A Relative Gap Analysis of Machine-Learning Frameworks for Heart-Disease Prediction and the Case for an Optimization-Driven Lightweight Transformer

Dr. Ashwini Shinde[1], Dr. Sashikant Gupta[2],

[1] pdf.Ashwini@lincoln.edu.my ; [2] shashigupta@lincoln.edu.my ;

Abstract: Ten years of machine-learning work on heart-disease prediction has chased a single number classification accuracy on one small cohort, usually the 303-patient Cleveland set. The best of these studies now report accuracies near 98%, and the newest add explainability and synthetic oversampling on top. We argue that this progress measures the wrong thing. Three properties decide whether a risk model ever reaches a patient in a low-resource clinic-does it hold across populations it was not tuned on, can a clinician see why it decided, and will it run on the phone in a health worker's pocket-and almost none of the accuracy first literature measures any of them. This paper reads the field as a search that optimized one axis and left three unpriced. We survey the accuracy-first tradition, the parallel track of metaheuristic feature selection, and the small but growing green-AI literature, and we name five concrete gaps that follow. We then test a recent design, SFO-LiteTransMLP, against all five. Because it selects features with Synergistic Fibroblast Optimization, feeds them to a two-layer transformer of 0.42 million parameters, carries SHAP inside the pipeline rather than beside it, and is validated across four cohorts with leave-one-dataset-out transfer, it is the first model in this line that can be judged on every axis at once-and it clears the bar on all of them at roughly one-fiftieth the parameter cost of its nearest deep competitor

Keywords: heart-disease prediction; research gap analysis; metaheuristic feature selection; lightweight transformer; explainable AI; sustainable machine learning, cross-cohort generation.

Introduction

A model that scores 97% on the Cleveland heart-disease cohort and 78% on the large Kaggle cardiovascular set is not two models. It is one model meeting two populations, and the nineteen-point gap between those numbers is the part of the story a patient would care about. Most published work on machine-learning heart-disease prediction reports the first number and never records the second, because the second requires testing on a cohort the model was not tuned to fit.

Cardiovascular disease still kills more people than any other single cause, roughly 17.9 million a year, and the burden falls hardest on the low- and middle-income health systems least able to absorb it [1]. That fact has driven a decade of prediction research. The trajectory is easy to read: individual classifiers gave way to ensembles, ensembles to hybrid neural networks, and hybrids to attention-based deep models, with the headline accuracy on Cleveland climbing from the mid-eighties into the high nineties along the way [2]–[7]. On its own terms the field has succeeded. Our claim is that its terms were too narrow.

The accuracy-first literature optimizes the first requirement and treats the other three as optional extras-addressed, when at all, by a paragraph of future work.

Related work

A. The accuracy-first tradition on small cohorts

This study used the Cleveland Heart Disease dataset, which included 303 people and thirteen clinical features. While Latha and Jeeva's collaborative work is a clear early example-bagging and boosting over weak base learners boosted accuracy on Cleveland to the mid-eighties [3], Mohan et al.'s mix of feature selection and a linear-plus-tree classifier advanced the same dataset over 88% [4]. Both papers describe a single split of a single cohort. While Arabasadi et al. used a genetic technique to seed a neural network's weights and reached almost 94% [5], Bharti et al. combined conventional and deep classifiers with outlier removal for a similar result [6]. Each result is real; none of them tells us how the model behaves on a patient recorded under a different protocol.

El-Sofany et al. represent the mature form of this tradition [2]. They select features three ways-chi-square, ANOVA, and mutual information-balance the classes with SMