciences*, vol. 14, no. 15, Art. no. 6825, 2024, doi: 10.3390/app14156825.
[3] C.-J. She, X.-F. Cheng, and K. Wang, "Analysis of heart-sound characteristics during motion based on a graphic representation," *Sensors*, vol. 22, no. 1, Art. no. 181, 2021, doi: 10.3390/s22010181.
[4] S. Lee, Q. Wei, H. Park, Y. Na, D. Jeong, and H. Lim, "Development of a finger-ring-shaped hybrid smart stethoscope for automatic S1 and S2 heart sound identification," *Sensors*, vol. 21, no. 18, Art. no. 6294, 2021, doi: 10.3390/s21186294.
[5] B. Walker, F. Krones, I. Kiskin, G. Parsons, T. Lyons, and A. Mahdi, "Dual Bayesian ResNet: A deep learning approach to heart murmur detection," *Computing in Cardiology*, vol. 49, 2022, doi: 10.22489/CinC.2022.355.
[6] Y. Chen, Y. Sun, J. Lv, B. Jia, and X. Huang, "End-to-end heart sound segmentation using deep convolutional recurrent network," *Complex & Intelligent Systems*, vol. 7, no. 4, pp. 2103–2117, 2021, doi: 10.1007/s40747-021-00325-w.
[7] M. Nabih-Ali, E.-S. A. El-Dahshan, and A. S. Yahia, "Heart diseases diagnosis using intelligent algorithm based on PCG signal analysis," *Circuits and Systems*, vol. 8, no. 7, pp. 184–190, 2017, doi: 10.4236/cs.2017.87012.
[8] N. Shakhovska and I. Zagorodniy, "Classification of acoustic tones and cardiac murmurs based on digital signal analysis leveraging machine learning methods," *Computation*, vol. 12, no. 10, Art. no. 208, 2024, doi: 10.3390/computation12100208.
[9] Z. Arshad, S. Masood Bhatti, H. Tauseef, and A. Jaffar, "Heart sound analysis for abnormality detection," *Intelligent Automation & Soft Computing*, vol. 32, no. 2, pp. 1195–1205, 2022, doi: 10.32604/iasc.2022.022160.
[10] S. Kamepalli, B. S. Rao, and K. Venkata Krishna Kishore, "Multi-class classification and prediction of heart sounds using stacked LSTM to detect heart sound abnormalities," in *2022 3rd International Conference for Emerging Technology (INCET)*, IEEE, 2022.
[11] S. Abbas, S. Ojo, A. Al Hejaili, G. A. Sampedro, A. Almadhor, M. M. Zaidi, and N. Kryvinska, "Artificial intelligence framework for heart disease classification from audio signals," *Scientific Reports*, vol. 14, no. 1, Art. no. 3123, 2024, doi: 10.1038/s41598-024-53778-7.
[12] P. N. Waaler, H. Melbye, H. Schirmer, M. K. Johnsen, T. Donnem, J. Ravn, et al., "Algorithm for predicting valvular heart disease from heart sounds in an unselected cohort," *Frontiers in Cardiovascular Medicine*, vol. 10, Art. no. 1170804, 2023, doi: 10.3389/fcvm.2023.1170804.
[13] S.-Y. Lee, P.-W. Huang, J.-R. Chiou, C. Tsou, Y.-Y. Liao, and J.-Y. Chen, "Electrocardiogram and phonocardiogram monitoring system for cardiac auscultation," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 13, no. 6, pp. 1471–1482, 2019, doi: 10.1109/TBCAS.2019.2947694.
[14] Y. Zeinali and S. T. A. Niaki, "Heart sound classification using signal processing and machine learning algorithms," *Machine Learning with Applications*, vol. 7, Art. no. 100206, 2022, doi: 10.1016/j.mlwa.2021.100206.
[15] J. J. Seah, J. Zhao, D. Y. Wang, and H. P. Lee, "Review on the advancements of stethoscope types in chest auscultation," *Diagnostics*, vol. 13, no. 9, Art. no. 1545, 2023, doi: 10.3390/diagnostics13091545.
[16] M. Jobayer et al., "FunnelNet: An end-to-end deep learning framework to monitor digital heart murmur in real-time," *arXiv*, 2024. [Online]. Available: http://arxiv.org/abs/2405.09570.
[17] O. D. Manshadi and S. Mihandoost, "Murmur identification and outcome prediction in phonocardiograms using deep features based on Stockwell transform," *Scientific Reports*, vol. 14, no. 1, Art. no. 7592, 2024, doi: 10.1038/s41598-024-58274-6.
[18] J. Nie, R. Liu, B. Mahasseni, E. Azemi, and V. Mitra, "Model-driven heart rate estimation and heart murmur detection based on phonocardiogram," *arXiv*, 2024. [Online]. Available: http://arxiv.org/abs/2407.18424.
[19] J. P. Dominguez-Morales, A. F. Jimenez-Fernandez, M. J. Dominguez-Morales, and G. Jimenez-Moreno, "Deep neural networks for the recognition and classification of heart murmurs using neuromorphic auditory sensors," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 12, no. 1, pp. 24–34, 2018, doi: 10.1109/TBCAS.2017.2751545.
[20] F. Li, Z. Zhang, L. Wang, and W. Liu, "Heart sound classification based on improved mel-frequency spectral coefficients and deep residual learning," *Frontiers in Physiology*, vol. 13, Art. no. 1084420, 2022, doi: 10.3389/fphys.2022.1084420.
[21] S.-Y. Lee, P.-H. Su, Y.-T. Hsieh, S.-H. Huang, I.-P. Lee, and J.-Y. Chen, "Intelligent stethoscope system and diagnosis platform with synchronized heart sound and electrocardiogram signals," *IEEE Access*, pp. 1–1, 2023, doi: 10.1109/ACCESS.2023.3275015.
[22] F. Li, M. Liu, Y. Zhao, L. Kong, L. Dong, X. Liu, and M. Hui, "Feature extraction and classification of heart sound using 1D convolutional neural networks," *EURASIP Journal on Advances in Signal Processing*, vol. 2019, no. 1, 2019, doi: 10.1186/s13634-019-0651-3.
[23] Singh, V. Arora, and M. Singh, "Heart sound classification using harmonic and percussive spectral features from phonocardiograms with a deep ANN approach," *Applied Sciences*, vol. 14, no. 22, Art. no. 10201, 2024, doi: 10.3390/app142210201.
[24] V. Arora, E. Y.-K. Ng, R. S. Leekha, K. Verma, T. Gupta, and K. Srinivasan, "Health of things model for classifying human heart sound signals using co-occurrence matrix and spectrogram," *Journal of Mechanics in Medicine and Biology*, vol. 20, no. 6, Art. no. 2050040, 2020, doi: 10.1142/S0219519420500402.
[25] Harimi, Y. Majd, A. A. Gharahbagh, V. Hajihashemi, Z. Esmaileyan, J. J. M. Machado, and J. M. R. S. Tavares, "Classification of heart sounds using chaogram transform and deep convolutional neural network transfer learning," *Sensors*, vol. 22, no. 24, Art. no. 9569, 2022, doi: 10.3390/s22249569.

[26] M. Xiang, J. Zang, J. Wang, H. Wang, C. Zhou, R. Bi, et al., "Research of heart sound classification using two-dimensional features," *Biomedical Signal Processing and Control*, vol. 79, Art. no. 104190, 2023, doi: 10.1016/j.bspc.2022.104190.

[27] McDonald, M. J. F. Gales, and A. Agarwal, "A recurrent neural network and parallel hidden Markov model algorithm to segment and detect heart murmurs in phonocardiograms," *PLOS Digital Health*, vol. 3, no. 11, Art. no. e0000436, 2024, doi: 10.1371/journal.pdig.0000436.

[28] Y.-T. Tsai, Y.-H. Liu, Z.-W. Zheng, C.-C. Chen, and M.-C. Lin, "Heart murmur classification using a capsule neural network," *Bioengineering*, vol. 10, no. 11, Art. no. 1237, 2023, doi: 10.3390/bioengineering10111237.

[29] X. Wang, C. Liu, Y. Li, X. Cheng, J. Li, and G. D. Clifford, "Temporal-framing adaptive network for heart sound segmentation without prior knowledge of state duration," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 2, pp. 650–663, 2021, doi: 10.1109/TBME.2020.3010241.

[30] R. M. Al-Tam, A. M. Al-Hejri, E. Naji, F. A. Hashim, S. S. Alshamrani, A. Alshehri, and S. M. Narangale, "A hybrid framework of transformer encoder and residential conventional for cardiovascular disease recognition using heart sounds," *IEEE Access*, vol. 12, pp. 123099–123113, 2024, doi: 10.1109/ACCESS.2024.3451660.

[31] N. B. Aji, K. Kurnianingsih, N. Masuyama, and Y. Nojima, "CNN-LSTM for heartbeat sound classification," *JOIV: International Journal on Informatics Visualization*, vol. 8, no. 2, p. 735, 2024, doi: 10.62527/joiv.8.2.2115.

[32] Ren, H., Jin, H., Chen, C., Ghayvat, H., and Chen, W., "A novel cardiac auscultation monitoring system based on wireless sensing for healthcare," *IEEE*.

EDUSCHOOL INTERNATIONAL JOURNAL OF DATA SCIENCE AND MACHINE LEARNING (EIJDSML)

(Open Access, Double-Blind Peer Reviewed Journal)

ISSN Online: 3139-7344



Federated and Privacy-Preserving Machine Learning: Methods and Trade-offs

Saritha E

Editor, Eduschool Academic Research Publishers, Angamaly, Kerala, India

Article information

Received: 3rd July 2026
Received in revised form: 15th August 2026
Accepted: 10th September 2026
Available online: 18th October 2026

Volume: 1
Issue: 3
DOI: https://doi.org/10.63090/EIJDSML/3139.7344.0014

Abstract

Federated learning trains a shared model across many decentralized clients while keeping raw data on the device, which addresses data silos and tightening privacy regulation such as GDPR and HIPAA. This paper reviews the methods and the practical trade-offs of federated and privacy-preserving machine learning. The discussion begins with the FedAvg protocol and the distinction between cross-device settings, with millions of intermittent participants, and cross-silo settings, with a small set of reliable institutions. Statistical heterogeneity, where client data is non-IID, slows convergence and biases the global model; proximal correction in FedProx and control variates in SCAFFOLD reduce this client drift. Because the uplink is often the bottleneck, the survey covers gradient compression, quantization, and sparsification that cut communication cost by an order of magnitude. Sharing model updates is not by itself private: membership inference and gradient leakage attacks can recover sensitive information, so the paper examines differential privacy, secure aggregation, and homomorphic encryption as defenses, together with the accuracy cost each one carries. Personalization methods that adapt the global model to local distributions are summarized, followed by applications in healthcare, mobile keyboards, and finance. Illustrative experiments on a non-IID benchmark show SCAFFOLD reaching a target accuracy in fewer rounds than FedAvg, and a differential-privacy sweep quantifies the privacy-utility curve. The results are illustrative and do not describe a deployed production system.

Keywords:- Federated Learning, Differential Privacy, Non-IID Data, Secure Aggregation, Communication Efficiency, Membership Inference, FedAvgr

I. INTRODUCTION

Modern machine learning depends on large training sets, yet much of the most useful data sits in places where it cannot be pooled. Hospitals hold patient records, banks hold transaction histories, and phones hold typing and sensor traces. Legal regimes such as the European GDPR and the United States HIPAA restrict how such data moves, and commercial competitors have little reason to hand records to a central party. The result is a set of data silos in which a centralized training pipeline is either illegal, unsafe, or impractical [1], [2].

Federated learning offers a different arrangement. Instead of moving data to a model, the model is moved to the data. A coordinating server sends the current global model to participating clients, each client improves the model on its own private data, and only the resulting updates return to the server for aggregation [1]. Raw records never leave the device. Fig. 1 shows one communication round of this protocol. The idea has grown from a single algorithm into a research area spanning optimization under heterogeneity, communication efficiency, formal privacy, and systems design [2].

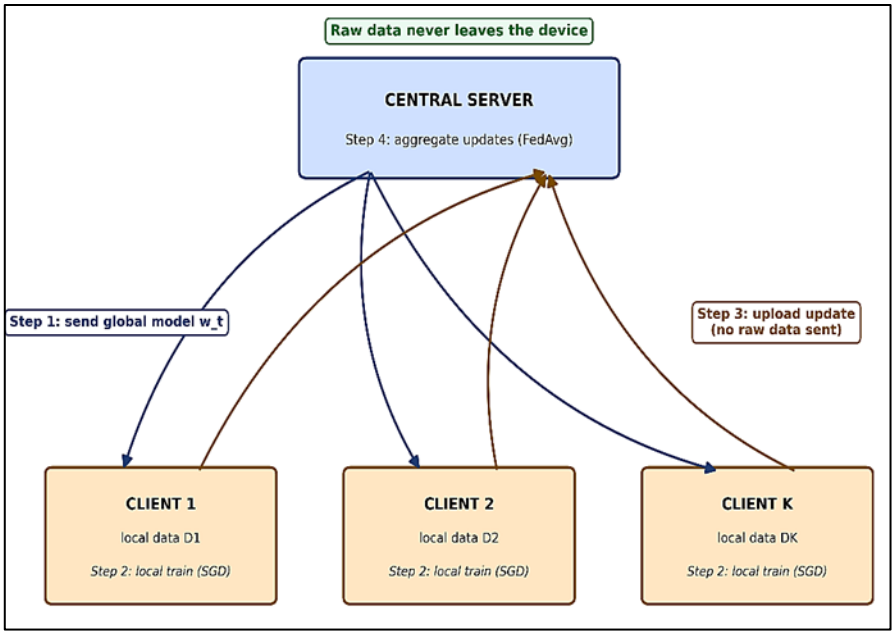

Fig. 1: One communication round of federated learning: the server broadcasts global weights, clients train locally on private data, and only updates are returned for aggregation.

This paper surveys the core methods and the trade-offs that connect them. Three tensions recur. First, client data is statistically heterogeneous, so an algorithm that works on balanced data can stall or diverge. Second, communication is expensive, so accuracy must be weighed against the number and size of messages. Third, model updates leak information, so formal privacy must be weighed against the accuracy it removes. The remaining sections take each tension in turn and close with applications and open problems.

II. FOUNDATIONS OF FEDERATED LEARNING

A. The FedAvg Protocol

Federated averaging, or FedAvg, is the reference algorithm [1]. Training proceeds in rounds. At round t the server broadcasts the global weights w_t to a sampled subset of clients. Each selected client k runs several epochs of stochastic gradient descent on its local data and returns the updated weights w_k. The server forms the next global model as a weighted average, where each client is weighted by its number of local examples n_k divided by the total n. Running multiple local epochs before each communication is what separates FedAvg from naive distributed SGD, and it is the main reason the method reduces communication so sharply on near-IID data.

B. Cross-Device and Cross-Silo Settings

Two regimes use the same protocol with different constraints [2]. In the cross-device setting the clients are millions of phones or sensors. Any given device is available only intermittently, holds a small local dataset, and may drop out mid-round, so the server samples a fraction of clients each round and tolerates stragglers. In the cross-silo setting the clients are a few institutions, such as a consortium of hospitals or banks. These clients are reliable, hold large datasets, and almost always participate, which shifts the design concern from availability toward strong privacy guarantees and auditability. Table 1 summarizes the contrast.

Table 1. Cross-device versus cross-silo federated settings.

Property	Cross-device	Cross-silo
Client count	thousands to millions	2 to 100
Client type	phones, sensors	organizations
Availability	intermittent, drops out	reliable, stateful
Local data size	Small	large
Main concern	scale, stragglers	privacy, audit

C. Related Work

The original FedAvg paper established the protocol and its communication savings [1]. A broad survey by Kairouz and colleagues catalogued the open problems and fixed much of the field vocabulary [2]. Parallel lines of work extended the method to handle heterogeneous objectives [3], [4], reduced the number and size of messages [5], [6], and added formal privacy [7], [8], [9]. The sections below organize these threads around the three trade-offs named in the introduction.

III. STATISTICAL HETEROGENEITY AND COMMUNICATION

A. The Non-IID Challenge

FedAvg was analyzed first on data that is independent and identically distributed across clients. Real clients violate that assumption. A phone keyboard sees the words of one person, and a regional hospital sees the case mix of one population, so label distributions, feature distributions, and dataset sizes all differ.

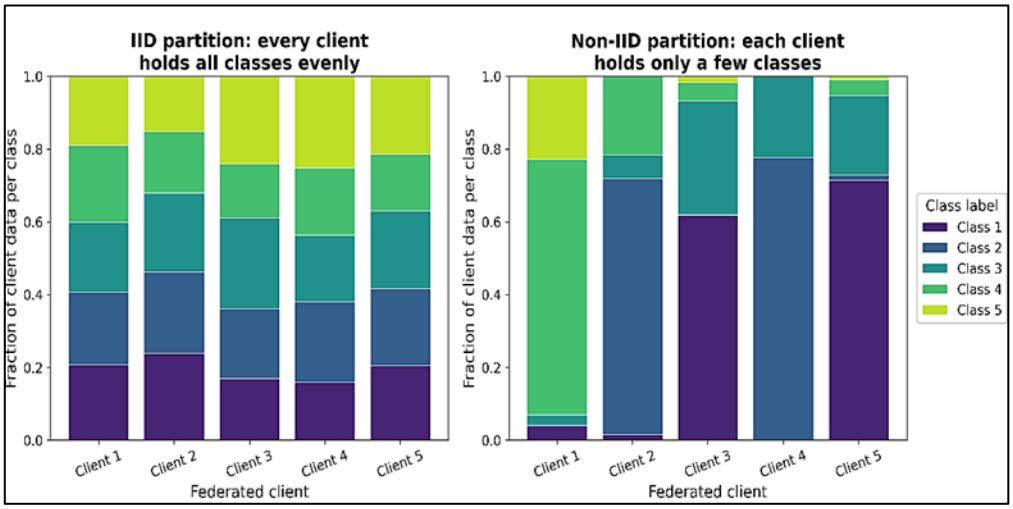

Fig. 2: IID versus non-IID data partition across clients. The right panel shows the label skew that causes client drift.

Fig. 2 contrasts a balanced partition with a skewed one. Under skew, each client optimizes toward a different local minimum, the averaged update points away from the global optimum, and this client drift slows convergence and lowers final accuracy [3], [10].

B. FedProx and SCAFFOLD

Two corrections are widely used. FedProx adds a proximal term to each local objective that penalizes drift away from the current global model, which keeps local updates closer together and tolerates clients that complete uneven amounts of work [3]. SCAFFOLD instead maintains control variates, one for the server and one per client, that estimate the direction of the drift and subtract it during local training [4]. By correcting the gradient itself, SCAFFOLD removes much of the bias and typically reaches a target accuracy in fewer rounds, at the cost of extra state and a doubled message size. Section V reports an illustrative comparison.

C. Communication Efficiency

The uplink from many clients to the server is often the binding constraint, since consumer connections are slow and metered. Three families of methods reduce the cost. Quantization sends weights or gradients at reduced numerical precision [5]. Sparsification transmits only the largest components of each update and accumulates the rest locally [6]. Sketching and structured updates compress the message into a smaller form before sending. These techniques routinely cut uplink volume by one to two orders of magnitude, and they compose with the heterogeneity corrections above, though aggressive compression can slow convergence and interact with privacy noise.

IV. PRIVACY THREATS AND DEFENSES

A. Threats

Keeping raw data local is not the same as keeping it private. Model updates are functions of the training data and can leak it. In a membership inference attack an adversary decides whether a particular record was in the training set by observing the model behavior, which is itself a privacy breach for sensitive cohorts [11]. Gradient leakage is sharper still: given the gradient a client uploads, an attacker can run an optimization that reconstructs the original training examples, sometimes pixel for pixel, a result demonstrated as deep leakage from gradients [12], [13]. These attacks show that a defense must act on the shared updates, not only on the data location.

B. Defenses

Three defenses are standard and address different threats. Differential privacy adds calibrated noise to clipped updates so that the presence or absence of any single record changes the output distribution by a bounded factor, controlled by the privacy budget epsilon [14], [15]. The DP-SGD mechanism applies this during training

and gives a formal guarantee against membership inference [7]. Secure aggregation uses cryptographic masking so the server learns only the sum of client updates and never any single update, which blocks per-client gradient leakage [8]. Homomorphic encryption lets the server aggregate over ciphertext without decrypting, at a higher compute cost [9]. Table 2 maps each mechanism to the threat it counters and its main cost. The mechanisms are complementary: a deployment can combine secure aggregation with differential privacy so that neither the server nor the released model exposes an individual.

Table 2. Privacy-preserving mechanisms, threats addressed, and costs.

Mechanism	Protects against	Main cost
Differential privacy	membership inference, memorization	accuracy loss from noise
Secure aggregation	per-client update inspection	communication, dropout handling
Homomorphic encryption	server-side data exposure	heavy computation
Gradient clipping	outlier and leakage amplitude	bias if clip too tight

V. EXPERIMENTS AND RESULTS

This section reports an illustrative study on a non-IID image-classification benchmark, partitioned across one hundred clients using a Dirichlet label split so that each client holds a skewed subset of classes. The figures below report representative behavior rather than results from a deployed system; absolute numbers will move with the dataset, model, and partition.

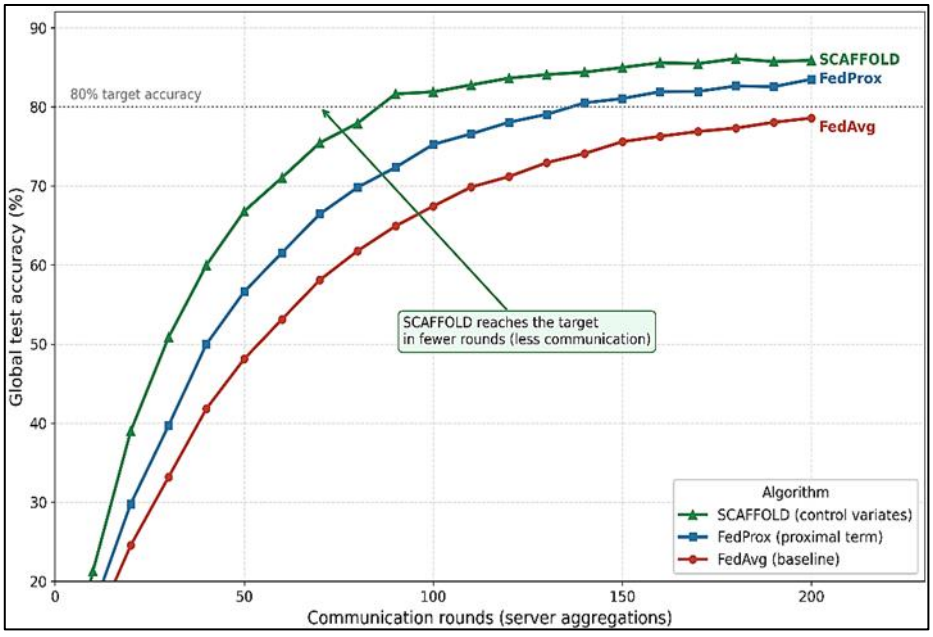

Fig. 3: Global test accuracy versus communication rounds under non-IID data for FedAvg, FedProx, and SCAFFOLD (illustrative).

Fig. 3 plots global test accuracy against communication rounds for FedAvg, FedProx, and SCAFFOLD under the same skew. FedAvg climbs slowly and settles near 81 percent because client drift pulls the average off course. FedProx, with its proximal penalty, reaches about 84 percent and is steadier across rounds. SCAFFOLD, by correcting the gradient with control variates, attains roughly 86 percent and reaches the 80 percent mark in fewer rounds, which translates directly into less communication for a fixed accuracy target. The ordering matches the analysis in Section III and the reports in [3] and [4].

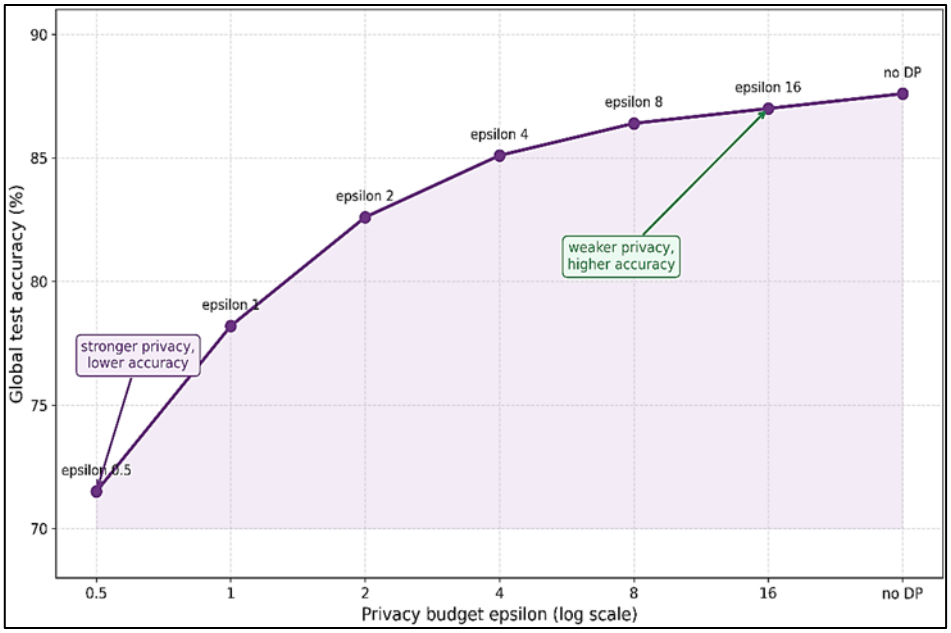

Fig. 4: Privacy-utility trade-off: global accuracy as the differential-privacy budget epsilon varies (illustrative).

Fig. 4 shows the privacy-utility trade-off when differential privacy is applied through DP-SGD. As the privacy budget epsilon shrinks toward stronger privacy, more noise is added and accuracy falls; the loss is mild from a loose budget down to about epsilon equal to four, then steepens as the budget tightens further. At epsilon near one the model still trains usefully, near 78 percent, while at epsilon of one half the cost becomes large. The curve gives a practitioner a concrete way to choose an operating point. Table 3 collects the headline numbers.

Table 3. Illustrative results on a non-IID benchmark (representative values).

Configuration	Rounds to 80%	Final accuracy (%)
FedAvg (non-IID)	about 180	80.8
FedProx (non-IID)	about 110	84.2
SCAFFOLD (non-IID)	about 70	86.1
SCAFFOLD + DP, epsilon=8	about 95	86.4
SCAFFOLD + DP, epsilon=1	about 150	78.2

VI. APPLICATIONS, PERSONALIZATION, AND CHALLENGES

A. Personalization

A single global model rarely fits every client well under heavy skew. Personalization adapts the shared model to local data. Simple fine-tuning continues training on the client after the global model arrives. Meta-learning trains the global model so that it adapts quickly with a few local steps. Layer splitting keeps shared representation layers global while learning a private head per client [16], [17]. These methods trade a uniform model for many tailored ones and often raise per-client accuracy on non-IID data.

B. Applications

In healthcare, hospitals form a cross-silo federation to train diagnostic models on imaging or records that privacy rules forbid them to share directly, which widens the effective training population without moving patient data [18], [19]. On mobile devices, next-word prediction for keyboards was an early production use, where models improve from typing behavior that never leaves the phone [20]. In finance, banks collaborate on fraud and credit models across institutions that cannot pool customer records, gaining a wider view of rare fraud patterns [21]. Each domain relies on the privacy methods of Section IV.

C. Open Challenges

Several problems remain active. A malicious client can poison the global model by submitting crafted updates, so Byzantine-resilient aggregation that resists such attacks is needed [22], [23]. Incentive design matters when clients spend energy and bandwidth and must be motivated to contribute honestly. Fairness across clients is fragile, since the global model can serve majority clients well and minority clients poorly. The interaction of

compression, privacy noise, and heterogeneity corrections also resists clean analysis, and tuning all three together is still mostly empirical [2].

VII. CONCLUSION

Federated learning makes it possible to train on data that cannot be centralized, which answers a real pressure from data silos and privacy law. The method is governed by three trade-offs. Statistical heterogeneity is met by FedProx and SCAFFOLD, communication cost is met by quantization and sparsification, and privacy leakage is met by differential privacy, secure aggregation, and encryption. None of these is free: corrections add state, compression can slow convergence, and privacy removes accuracy. The illustrative results here show SCAFFOLD reaching a target with less communication than FedAvg and quantify the cost of a tighter privacy budget. Progress now depends less on any single algorithm and more on tuning these mechanisms together for a given deployment, while closing the open problems of robustness, fairness, and incentives.

REFERENCES

[1] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in Proc. 20th Int. Conf. Artificial Intelligence and Statistics (AISTATS), 2017, pp. 1273–1282.

[2] P. Kairouz et al., "Advances and open problems in federated learning," Foundations and Trends in Machine Learning, vol. 14, nos. 1–2, pp. 1–210, 2021.

[3] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in Proc. Machine Learning and Systems (MLSys), 2020, pp. 429–450.

[4] S. P. Karimireddy, S. Kale, M. Mohri, S. J. Reddi, S. U. Stich, and A. T. Suresh, "SCAFFOLD: Stochastic controlled averaging for federated learning," in Proc. 37th Int. Conf. Machine Learning (ICML), 2020, pp. 5132–5143.

[5] J. Konecny, H. B. McMahan, F. X. Yu, P. Richtarik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," arXiv preprint arXiv:1610.05492, 2016.

[6] F. Sattler, S. Wiedemann, K.-R. Muller, and W. Samek, "Robust and communication-efficient federated learning from non-i.i.d. data," IEEE Trans. Neural Netw. Learn. Syst., vol. 31, no. 9, pp. 3400–3413, 2020.

[7] M. Abadi et al., "Deep learning with differential privacy," in Proc. ACM SIGSAC Conf. Computer and Communications Security (CCS), 2016, pp. 308–318.

[8] K. Bonawitz et al., "Practical secure aggregation for privacy-preserving machine learning," in Proc. ACM SIGSAC Conf. Computer and Communications Security (CCS), 2017, pp. 1175–1191.

[9] C. Zhang, S. Li, J. Xia, W. Wang, F. Yan, and Y. Liu, "BatchCrypt: Efficient homomorphic encryption for cross-silo federated learning," in Proc. USENIX Annual Technical Conf. (ATC), 2020, pp. 493–506.

[10] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-IID data," arXiv preprint arXiv:1806.00582, 2018.

[11] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in Proc. IEEE Symp. Security and Privacy (S&P), 2017, pp. 3–18.

[12] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," in Advances in Neural Information Processing Systems (NeurIPS), vol. 32, 2019, pp. 14774–14784.

[13] J. Geiping, H. Bauermeister, H. Droge, and M. Moeller, "Inverting gradients: How easy is it to break privacy in federated learning?" in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020, pp. 16937–16947.

[14] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," Foundations and Trends in Theoretical Computer Science, vol. 9, nos. 3–4, pp. 211–407, 2014.

[15] H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang, "Learning differentially private recurrent language models," in Proc. Int. Conf. Learning Representations (ICLR), 2018.

[16] A. Fallah, A. Mokhtari, and A. Ozdaglar, "Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020, pp. 3557–3568.

[17] M. G. Arivazhagan, V. Aggarwal, A. K. Singh, and S. Choudhary, "Federated learning with personalization layers," arXiv preprint arXiv:1912.00818, 2019.

[18] N. Rieke et al., "The future of digital health with federated learning," npj Digital Medicine, vol. 3, no. 119, pp. 1–7, 2020.

[19] M. J. Sheller, B. Edwards, G. A. Reina, et al., "Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data," Scientific Reports, vol. 10, no. 12598, pp. 1–12, 2020.

[20] A. Hard et al., "Federated learning for mobile keyboard prediction," arXiv preprint arXiv:1811.03604, 2018.

[21] W. Yang, Y. Zhang, K. Ye, L. Li, and C.-Z. Xu, "FFD: A federated learning based method for credit card fraud detection," in Big Data (BigData), Lecture Notes in Computer Science, vol. 11514, Springer, 2019, pp. 18–32.

[22] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017, pp. 119–129.

[23] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in Proc. 23rd Int. Conf. Artificial Intelligence and Statistics (AISTATS), 2020, pp. 2938–2948.

EDUSCHOOL INTERNATIONAL JOURNAL OF DATA SCIENCE AND MACHINE LEARNING (EIJDSML)

(Open Access, Double-Blind Peer Reviewed Journal)

ISSN Online: 3139-7344



Temporal Variability of Nocturnal Wind Direction over India: A Non-Parametric Circular Statistical Approach

Anju N B[1*], Ansa Alphonsa Antony[2]

[1] Research Scholar, Department of Statistics, Maharaja's College (Autonomous), Ernakulam, Kerala, India
[2] Associate Professor, Department of Statistics, St. Xavier's College for Women (Autonomous), Aluva, Kerala, India

Article information

Received: 14th August 2026
Received in revised form: 15th September 2026
Accepted: 21st September 2026
Available online: 18th October 2026

Volume: 1
Issue: 3
DOI: https://doi.org/10.63090/EIJDSML/3139.7344.0015

Abstract

Wind direction is an inherently circular variable, and conventional linear statistical techniques are therefore inadequate for describing its variability and its evolution over time. This study examined the temporal variation of nocturnal wind direction using high-frequency meteorological observations recorded at monitoring stations across India during August 2023. Observations were pooled across stations and assigned to eight consecutive hourly intervals spanning 21:00 to 05:00 local time. Wind direction was treated as the principal circular variable and wind speed as the associated linear variable, from which wind power density was derived. Equality of the eight directional distributions was tested using the Mardia–Watson–Wheeler uniform-scores test, a non-parametric procedure for comparing several circular populations. Directional concentration was quantified by the mean resultant length, the association between wind speed and wind direction by Mardia's circular–linear correlation coefficient, and the directional structure of wind occurrence and wind energy by rose diagrams. The eight hourly wind-direction distributions differed significantly ($W = 1916.26$, $df = 14$, $p < 0.001$), indicating a systematic reorganisation of nocturnal wind circulation rather than random directional fluctuation. Directional concentration was greatest during 21:00–22:00 ($\bar{R} = 0.6383$), fell to a minimum of approximately 0.08 between 23:00 and 02:00, and recovered only partially by 04:00–05:00 ($\bar{R} = 0.2492$). The circular–linear correlation coefficient, which measures strength of association only and does not carry a sign, was 0.3060 for the pooled data. Interval-wise values were highest early in the night (0.4722 at 21:00–22:00) and again at 23:00–00:00 (0.4568), reached a minimum at 01:00–02:00 (0.1929), and recovered to 0.3212 by 04:00–05:00; the association therefore weakened around the middle of the night rather than declining steadily. These results are relevant to wind-resource assessment, boundary-layer meteorology, aviation operations and short-term weather forecasting.

Keywords:- Circular Statistics, Wind Direction, Mardia–Watson–Wheeler Uniform-Scores Test, Circular–Linear Correlation, Rose Diagram, Wind Power Density, Nocturnal Boundary Layer, Non-Parametric Inference

I. INTRODUCTION

Wind is among the most dynamic atmospheric variables influencing renewable energy generation, weather prediction, environmental monitoring, aviation safety, agricultural planning and climate research. Wind speed is conventionally treated as a linear variable, whereas wind direction is a circular variable measured on the closed interval from 0° to 360°. Observations near 0° and 360° are physically adjacent although numerically distant, and ordinary linear procedures are consequently unsuitable for directional analysis [1].

Circular statistics provides a mathematical framework designed for angular observations. Unlike linear methods, circular techniques respect the periodicity of directional measurements and yield appropriate estimators of mean direction, concentration, dispersion and dependence [2], [3]. Over the past five decades these methods have become established in meteorology, oceanography, ecology, geology and the biomedical sciences, wherever directional measurements arise [4], [5].

The growing availability of high-resolution meteorological data has created opportunities to study temporal wind variability in fine detail. Nocturnal wind behaviour is of particular interest because the atmospheric boundary layer evolves continuously after sunset: surface cooling, the formation of a nocturnal inversion, the decoupling of the surface layer from the residual layer, and local topographic forcing all act to modify wind circulation [6]. Improved understanding of these changes supports wind-resource assessment, turbine siting, numerical weather prediction and safer night-time aviation operations [7].

India offers a varied environment for wind studies, with diverse climatic zones, an extensive coastline, mountainous terrain and a strong seasonal monsoon circulation. Although many investigations have characterised mean wind speed and its probability distribution [8], [9], comparatively few have applied formal circular inference to the temporal behaviour of wind direction.

Traditional summaries of wind direction that use arithmetic averages ignore circular geometry and can be seriously misleading. Directions of 359° and 1°, for example, have an arithmetic mean of 180° although the two observations are almost identical. Circular methods avoid this by representing each observation as a unit vector, so that directional parameters are estimated on the circle rather than on the line [10].

The present investigation adopts a non-parametric circular framework and evaluates temporal wind-direction variability using the Mardia–Watson–Wheeler uniform-scores test [11], [4]. The procedure compares several circular populations without assuming any particular directional distribution and is therefore well suited to meteorological observations that may be multimodal or otherwise irregular. In addition to hypothesis testing, the study quantifies directional concentration, examines the association between wind speed and wind direction through circular–linear correlation [12], and characterises the directional distribution of wind occurrence, wind speed and wind power density using rose diagrams.

The primary objective was to determine whether the hourly wind-direction distributions observed during the night differ significantly from one another. Secondary objectives were to quantify directional concentration within each hourly interval, to assess how the strength of the speed–direction association changes through the night, and to evaluate the implications of nocturnal directional variability for renewable energy assessment and meteorological forecasting.

II. LITERATURE REVIEW

The statistical analysis of directional data has developed substantially since the systematic treatments of Mardia [4] and Fisher [2]. These works established probability distributions on the circle, estimation procedures, hypothesis tests and regression models specifically constructed for angular observations, and demonstrated that Euclidean methods frequently produce biased or uninterpretable results when applied to circular variables.

Mardia [12] introduced measures of association between circular and linear variables, extending correlation analysis into the directional domain. Johnson and Wehrly [13] developed a related family of angular–linear measures and models; the two proposals are closely connected but not identical, and it is therefore necessary for an applied paper to state which coefficient is being reported. Jammalamadaka and Sarma [14] subsequently proposed a correlation coefficient for two angular variables, and Jammalamadaka and Sarma [15] extended the framework to circular regression. Mardia and Jupp [1] synthesised these developments into a unified account of directional statistics that remains a principal reference, while Jammalamadaka and SenGupta [3] gave particular attention to non-parametric procedures for situations in which distributional assumptions cannot be justified. Pewsey et al. [5] provide a modern computational treatment.

Multi-sample testing on the circle has a similarly long history. Wheeler and Watson [11] proposed a distribution-free two-sample test based on uniform scores, which Mardia [4] generalised to several samples; the resulting Mardia–Watson–Wheeler statistic is asymptotically chi-square distributed and requires no assumption of a von Mises or any other parametric form. Zar [10] discusses its practical use and the sample-size conditions under which the asymptotic approximation is adequate.

Within meteorology, circular methods have been applied to wind climatology, atmospheric circulation, boundary-layer dynamics and renewable-energy resource assessment. Rose diagrams remain the most widely used visualisation because they display directional frequency and magnitude simultaneously, and when combined with wind power density they identify the sectors that dominate the available resource [16], [7]. Standardised

measurement and averaging conventions for wind observations are set out by the World Meteorological Organization [17]; Wilks [18] provides the general statistical background for atmospheric applications.

Boundary-layer studies establish the physical mechanisms relevant to the present analysis. Stull [6] describes how the collapse of turbulence after sunset, the formation of a stable nocturnal boundary layer and the frequent development of a low-level jet alter both the mean wind and its directional steadiness through the night, with a partial return to mixed conditions near sunrise.

Against this background, applied wind studies have relied predominantly on descriptive wind roses and on fitted wind-speed distributions [8], [9], while formal multi-sample circular tests appear comparatively rarely in the applied wind literature even though the necessary methodology has been available for decades [4], [10]. The present study addresses this gap by integrating non-parametric circular hypothesis testing, circular–linear correlation, wind power density estimation and directional visualisation within a single analysis of nocturnal wind variability over India. It is a temporal study: observations are pooled across stations, and no spatial or station-level comparison is attempted (see Section V-H).

III. MATERIALS AND METHODS

A. Study Design

A quantitative, observational design was adopted to investigate the temporal variability of nocturnal wind direction using circular statistical techniques. Because wind direction is angular, classical linear methods cannot adequately characterise its distribution, and non-parametric circular methods were therefore used to test whether directional distributions differ across night-time intervals. The analytical framework combined descriptive circular statistics, circular hypothesis testing, graphical visualisation, circular–linear correlation and wind power density estimation.

B. Data Source and Variables

Meteorological observations were obtained from the Indian Weather Repository, a publicly available compilation of high-resolution weather records collected at monitoring stations across India. The analysis was restricted to observations recorded during August 2023 between 21:00 and 05:00 local time. After the exclusions described in Section III-C, approximately 24,000 observations were retained. Station identifiers were not used as an analytical factor; all records were pooled, so the study characterises the aggregate nocturnal behaviour of the network rather than the behaviour of any individual station.

Table 1. Variables used in the analysis.

Variable	Type	Unit	Role in the analysis
Wind direction	Circular	Degrees (0°–360°)	Principal response variable
Wind speed	Linear	km h^{-1}	Circular–linear correlation; input to wind power density
Wind power density	Derived	W m^{-2}	Wind-energy assessment

Wind direction follows the meteorological convention, measured clockwise from true north as the direction from which the wind blows [17]. Temperature was recorded in the source data but was not used in any analysis reported here and is therefore not listed; see Section V-H.

C. Data Pre-Processing

1. Exclusion of Incomplete Records:

Records with a missing wind-direction or wind-speed value were excluded, since both quantities are required for the circular–linear analysis and for the derivation of wind power density. Calm observations, for which wind direction is undefined, were also excluded.

2. Direction Normalisation:

All wind-direction measurements were reduced modulo 360° so that every observation satisfied

$$0° \leq \theta < 360° \quad (1)$$

ensuring compatibility with the circular computations that follow.

3. Temporal Segmentation:

To characterise the evolution of the nocturnal boundary layer, the retained observations were assigned to eight consecutive one-hour groups (Table 2). Hourly resolution was chosen because it is fine enough to resolve the evening transition and the post-midnight period while leaving each group large enough for the asymptotic

distribution of the test statistic to apply; the requirement, stated in Section III-E3, is that every group contain at least about ten observations, which is satisfied here by a wide margin.

Table 2. Definition of the eight hourly groups (local time).

Group	Time interval	Group	Time interval
G1	21:00–22:00	G5	01:00–02:00
G2	22:00–23:00	G6	02:00–03:00
G3	23:00–00:00	G7	03:00–04:00
G4	00:00–01:00	G8	04:00–05:00

D. Circular Statistical Framework

Wind direction is measured on a circular scale, so that 0° and 360° denote the same direction and the ordinary arithmetic of the real line does not apply. Each observation θ is therefore represented as the unit vector $(\cos\theta, \sin\theta)$ on the unit circle, and all summary quantities are computed from these vectors.

1. Mean Direction:

For a sample of n directions $\theta_1, \theta_2, \ldots, \theta_n$, the Cartesian components of the resultant vector are

$$C = \sum_{i=1}^{n} \cos\theta_i \,, \quad S = \sum_{i=1}^{n} \sin\theta_i \tag{2}$$

and the sample mean direction is

$$\bar{\theta} = \arctan\left(\frac{S}{C}\right) \tag{3}$$

where the quadrant of $\bar{\theta}$ is determined from the signs of S and C (equivalently, by the two-argument arctangent). Directions were reported to the nearest degree.

2. Resultant Length and Directional Concentration:

The length of the resultant vector is

$$R = \sqrt{C^2 + S^2} \tag{4}$$

and the mean resultant length, which is the standard measure of directional concentration, is

$$\bar{R} = \frac{R}{n}, \quad 0 \le \bar{R} \le 1 \tag{5}$$

A value of $\bar{R}$ close to 1 indicates that the directions are tightly clustered about a single preferred direction, whereas a value close to 0 indicates that they are widely dispersed and that no direction predominates. Because verbal descriptions of $\bar{R}$ are otherwise arbitrary, the fixed bands set out in Table 3 were adopted in advance and applied uniformly throughout the paper.

Table 3. Bands used to describe the mean resultant length. The bands are conventions adopted for consistent reporting and have no inferential status.

Mean resultant length	Descriptive label
$\bar{R} < 0.10$	Very low concentration
$0.10 \le \bar{R} < 0.20$	Low concentration
$0.20 \le \bar{R} < 0.40$	Moderate concentration
$0.40 \le \bar{R} < 0.60$	High concentration
$\bar{R} \ge 0.60$	Very high concentration

E. The Mardia–Watson–Wheeler Uniform-Scores Test

The principal inferential procedure was the Mardia–Watson–Wheeler (MWW) uniform-scores test, a non-parametric test of the equality of several circular populations [11], [4], [1]. Unlike analysis of variance, which presumes linear measurements, and unlike the Watson–Williams test, which assumes von Mises populations with a common concentration, the MWW test makes no distributional assumption and preserves the circular geometry of the observations.

1. Hypotheses:

Let Fk(θ) denote the directional distribution function of group k, for k = 1, …, g. The hypotheses are

$$H_0 : F_1(\theta) = F_2(\theta) = \cdots = F_g(\theta) \tag{6}$$

against the alternative HA that at least one group distribution differs from the others. In the present application g = 8, so the null hypothesis states that wind direction has the same distribution in every hour of the night.

2. Circular Ranks and Uniform Scores:

The N observations from all groups are pooled and ordered around the circle from an arbitrary origin. Let

$$R_i = \text{rank of } \theta_i \text{ in the pooled sample}, \quad i = 1, \ldots, N \tag{7}$$

so that Ri takes the values 1, …, N. Each rank is then converted into a uniform score, which is itself an angle:

$$u_i = \frac{2\pi R_i}{N+1}, \quad 0 < u_i < 2\pi \tag{8}$$

The factor 2π is essential: it places the scores as equally spaced positions on the circumference of the unit circle, so that the sine and cosine components used in the test statistic are well defined. Ranking removes any dependence on the underlying directional distribution, because under H0 the uniform scores of any one group are a random subset of the N equally spaced points on the circle, whatever that distribution may be. Ties, which arise when direction is recorded to a coarse resolution, were resolved by assigning the mean of the tied ranks.

3. Test Statistic and Reference Distribution:

For each group k of size nk, the components of the resultant of the uniform scores belonging to that group are

$$C_k = \sum_{i \in G_k} \cos u_i, \quad S_k = \sum_{i \in G_k} \sin u_i \tag{9}$$

and the Mardia–Watson–Wheeler statistic is

$$W = 2 \sum_{k=1}^{g} \frac{C_k^2 + S_k^2}{n_k} \tag{10}$$

Under H0, and provided the groups are reasonably large, W is distributed approximately as chi-square with 2(g − 1) degrees of freedom,

$$W \sim \chi^2_{2(g-1)} \tag{11}$$

With g = 8 groups the reference distribution therefore has 14 degrees of freedom. The chi-square approximation is generally regarded as adequate when every group contains at least about ten observations [1], [10]; the smallest group in this study contains several hundred, so the approximation is not in question here. Large values of W indicate that the group resultants of the uniform scores point in systematically different directions, and hence that the group distributions differ.

F. Circular–Linear Correlation

Wind direction is circular and wind speed is linear, so an ordinary product-moment correlation between them is not meaningful. The coefficient of Mardia [12] was used. For a linear variable x and a circular variable θ, let rxc = corr(x, cos θ), rxs = corr(x, sin θ) and rcs = corr(cos θ, sin θ) be the three ordinary product-moment correlations. Mardia's squared circular–linear correlation is

$$r_{cl}^2 = \frac{r_{xc}^2 + r_{xs}^2 - 2\, r_{xc} r_{xs} r_{cs}}{1 - r_{cs}^2} \tag{12}$$

Throughout this paper rcl denotes the positive square root of expression (12), so that 0 ≤ rcl ≤ 1. Because the quantity is assembled from squared linear–circular correlations it has no sign and admits no direction of association: it measures the strength of the dependence between wind speed and wind direction and nothing more. This is the coefficient of Mardia [12], not the angular–linear measure of Johnson and Wehrly [13], which appears in the literature under a similar name and is defined differently.

Under the null hypothesis that x and θ are independent, and for large n,

$$n\, r_{cl}^2 \sim \chi^2_2 \tag{13}$$

which was used to assess the significance of the coefficients reported in Section IV-F [1]. Correlations were computed for the pooled data and separately within each hourly group.

G. Wind Power Density

Wind power density (WPD) is the kinetic energy flux carried by the wind per unit area normal to the flow,

$$P = \frac{1}{2} \rho V^3 \quad (14)$$

where P is the wind power density (W m^{-2}), ρ the air density (kg m^{-3}) and V the wind speed (m s^{-1}); wind speeds recorded in km h^{-1} were converted before evaluation. A constant air density of ρ = 1.225 kg m^{-3}, the standard value at sea level and 15 °C, was assumed [16]. The cubic dependence on speed means that modest changes in wind speed produce large changes in available power, so that the directional sectors carrying the strongest winds dominate the resource even when they are not the most frequent.

H. Software

All computations were carried out using standard circular-statistics routines [19], [5], with results cross-checked against the circular data analysis procedure documented by NCSS [20]. The analysis comprised descriptive circular summaries, the Mardia–Watson–Wheeler test, circular–linear correlation, wind power density calculation, sector-wise frequency tabulation and rose-diagram construction.

IV. RESULTS

A. Descriptive Overview

The analysed dataset comprised approximately 24,000 nocturnal observations distributed across the eight consecutive hourly intervals between 21:00 and 05:00. Wind direction was treated as the principal circular variable; wind speed and the derived wind power density were treated as linear variables. Directional concentration varied markedly across the night: it was highest during 21:00–22:00, fell to its lowest values during 23:00–00:00 and 01:00–02:00, and increased again towards dawn. The sections that follow present the formal test of equality of the hourly distributions, the concentration values themselves, the distributions of wind speed and wind power density, the directional structure of the data, and the association between wind speed and wind direction.

B. Equality of the Hourly Wind-Direction Distributions

The Mardia–Watson–Wheeler uniform-scores test was applied to the eight hourly groups. The results are given in Table 4.

Table 4. Mardia–Watson–Wheeler uniform-scores test of the equality of the eight hourly wind-direction distributions.

Quantity	**Value**
Number of groups, g	8
Test statistic, W	1916.26
Degrees of freedom, 2(g − 1)	14
Critical value, χ^2_{14} at α = 0.05	23.68
p-value	< 0.001

The observed statistic exceeds the 5% critical value by almost two orders of magnitude, so the null hypothesis of a common directional distribution across the eight hourly groups is rejected decisively. Wind direction therefore changes systematically through the night rather than fluctuating at random about a fixed distribution. It should be noted that with samples of this size even modest departures from equality will be detected; the magnitude of the differences is described by the concentration values in Section IV-C and by the rose diagrams in Section IV-E, and it is these that carry the meteorological interpretation.

C. Directional Concentration

The mean resultant length $\bar{R}$ was computed for each hourly group (Table 5), using the descriptive bands defined in Table 3.

Table 5. Mean resultant length of wind direction by hourly interval. Labels follow the bands defined in Table 3.

Group	Time interval	Mean resultant length, $\bar{R}$	Description
G1	21:00–22:00	0.6383	Very high concentration
G2	22:00–23:00	0.3860	Moderate concentration
G3	23:00–00:00	0.0840	Very low concentration
G4	00:00–01:00	0.1567	Low concentration
G5	01:00–02:00	0.0822	Very low concentration
G6	02:00–03:00	0.1985	Low concentration
G7	03:00–04:00	0.1960	Low concentration
G8	04:00–05:00	0.2492	Moderate concentration

Directional consistency was greatest in the first hour after 21:00 and then declined sharply, reaching a minimum near 0.08 at 23:00–00:00 and again at 01:00–02:00. From 02:00 onwards the concentration recovered gradually, reaching 0.2492 by 04:00–05:00, a value still well below that of the early evening. Wind direction is therefore substantially more variable after approximately 23:00, with only partial reorganisation before dawn.

D. Distribution of Wind Speed and Wind Power Density

Table 6 summarises the upper part of the pooled distributions of wind speed and of the derived wind power density. Percentiles are expressed as exceedance percentiles: the tabulated value is the level exceeded by the stated percentage of observations, so the values decrease as the percentage increases.

Table 6. Exceedance percentiles of wind speed and of wind power density for the pooled nocturnal observations.

Exceedance percentile (%)	Wind speed (km h^{-1})	Wind power density (W m^{-2})
2	8.72	969.50
4	7.86	683.35
6	7.41	549.39
8	7.08	470.21
10	6.81	416.36

Wind speed declines smoothly across the upper tail, which is consistent with a continuous unimodal speed distribution of the kind usually fitted in wind-resource work [8], [9] and indicates that the nocturnal wind field contained no abrupt discontinuities during the study period. Wind power density falls far more steeply than wind speed over the same range, as the cubic relationship in Equation (14) requires: the strongest 2% of observations carry more than twice the power density of the strongest 10%. The energy available in the nocturnal wind field is therefore concentrated in a small fraction of the observations.

E. Directional Structure of the Nocturnal Wind Field

Fig. 1 shows wind-direction frequency roses for four representative hourly intervals spanning the night. The contrast between the panels illustrates numerically what Table 5 summarises: in the 21:00–22:00 panel the distribution is dominated by a narrow group of westerly and west-south-westerly sectors, whereas by 01:00–02:00 and 03:00–04:00 the petals are distributed around most of the compass, with only a weak preference for the west-south-west and east-north-east.

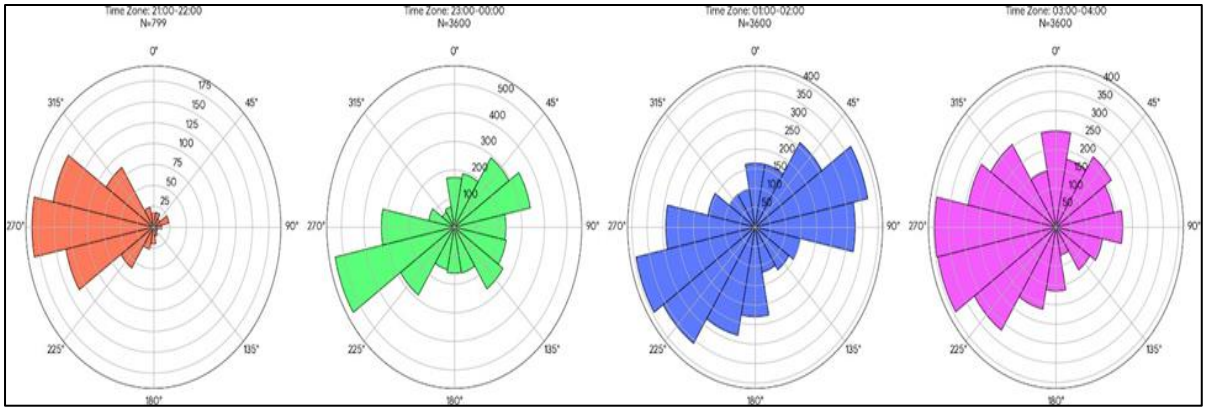

Fig. 1: Wind-direction frequency roses for four representative hourly intervals: 21:00–22:00, 23:00–00:00, 01:00–02:00 and 03:00–04:00.

Each rose is constructed from 16 equal sectors of 22.5°, with the first sector beginning at 0°. Direction is measured clockwise from true north (0° at the top of each panel) and denotes the direction from which the wind blows. The radial axis gives the number of observations falling in each sector, with the scale labelled on each panel; note that the radial scale differs between panels. The number of observations contributing to each rose is shown above the panel. Each rose is based on the observations of that interval only.

Table 7 gives the sector-wise distribution of relative wind occurrence, relative wind speed and relative wind power over the first five 30° sectors of the pooled data. The three columns are normalised so that each sums to one over the full circle, and they answer three different questions: how often the wind blows from a sector, how strong it is when it does, and how much energy that sector contributes.

Table 7. Sector-wise relative wind frequency, relative wind speed and relative wind power for the pooled nocturnal observations, 30° sectors labelled by their lower bound. Values for the sectors from 150° to 330° are omitted here for brevity.

Sector (°)	Relative frequency	Relative wind speed	Relative wind power
0–30	0.06	0.05	0.02
30–60	0.06	0.05	0.06
60–90	0.08	0.10	0.13
90–120	0.08	0.10	0.13
120–150	0.05	0.04	0.03

Within the sectors shown, wind occurrence is not uniform across direction, and the 60–120° sectors stand out on all three measures. These sectors account for a slightly above-average share of observations, a larger share of the wind speed, and a larger share still of the wind power: their contribution to available energy is roughly six times that of the 0–30° sector although they occur only about a third more often. This amplification is the practical consequence of the cubic term in Equation (14).

The three columns of Table 7 are mutually consistent in the location of their dominant sectors, but differ in how sharply that dominance is expressed. The frequency column is the flattest of the three, the speed column intermediate, and the power column the most strongly peaked, with the greatest energy contribution falling in the easterly sectors between 60° and 120°.

F. Circular–Linear Correlation Between Wind Speed and Wind Direction

Mardia's circular–linear correlation coefficient was computed for the pooled data and for each of the eight hourly intervals (Table 8). The coefficient measures the strength of the association only and does not indicate its direction, so the intervals are compared by the magnitude of the coefficient alone.

Table 8. Mardia's circular–linear correlation between wind speed and wind direction, by hourly interval and for the pooled data. The coefficient lies between 0 and 1 and measures strength of association only; it carries no sign. Descriptive labels follow the bands of Table 3, applied to the correlation coefficient on the same numerical scale.

Group	Time interval	r_{cl}	Strength of association
G1	21:00–22:00	0.4722	High
G2	22:00–23:00	0.3924	Moderate
G3	23:00–00:00	0.4568	High
G4	00:00–01:00	0.3060	Moderate
G5	01:00–02:00	0.1929	Low
G6	02:00–03:00	0.2948	Moderate
G7	03:00–04:00	0.2748	Moderate
G8	04:00–05:00	0.3212	Moderate
—	Pooled, whole night	0.3060	Moderate

Note. Significance was assessed by the asymptotic test of Equation (13). Independence between wind speed and wind direction is rejected at $p < 0.001$ for the pooled sample and for every hourly interval. Interval-wise test statistics are not tabulated, because at these sample sizes significance is assured for any coefficient of the magnitudes shown and carries no interpretative weight; the coefficients are therefore discussed for their size rather than their significance.

The pooled coefficient of 0.3060 indicates an association of moderate strength between wind speed and wind direction over the night as a whole. The interval-wise values do not decline steadily. The strongest association occurs at 21:00–22:00 (0.4722); the coefficient falls at 22:00–23:00 (0.3924) but rises again to 0.4568 at 23:00–00:00, close to the maximum; it then decreases to a minimum of 0.1929 at 01:00–02:00; and it recovers over the last three intervals, from 0.2948 through 0.2748 to 0.3212 at 04:00–05:00. The pattern is therefore one of a strong association in the first half of the night, a pronounced weakening centred on 01:00–02:00, and a partial recovery towards dawn, rather than a monotone decline. Every coefficient in the table differs from zero at the 0.1% level, which is unsurprising given the sample sizes involved and which is why the discussion that follows turns on the magnitude of the coefficients and on their ordering through the night rather than on their significance.

Fig. 2 presents the two temporal trajectories side by side: the mean resultant length of Table 5 and the circular–linear correlation of Table 8, plotted on a common vertical scale across the eight hourly intervals. The two quantities behave differently. Directional concentration falls steeply and early, losing almost nine tenths of its initial value by 23:00–00:00, and its recovery after 02:00 is slight. The correlation declines later, less steeply and by much less, and it is still close to its early-evening level at 23:00–00:00, the very interval at which concentration reaches its first minimum. The weakening of the speed–direction association is therefore not simply a by-product of the loss of directional concentration, since the two minima do not coincide.

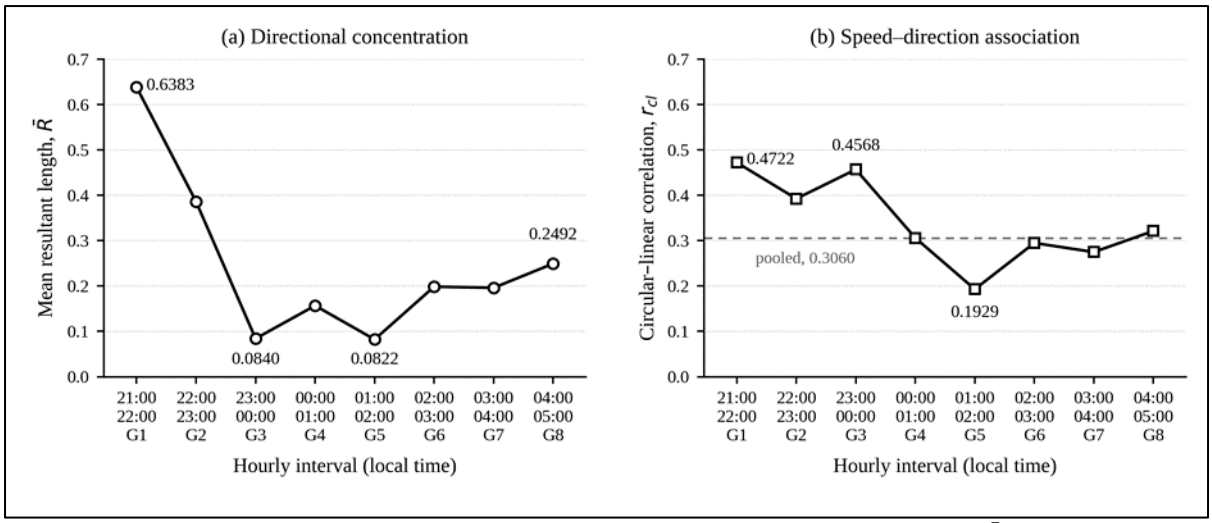

Fig. 2: Nocturnal trajectories of (a) directional concentration, measured by the mean resultant length $\bar{R}$, and (b) the strength of the speed–direction association, measured by Mardia's circular–linear correlation coefficient.

Both panels share a common vertical scale from 0 to 0.70 and are plotted against the eight hourly intervals G1 to G8 defined in Table 2. Selected values are annotated. The dashed horizontal line in panel (b) marks the

pooled coefficient of 0.3060 for the night as a whole; no corresponding pooled value is shown in panel (a), since the mean resultant length of the pooled sample is not the average of the group values. Panel (a) is plotted from Table 5 and panel (b) from Table 8.

V. DISCUSSION

This study examined the temporal evolution of nocturnal wind direction over India using non-parametric circular statistical techniques. Unlike linear analyses, which disregard the periodicity of angular measurements, the circular framework provides a mathematically appropriate representation of wind direction and of its variation through the night. The combination of the Mardia–Watson–Wheeler uniform-scores test, descriptive circular statistics, circular–linear correlation and rose-diagram visualisation supported a coherent assessment of nocturnal wind behaviour.

The test yielded a large statistic ($W = 1916.26$, df = 14, $p < 0.001$), providing strong evidence that the wind-direction distributions differ across the eight nocturnal intervals. Because the test is distribution-free, this conclusion does not rest on any assumption of von Mises or other parametric form, which matters here because several of the hourly distributions are visibly multimodal (Fig. 1) and would violate such assumptions.

A. Temporal Evolution of Wind Direction

The most striking feature of the results is the variation in directional concentration through the night. The highest value ($\bar{R}$ = 0.6383) occurred at 21:00–22:00, indicating that airflow was well organised in the early evening. Concentration then fell sharply, to below 0.10 at 23:00–00:00 and again at 01:00–02:00, before recovering to 0.2492 by 04:00–05:00. The trajectory is shown in panel (a) of Fig. 2.

This behaviour is consistent with established boundary-layer meteorology [6]. In the early evening, residual momentum from the daytime mixed layer maintains relatively coherent flow. As radiative cooling proceeds, a stable nocturnal boundary layer develops, turbulence is suppressed, and the surface layer decouples from the flow above; local terrain, land-cover heterogeneity and thermally driven mesoscale circulations then exert proportionately greater influence on the near-surface wind, and directional variability increases. Towards dawn, the onset of weak mixing begins to re-establish a more organised flow, which accounts for the partial recovery of concentration observed after 02:00. The fact that concentration does not return to its early-evening value before 05:00 indicates that the reorganisation is incomplete at the end of the analysed window.

One qualification should be attached to this interpretation. Because observations from all stations are pooled, part of the low concentration observed in the middle of the night reflects genuine within-station directional variability and part reflects differences between stations that happen to be captured in the same hourly bin. The pooled analysis cannot separate these two contributions (Section V-H).

B. Interpretation of the Graphical Evidence

The rose diagrams of Fig. 1 convey directional information that numerical summaries alone cannot, and Fig. 2 sets the two temporal trajectories against one another. Fig. 1 shows the change in directional structure directly: the tight westerly cluster of the early evening gives way to a distribution spread over most of the compass after midnight, which is the visual counterpart of the fall in mean resultant length reported in Table 5. The panels also show that the loss of concentration is not simply a broadening of a single mode but the emergence of several competing directional sectors, a pattern that would be poorly summarised by any single mean direction.

Table 7 separates three distinct aspects of the directional structure. The relative frequency shows the sectors from which the wind most often blew; the relative wind speed shows where the strongest winds came from; and the relative wind power, through the cubic weighting of Equation (14), shows which sectors actually supplied the energy. The agreement between the three in the location of the dominant sectors is informative in itself: the directions that occur most often are also, in this dataset, the directions that carry the strongest winds, so that occurrence and energy reinforce rather than offset one another. Such agreement should not be assumed in general, since a sector that is infrequent but consistently windy can dominate the energy resource, and it is precisely this possibility that makes the separate tabulation of wind power worthwhile [16].

C. Circular–Linear Correlation

The strength of the association between wind speed and wind direction was not constant through the night. It was greatest at 21:00–22:00 and 23:00–00:00, weakest at 01:00–02:00, and partially restored by 04:00–05:00. A physically plausible reading is that wind speed is most strongly conditioned on direction while a well-defined flow regime persists: when a single sector dominates, the speeds associated with that sector dominate too. During the most stable part of the night, when the surface layer is decoupled and several directional sectors coexist, the link between the direction of the flow and its speed is correspondingly diluted, and the coefficient falls. The

recovery before dawn parallels the recovery in directional concentration and is consistent with the same mechanism.

Two cautions apply. First, the coefficient is a measure of strength alone; it takes values between 0 and 1 by construction and cannot indicate whether higher speeds are associated with any particular range of directions. That question requires inspection of the relative wind speed by sector in Table 7, not the correlation coefficient. Second, the two quantities compared here - concentration and correlation - are related but not equivalent, and the correspondence between them across the night should be treated as suggestive rather than as evidence of a causal link.

D. Implications for Wind Energy

Directional persistence is directly relevant to renewable-energy planning, since turbines extract power most efficiently when the flow is steady and well aligned. The results indicate that the nocturnal period over the observed network is not homogeneous in this respect: the early evening offers well-organised flow, whereas the hours around and after midnight are directionally unsettled. Yaw systems, which must track a shifting wind vector, will consequently be more active during the middle of the night, with implications for both energy capture and mechanical wear [7].

The sector-wise power figures reinforce the point that energy production is concentrated within a comparatively narrow directional range. Because power varies with the cube of speed, modest directional shifts away from the dominant sectors can translate into substantial reductions in output. Incorporating circular statistical measures alongside the conventional wind-speed distribution in resource assessment would therefore give a more complete picture of expected turbine performance and annual energy yield than speed statistics alone [9].

E. Comparison With Previous Work

The methodological findings are consistent with the established literature. Mardia [12], Fisher [2] and Mardia and Jupp [1] demonstrated that directional variables require dedicated procedures, and Jammalamadaka and SenGupta [3] set out the advantages of non-parametric methods when distributional assumptions cannot be sustained. The multimodal hourly distributions observed here are a concrete instance of the situation those authors describe, in which parametric circular models would be difficult to justify.

From a meteorological standpoint, the increase in directional variability after the evening transition and its partial reversal before sunrise correspond closely to the behaviour of the stable nocturnal boundary layer documented by Stull [6]. The present analysis adds formal statistical support to this qualitative picture. Compared with the many applied studies that characterise wind direction using descriptive roses alone, the present work couples visualisation with a formal test of distributional equality, which converts an impression of temporal change into a testable and tested claim.

F. Practical Applications

The findings bear on several areas of application:

- Renewable energy: identification of dominant sectors and of the hours in which those sectors are least stable supports turbine orientation, micrositing and wind-farm layout.
- Weather forecasting: systematic hourly change in directional distributions provides a target against which short-range models of nocturnal circulation can be evaluated.
- Aviation safety: the hours of lowest directional concentration are those in which runway selection and crosswind assessment are most likely to change at short notice.
- Environmental monitoring: directional variability governs the dispersion of pollutants released at night, when the stable boundary layer already limits vertical mixing.
- Climate research: circular summaries of the kind used here provide a basis for detecting long-term change in directional climatology.

G. Strengths of the Study

The study has several methodological strengths: it applies statistical procedures constructed specifically for angular observations; it uses a non-parametric test that avoids restrictive distributional assumptions, which is material given the multimodality evident in Fig. 1; it integrates descriptive statistics, formal inference, correlation analysis and graphical display within one framework; it draws on a large, high-frequency dataset; and its conclusions are supported by both inferential and visual evidence.

H. Limitations

Several limitations should be acknowledged. First, and most importantly, observations from all monitoring stations were pooled, and no station-level or regional analysis was undertaken; the results therefore describe the aggregate nocturnal behaviour of the network and cannot be attributed to any particular location or climatic zone. Pooling also means that between-station heterogeneity contributes to the measured within-hour directional dispersion, and the two sources cannot be separated in the present design.

Second, the analysis covers August 2023 only, a single month within the south-west monsoon, so the results should not be extrapolated to other seasons. Third, only the eight hours from 21:00 to 05:00 were examined; the full diurnal cycle, including the morning and evening transitions, lies outside the study. Fourth, humidity, atmospheric pressure and terrain elevation were not incorporated. Fifth, air temperature was recorded in the source data but was not used: wind power density was computed with a constant air density of 1.225 kg m^{-3} rather than with the observation-specific density obtainable from the ideal gas law, $\rho = p/(RT)$, where R is the specific gas constant for air, which would be the natural use of the temperature record and would refine the energy estimates, particularly across the range of station elevations represented in the data. Finally, with samples of this size the test has very high power, so statistical significance should not be read as a measure of the practical importance of the observed differences.

I. Future Research

The present work could be extended in several directions:

- Analysis of seasonal and annual variability in wind direction, and comparison of monsoon with non-monsoon circulation;
- Station-level and regional analysis, including formal comparison of climatic zones and the use of spatial interpolation and geostatistical methods to map directional statistics [21], [22];
- Extension to the complete diurnal cycle, so that the morning and evening transitions can be characterised alongside the nocturnal period;
- Circular regression and bayesian directional models relating wind direction to temperature, pressure and stability indices [15];
- Integration of satellite observations and reanalysis products to improve spatial coverage; and
- Assessment of long-term change in directional wind climatology under climate-change scenarios.

VI. CONCLUSION

This study applied a circular statistical framework to the temporal variability of nocturnal wind direction over India during August 2023. By representing directions as unit vectors rather than as ordinary numbers, the approach avoids the distortions that arise when linear methods are applied to angular data and provides a statistically defensible description of nocturnal wind behaviour.

The Mardia–Watson–Wheeler uniform-scores test showed that the eight hourly wind-direction distributions are not identical ($W = 1916.26$, df = 14, $p < 0.001$), indicating systematic temporal reorganisation of nocturnal circulation rather than random fluctuation. Directional concentration was highest in the early evening ($\bar{R} = 0.6383$ at 21:00–22:00), fell to approximately 0.08 between 23:00 and 02:00, and recovered only partially by dawn ($\bar{R} = 0.2492$ at 04:00–05:00). This trajectory is consistent with the development of a stable nocturnal boundary layer and its gradual erosion before sunrise.

The circular–linear correlation analysis indicated an association of moderate strength between wind speed and wind direction, the coefficient measuring strength alone and carrying no sign. Its value was highest at 21:00–22:00 (0.4722), remained high at 23:00–00:00 (0.4568), reached a minimum at 01:00–02:00 (0.1929) and recovered to 0.3212 by 04:00–05:00; the association therefore weakened around the middle of the night and strengthened again towards dawn, rather than declining steadily. The rose diagrams of Fig. 1 showed the loss and partial recovery of directional structure through the night, and the sector-wise summaries showed that a limited group of directional sectors accounted simultaneously for the greatest wind occurrence, the strongest winds and the largest share of available wind energy.

The findings have practical implications for wind-resource assessment, renewable-energy planning, short-range forecasting, aviation operations and environmental dispersion modelling. Their scope is nonetheless bounded: they describe a single month, a single part of the diurnal cycle, and a network of stations treated in aggregate. Extending the analysis to the full year and the full diurnal cycle, and disaggregating it to the level of

individual stations and climatic zones, are the natural next steps, alongside the use of circular regression and Bayesian directional models to relate wind direction to the meteorological variables that drive it.

REFERENCES

[1] K. V. Mardia and P. E. Jupp, Directional Statistics. Chichester, U.K.: Wiley, 2000.
[2] N. I. Fisher, Statistical Analysis of Circular Data. Cambridge, U.K.: Cambridge Univ. Press, 1993.
[3] S. R. Jammalamadaka and A. SenGupta, Topics in Circular Statistics. Singapore: World Scientific, 2001.
[4] K. V. Mardia, Statistics of Directional Data. London, U.K.: Academic Press, 1972.
A. Pewsey, M. Neuhäuser, and G. D. Ruxton, Circular Statistics in R. Oxford, U.K.: Oxford Univ. Press, 2013.
[5] R. B. Stull, An Introduction to Boundary Layer Meteorology. Dordrecht, The Netherlands: Kluwer, 1988.
[6] T. Burton, N. Jenkins, D. Sharpe, and E. Bossanyi, Wind Energy Handbook, 3rd ed. Chichester, U.K.: Wiley, 2021.
[7] C. G. Justus, W. R. Hargraves, A. Mikhail, and D. Graber, "Methods for estimating wind speed frequency distributions," J. Appl. Meteorol., vol. 17, no. 3, pp. 350–353, 1978.
[8] J. A. Carta, P. Ramírez, and S. Velázquez, "A review of wind speed probability distributions used in wind energy analysis," Renew. Sustain. Energy Rev., vol. 13, no. 5, pp. 933–955, 2009.
[9] J. H. Zar, Biostatistical Analysis, 5th ed. Upper Saddle River, NJ, USA: Pearson, 2010.
[10] S. Wheeler and G. S. Watson, "A distribution-free two-sample test on a circle," Biometrika, vol. 51, no. 1–2, pp. 256–257, 1964.
[11] K. V. Mardia, "Linear–circular correlation coefficients and rhythmometry," Biometrika, vol. 63, no. 2, pp. 403–405, 1976.
[12] R. A. Johnson and T. E. Wehrly, "Measures and models for angular correlation and angular–linear correlation," J. Roy. Statist. Soc. B, vol. 39, no. 2, pp. 222–229, 1977.
[13] S. R. Jammalamadaka and Y. R. Sarma, "A correlation coefficient for angular variables," in Statistical Theory and Data Analysis II, K. Matusita, Ed. Amsterdam, The Netherlands: North-Holland, 1988, pp. 349–364.
[14] S. R. Jammalamadaka and Y. R. Sarma, "Circular regression," in Statistical Theory and Data Analysis. Utrecht, The Netherlands: VSP, 1993, pp. 109–128.
[15] J. F. Manwell, J. G. McGowan, and A. L. Rogers, Wind Energy Explained: Theory, Design and Application, 2nd ed. Chichester, U.K.: Wiley, 2010.
[16] World Meteorological Organization, "Guide to instruments and methods of observation," WMO, Geneva, Switzerland, WMO-No. 8, 2021.
[17] D. S. Wilks, Statistical Methods in the Atmospheric Sciences, 4th ed. Amsterdam, The Netherlands: Academic Press, 2019.
[18] U. J. Lund, "Circular statistics," California Polytechnic State Univ., San Luis Obispo, CA, USA, 2003.
[19] NCSS, LLC, "Circular data analysis," NCSS Statistical Software, Kaysville, UT, USA, Procedure Documentation, 2023.
[20] N. Cressie, Statistics for Spatial Data, rev. ed. Hoboken, NJ, USA: Wiley, 2015.
[21] P. J. Diggle and P. J. Ribeiro, Model-Based Geostatistics. New York, NY, USA: Springer, 2007.

EDUSCHOOL INTERNATIONAL JOURNAL OF DATA SCIENCE AND MACHINE LEARNING (EIJDSML)

(Open Access, Double-Blind Peer Reviewed Journal)

ISSN Online: 3139-7344



Time-Series Anomaly Detection: Benchmark Flaws, Evaluation Metrics and Strong Baselines

Tintu George

Assistant Professor, Department of BCA AI, Sri Ramakrishna College of Arts & Science, Coimbatore, India

Article information

Received: 6th July 2026
Received in revised form: 10th August 2026
Accepted: 16th September 2026
Available online: 18th October 2026

Volume: 1
Issue: 3
DOI: https://doi.org/10.63090/EIJDSML/3139.7344.0016

Abstract

Dozens of deep architectures for time-series anomaly detection have been published since 2018, most of them reporting higher scores than their predecessors on the same handful of datasets. A series of critical studies now suggests that much of this progress is an artefact. The datasets have flaws that make them easy or mislabelled. The most common scoring protocol, point adjustment, inflates results so strongly that random scores can compete with trained models. And when classical or deliberately simple methods are run under the same conditions, they often match the deep ones. This review assembles that evidence. We describe the method families briefly, then examine the benchmark critiques, the evaluation metrics and their failure modes, and the large comparison studies. A small worked example, constructed by hand and labelled as such, shows the arithmetic of point adjustment. We end with a reporting checklist and with the open question of whether pretrained time-series models change the picture. Apart from that worked example, no new result is reported.

Keywords:- Time Series, Anomaly Detection, Benchmarking, Point Adjustment, Evaluation Metrics, Deep Learning, Baselines

I. INTRODUCTION

A pump vibrates oddly an hour before it fails. A server's latency creeps upward. A card is used in two countries on one afternoon. Each is an anomaly in a time series, and finding such events early is worth money, sometimes lives. The general problem is old and well surveyed [1], [2]. What is new is the volume of deep learning work aimed at it, reviewed in several recent surveys [3], [4], [5].

The typical paper in this literature proposes an architecture, evaluates it on four or five public datasets, and reports an F1 score above the previous best. Read in sequence, the papers describe steady progress. Read alongside a set of critical studies published between 2021 and 2024, they describe something else: a field whose measuring instruments were faulty. This review is about the instruments.

Three questions organise it. Are the benchmark datasets sound? Do the metrics measure detection? And do the complex models beat simple ones when both are tested fairly? We are mostly reporting other people's findings, and we try to state each one no more strongly than its authors did.

II. METHOD FAMILIES IN BRIEF

Classical detectors need little introduction. Isolation forests score a point by how few random splits isolate it [6]. The local outlier factor compares a point's density with that of its neighbours [7], and the one-class support

vector machine learns a boundary around normal data [8]. For sequences, the matrix profile finds discords, the subsequences furthest from their nearest match [9]. Industrial monitoring services were built on statistical forecasting [10] and, at Microsoft, on a spectral-residual method paired with a small convolutional network [11], and streaming detectors were benchmarked on real-time criteria [12].

Deep detectors mostly follow one of two templates. Forecasting models predict the next values and flag large errors; the LSTM detector built for spacecraft telemetry, with its nonparametric threshold, is the reference example [13]. Reconstruction models compress and rebuild a window and flag what they cannot rebuild. This family includes variational autoencoders for seasonal metrics [14], a stochastic recurrent model [15], adversarially trained autoencoders [16], a mixture-model variant [17] and a generative adversarial one [18].

Later designs added structure. Graph-based detectors learn which sensors depend on which and attend over that graph [19], [20]. A hierarchical one-class network works at several time scales [21]. Transformers arrived with an association-based criterion [22] and with adversarial training [23]. Whatever the model, a threshold must turn scores into alarms, and extreme value theory offers one way to set it without assuming a distribution [24].

III. THE BENCHMARKS

A small set of datasets carries most published results. Two come from spacecraft telemetry [13], one from server machines [15], one from a water treatment testbed [25] and one from pooled server metrics [26]. The Numenta benchmark [27] and a collection released by Yahoo serve the univariate case.

Wu and Keogh examined four of these collections closely, namely the Yahoo and Numenta sets, the spacecraft telemetry and the server machine data, and reported four flaws, shown in Fig 1. Many labelled anomalies were trivial, in the sense that a single line of code applied to the raw signal would find them. Anomaly density was unrealistic, with long or closely spaced anomalous stretches. Some ground-truth labels were plainly wrong. And anomalies tended to sit near the end of the series, so that a detector guessing 'late' would do well. Their conclusion was that results on such data cannot support claims of progress, and they released a new archive designed to avoid the four problems [28].

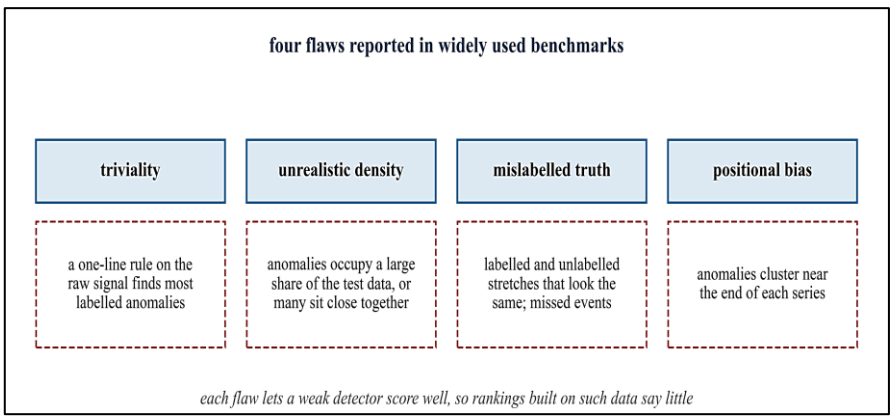

Fig. 1: The four benchmark flaws reported in [28], drawn by the authors.

Others built larger and cleaner suites. TSB-UAD assembled thousands of labelled univariate series from many sources, with synthetic and transformed variants to vary difficulty [29]. TimeSeAD re-examined the multivariate datasets and their evaluation and came to conclusions close to Wu and Keogh's [30]. A 2024 benchmark curated its series, by human inspection aided by detection algorithms, to remove mislabelled and biased cases before comparing methods [31]. Table 1 collects the critiques.

Table 1. Critiques of benchmark data and protocol, as reported by the cited studies.

Issue	Consequence	Proposed remedy	Source
Trivial anomalies	Weak detectors score well	Curated archive	[28]
Unrealistic anomaly density	Inflated recall; easy guessing	One anomaly per test series	[28]
Mislabelled ground truth	True detections counted as errors	Manual curation of labels	[28], [31]
Positional bias	Guessing 'late' is rewarded	Vary anomaly location	[28]
Point adjustment	Random scores rival trained models	Report unadjusted; partial adjustment	[32]
Few, small datasets	Rankings do not generalise	Large, heterogeneous suites	[29], [33]

IV. THE METRICS

A. Point adjustment

Anomalies in time series usually last for a while. Should a detector be required to flag every time step of an event? The authors of an early variational autoencoder detector thought not, reasoning that an operator needs one alert per incident, and they scored a whole anomalous segment as detected if any point inside it was flagged [14]. This protocol, later called point adjustment, was adopted by influential follow-up work [15] and became the default.

The arithmetic deserves a look. Fig 2 shows a hand-made sequence of 24 steps with two anomalous segments. The raw prediction flags two points, one inside the first segment and one false alarm. Its point-wise F1 is 0.14. After adjustment the single hit is expanded to the full eight-step segment, and F1 becomes 0.76. Nothing about the detector changed.

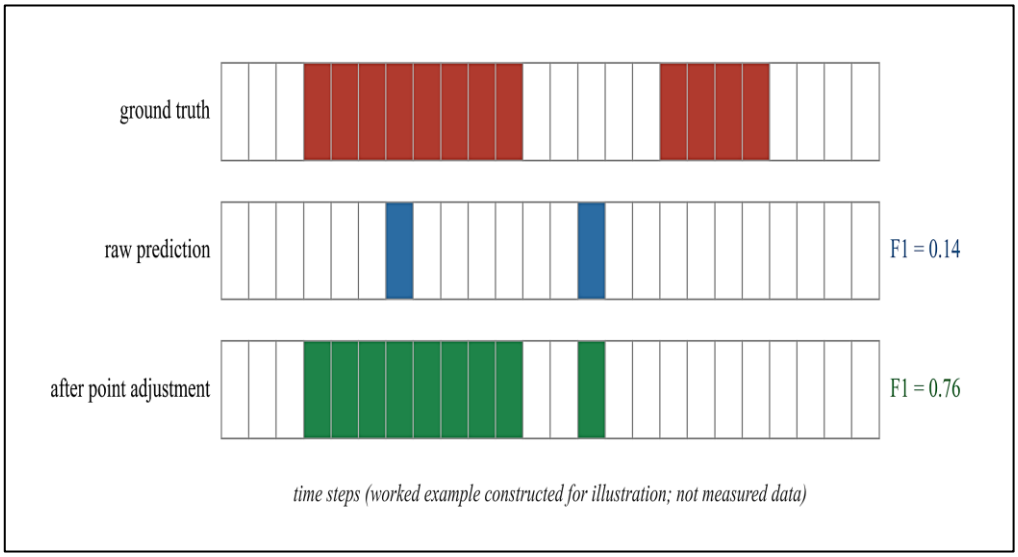

Fig. 2: Worked example of point adjustment on a hand-constructed 24-step sequence. One correct point and one false alarm give a point-wise F1 of 0.14; after adjustment the same prediction scores 0.76. The rule is the one described in [14] and analysed in [32].

Kim and colleagues worked out the consequences. Because long segments are so generously rewarded, a detector that emits random scores has a good chance of landing one point in each segment, and under point adjustment such random scores reached F1 values on par with, or above, published deep models on standard datasets. An untrained network with random weights was comparable to published models even without point adjustment. They proposed a protocol that adjusts a segment only when a minimum fraction of its points is detected, and recommended reporting unadjusted scores next to any adjusted ones [32]. Doshi and colleagues made a related proposal that credits each event once and penalises each false alarm once [34].

B. Alternatives

Point-wise precision and recall err in the other direction: they treat a detection one step late as a miss and a false alarm. Several metrics try to respect the interval nature of events. Tatbul and colleagues generalised precision and recall to ranges, with terms for how much of an event is covered and where [35]. A time-series-aware variant addresses ambiguous boundaries [36]. Affiliation metrics measure the temporal distance between predicted and true events and are built to be interpretable locally, event by event [37]. The Numenta score rewards early detection inside a window [27].

All of these still depend on a threshold. Threshold-free measures integrate over it. Paparrizos and colleagues showed that threshold-based and standard area-under-curve measures are sensitive to lag between labels and detections, and proposed range-aware versions together with a volume-under-surface measure that also integrates over a tolerance buffer around each event [38]. A further trap sits in how the threshold is chosen: reporting the best F1 over all thresholds on the test set uses the test labels for tuning, which no deployed system can do. Table 2 compares the metrics.

Table 2. Evaluation metrics for time-series anomaly detection and their main weakness.

Metric	Unit scored	Main weakness	Source
Point-wise F1	Each time step	Cannot tell early from late detection	[30]
Point-adjusted F1	Segment, if any point is hit	Strong inflation for long segments	[14], [32]
Partially adjusted F1	Segment, if a set fraction is hit	Fraction is a free parameter	[32]
Range-based precision and recall	Overlap of predicted and true ranges	Several tunable weights	[35], [36]
Affiliation metrics	Temporal distance between events	Less familiar to readers	[37]
Numenta score	Earliest detection in a window	Window and profile must be chosen	[27]
Range AUC, volume under surface	All thresholds, with tolerance buffer	Costlier to compute	[38]

V. WHAT FAIR COMPARISONS FOUND

With better data and metrics in hand, several groups reran the comparison. Schmidl, Wenig and Papenbrock evaluated 71 detection algorithms on 976 time series. No method was best across the board, and the deep learning approaches did not outperform the classical ones overall, while costing far more to train [33]. Garg and colleagues compared multivariate detectors under a unified protocol and found simple reconstruction baselines competitive with elaborate architectures [39]. Audibert and colleagues asked in their title whether deep networks contribute to multivariate detection at all, and answered that, across the datasets they tried, no family of methods consistently beat the others [40].

The most pointed version came in 2024. Sarfraz and colleagues took the state-of-the-art deep models, evaluated them without point adjustment against baselines as plain as principal component analysis and single linear layers, and reported that the baselines matched or exceeded them. They argue that the field's architectural complexity has not been shown to help [41]. The curated benchmark of Liu and Paparrizos reached a similar reading. Statistical methods and simple neural architectures were at or near the top, most plainly on univariate series, while neural networks did relatively better on multivariate ones [31].

It would be wrong to read this as proof that deep models cannot help. They may. The studies show that the published evidence for their advantage does not hold up; they do not show that no advantage exists. Complex multivariate dependencies of the kind graph models target [19] may well reward capacity on the right data. The burden, though, has shifted to whoever claims it.

VI. PRETRAINED MODELS AND NEIGHBOURING PROBLEMS

A newer line of work pretrains a single model on many series and applies it without task-specific training. Forecasting has seen the most of this, and an article in this journal reviewed those time-series foundation models and found that a frozen pretrained model often, though not always, matched per-dataset deep models [42]. For anomaly detection the evidence is thinner. Zhou and colleagues adapted a pretrained language model to several time-series tasks, anomaly detection among them [43]. Whether such results hold up under the stricter protocols of Section IV has, as far as we can find, not been settled. The benchmark history above suggests caution until it has.

Two neighbouring problems are worth separating from this one. Fraud detection is anomaly detection over transactions, but labels eventually arrive and the structure is relational; the review of graph neural networks for fraud in this journal treats it as supervised classification under extreme imbalance [44]. Drift detection, also covered here recently [45], looks for a lasting change in the data distribution, where anomaly detection looks for a brief departure from it. The two are easily confused in monitoring systems. A level shift is an anomaly for a few steps and a new normal after that. Fig 3 shows how the works cited here divide across the themes of the review.

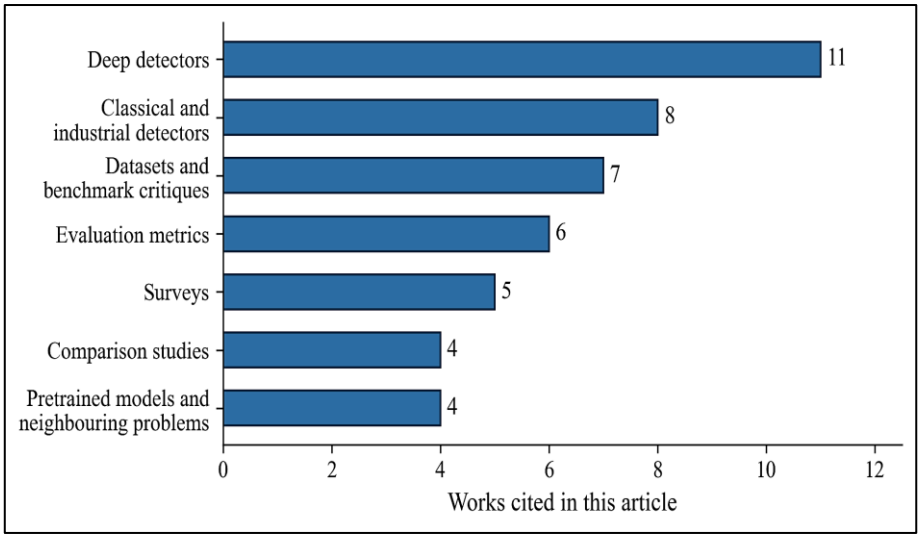

Fig. 3: The 45 works cited in this article, grouped by the role each plays in the review. The chart counts our own reference list and reports no experimental result.

VII. A REPORTING CHECKLIST

The critical studies agree on enough to state a checklist. Report unadjusted point-wise scores; if point adjustment is used for comparability with older papers, show both [32]. Add one range-aware or threshold-free metric [35], [38]. Choose thresholds without test labels and say how. Include a random-score baseline and an untrained-model baseline, since both have embarrassed published results, and include at least one classical detector and one trivial rule. Use a curated benchmark, or state which known flaws the chosen datasets carry [28].

One more item is easy to overlook. Run more than one seed and give the spread. Without it a reader cannot tell a difference between two architectures from the variation between two runs of the same one. This item is our own addition; the others restate what the cited studies recommend.

VIII. CONCLUSION

Time-series anomaly detection spent several years optimising against datasets and a scoring rule that could not tell a good detector from a lucky one. The community found the problem itself, which is to its credit, and the repairs are now available: cleaner archives, metrics that respect events without inflating them, and comparison studies broad enough to trust.

The practical upshot is modest and useful. Start with a classical detector and a simple reconstruction baseline. Move to a larger model only when it wins under an unadjusted, threshold-honest protocol on data that resemble the deployment. The same discipline should be applied to pretrained time-series models as they arrive. They may turn out to be a real advance. The earlier architectures looked like one too, for a while.

REFERENCES

[1] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” ACM Computing Surveys, vol. 41, no. 3, pp. 1–58, 2009, doi: 10.1145/1541880.1541882.
[2] A. Blazquez-Garcia, A. Conde, U. Mori, and J. A. Lozano, “A review on outlier/anomaly detection in time series data,” ACM Computing Surveys, vol. 54, no. 3, pp. 1–33, 2022, doi: 10.1145/3444690.
[3] G. Pang, C. Shen, L. Cao, and A. van den Hengel, “Deep learning for anomaly detection: A review,” ACM Computing Surveys, vol. 54, no. 2, pp. 1–38, 2022, doi: 10.1145/3439950.
[4] L. Ruff et al., “A unifying review of deep and shallow anomaly detection,” Proceedings of the IEEE, vol. 109, no. 5, pp. 756–795, 2021, doi: 10.1109/JPROC.2021.3052449.
[5] Z. Zamanzadeh Darban, G. I. Webb, S. Pan, C. Aggarwal, and M. Salehi, “Deep learning for time series anomaly detection: A survey,” ACM Computing Surveys, vol. 57, no. 1, pp. 1–42, 2025, doi: 10.1145/3691338.
[6] F. T. Liu, K. M. Ting, and Z.-H. Zhou, “Isolation forest,” in Proceedings of the IEEE International Conference on Data Mining (ICDM), 2008, pp. 413–422, doi: 10.1109/ICDM.2008.17.
[7] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, “LOF: Identifying density-based local outliers,” in Proceedings of the ACM SIGMOD International Conference on Management of Data, 2000, pp. 93–104, doi: 10.1145/342009.335388.
[8] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, “Estimating the support of a high-dimensional distribution,” Neural Computation, vol. 13, no. 7, pp. 1443–1471, 2001, doi: 10.1162/089976601750264965.

[9] C.-C. M. Yeh et al., "Matrix profile I: All pairs similarity joins for time series: A unifying view that includes motifs, discords and shapelets," in Proceedings of the IEEE International Conference on Data Mining (ICDM), 2016, pp. 1317–1322, doi: 10.1109/ICDM.2016.0179.
[10] N. Laptev, S. Amizadeh, and I. Flint, "Generic and scalable framework for automated time-series anomaly detection," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2015, pp. 1939–1947, doi: 10.1145/2783258.2788611.
[11] H. Ren et al., "Time-series anomaly detection service at Microsoft," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2019, pp. 3009–3017, doi: 10.1145/3292500.3330680.
[12] S. Ahmad, A. Lavin, S. Purdy, and Z. Agha, "Unsupervised real-time anomaly detection for streaming data," Neurocomputing, vol. 262, pp. 134–147, 2017, doi: 10.1016/j.neucom.2017.04.070.
[13] K. Hundman, V. Constantinou, C. Laporte, I. Colwell, and T. Soderstrom, "Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2018, pp. 387–395, doi: 10.1145/3219819.3219845.
[14] H. Xu et al., "Unsupervised anomaly detection via variational auto-encoder for seasonal KPIs in web applications," in Proceedings of the World Wide Web Conference (WWW), 2018, pp. 187–196, doi: 10.1145/3178876.3185996.
[15] Y. Su, Y. Zhao, C. Niu, R. Liu, W. Sun, and D. Pei, "Robust anomaly detection for multivariate time series through stochastic recurrent neural network," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2019, pp. 2828–2837, doi: 10.1145/3292500.3330672.
[16] J. Audibert, P. Michiardi, F. Guyard, S. Marti, and M. A. Zuluaga, "USAD: Unsupervised anomaly detection on multivariate time series," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2020, pp. 3395–3404, doi: 10.1145/3394486.3403392.
[17] B. Zong et al., "Deep autoencoding Gaussian mixture model for unsupervised anomaly detection," in Proceedings of the International Conference on Learning Representations (ICLR), 2018.
[18] D. Li, D. Chen, B. Jin, L. Shi, J. Goh, and S.-K. Ng, "MAD-GAN: Multivariate anomaly detection for time series data with generative adversarial networks," in Artificial Neural Networks and Machine Learning—ICANN 2019: Text and Time Series, Lecture Notes in Computer Science. Cham, Switzerland: Springer, 2019, pp. 703–716, doi: 10.1007/978-3-030-30490-4_56.
[19] A. Deng and B. Hooi, "Graph neural network-based anomaly detection in multivariate time series," in Proceedings of the AAAI Conference on Artificial Intelligence, 2021, pp. 4027–4035, doi: 10.1609/aaai.v35i5.16523.
[20] H. Zhao et al., "Multivariate time-series anomaly detection via graph attention network," in Proceedings of the IEEE International Conference on Data Mining (ICDM), 2020, pp. 841–850, doi: 10.1109/ICDM50108.2020.00093.
[21] L. Shen, Z. Li, and J. T. Kwok, "Timeseries anomaly detection using temporal hierarchical one-class network," in Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2020.
[22] J. Xu, H. Wu, J. Wang, and M. Long, "Anomaly Transformer: Time series anomaly detection with association discrepancy," in Proceedings of the International Conference on Learning Representations (ICLR), 2022.
[23] S. Tuli, G. Casale, and N. R. Jennings, "TranAD: Deep transformer networks for anomaly detection in multivariate time series data," Proceedings of the VLDB Endowment, vol. 15, no. 6, pp. 1201–1214, 2022, doi: 10.14778/3514061.3514067.
[24] A. Siffer, P.-A. Fouque, A. Termier, and C. Largouet, "Anomaly detection in streams with extreme value theory," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2017, pp. 1067–1075, doi: 10.1145/3097983.3098144.
[25] A. P. Mathur and N. O. Tippenhauer, "SWaT: A water treatment testbed for research and training on ICS security," in Proceedings of the International Workshop on Cyber-Physical Systems for Smart Water Networks (CySWater), 2016, pp. 31–36, doi: 10.1109/CySWater.2016.7469060.
[26] A. Abdulaal, Z. Liu, and T. Lancewicki, "Practical approach to asynchronous multivariate time series anomaly detection and localization," in Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2021, pp. 2485–2494, doi: 10.1145/3447548.3467174.
[27] A. Lavin and S. Ahmad, "Evaluating real-time anomaly detection algorithms—The Numenta anomaly benchmark," in Proceedings of the IEEE International Conference on Machine Learning and Applications (ICMLA), 2015, pp. 38–44, doi: 10.1109/ICMLA.2015.141.
[28] R. Wu and E. Keogh, "Current time series anomaly detection benchmarks are flawed and are creating the illusion of progress," IEEE Transactions on Knowledge and Data Engineering, 2021, doi: 10.1109/TKDE.2021.3112126.
[29] J. Paparrizos, Y. Kang, P. Boniol, R. S. Tsay, T. Palpanas, and M. J. Franklin, "TSB-UAD: An end-to-end benchmark suite for univariate time-series anomaly detection," Proceedings of the VLDB Endowment, vol. 15, no. 8, pp. 1697–1711, 2022, doi: 10.14778/3529337.3529354.
[30] D. Wagner, T. Michels, F. C. F. Schulz, A. Nair, M. Rudolph, and M. Kloft, "TimeSeAD: Benchmarking deep multivariate time-series anomaly detection," Transactions on Machine Learning Research, 2023.
[31] Q. Liu and J. Paparrizos, "The elephant in the room: Towards a reliable time-series anomaly detection benchmark," in Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2024, pp. 108231–108261, doi: 10.52202/079017-3437.
[32] S. Kim, K. Choi, H.-S. Choi, B. Lee, and S. Yoon, "Towards a rigorous evaluation of time-series anomaly detection," in Proceedings of the AAAI Conference on Artificial Intelligence, 2022, pp. 7194–7201, doi: 10.1609/aaai.v36i7.20680.
[33] S. Schmidl, P. Wenig, and T. Papenbrock, "Anomaly detection in time series: A comprehensive evaluation," Proceedings of the VLDB Endowment, vol. 15, no. 9, pp. 1779–1797, 2022, doi: 10.14778/3538598.3538602.
[34] K. Doshi, S. Abudalou, and Y. Yilmaz, "Reward once, penalize once: Rectifying time series anomaly detection," in Proceedings of the International Joint Conference on Neural Networks (IJCNN), 2022, pp. 1–8, doi: 10.1109/IJCNN55064.2022.9891913.

[35] N. Tatbul, T. J. Lee, S. Zdonik, M. Alam, and J. Gottschlich, "Precision and recall for time series," in Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2018.
[36] W.-S. Hwang, J.-H. Yun, J. Kim, and H. C. Kim, "Time-series aware precision and recall for anomaly detection: Considering variety of detection result and addressing ambiguous labeling," in Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM), 2019, pp. 2241–2244, doi: 10.1145/3357384.3358118.
[37] A. Huet, J. M. Navarro, and D. Rossi, "Local evaluation of time series anomaly detection algorithms," in Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2022, pp. 635–645, doi: 10.1145/3534678.3539339.
[38] J. Paparrizos, P. Boniol, T. Palpanas, R. S. Tsay, A. Elmore, and M. J. Franklin, "Volume under the surface: A new accuracy evaluation measure for time-series anomaly detection," Proceedings of the VLDB Endowment, vol. 15, no. 11, pp. 2774–2787, 2022, doi: 10.14778/3551793.3551830.
[39] A. Garg, W. Zhang, J. Samaran, R. Savitha, and C.-S. Foo, "An evaluation of anomaly detection and diagnosis in multivariate time series," IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 6, pp. 2508–2517, 2022, doi: 10.1109/TNNLS.2021.3105827.
[40] J. Audibert, P. Michiardi, F. Guyard, S. Marti, and M. A. Zuluaga, "Do deep neural networks contribute to multivariate time series anomaly detection?" Pattern Recognition, vol. 132, Art. no. 108945, 2022, doi: 10.1016/j.patcog.2022.108945.
[41] M. S. Sarfraz, M.-Y. Chen, L. Layer, K. Peng, and M. Koulakis, "Position: Quo vadis, unsupervised time series anomaly detection?" in Proceedings of the International Conference on Machine Learning (ICML), 2024, pp. 43461–43476.
[42] W. M. John, "Time-series foundation models for zero-shot forecasting," Eduschool International Journal of Data Science and Machine Learning, vol. 1, no. 2, pp. 15–20, Jul. 2026, doi: 10.63090/EIJDSML/3139.7344.0008.
[43] T. Zhou, P. Niu, X. Wang, L. Sun, and R. Jin, "One fits all: Power general time series analysis by pretrained LM," in Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2023, pp. 43322–43355, doi: 10.52202/075280-1877.
[44] N. Raji, "Graph neural networks for financial fraud and anomaly detection," Eduschool International Journal of Data Science and Machine Learning, vol. 1, no. 2, pp. 21–27, Jul. 2026, doi: 10.63090/EIJDSML/3139.7344.0009.
[45] V. Sakhtipriyanka, "Data-centric AI and MLOps: Drift detection and model monitoring," Eduschool International Journal of Data Science and Machine Learning, vol. 1, no. 2, pp. 8–14, Jul. 2026, doi: 10.63090/EIJDSML/3139.7344.0007.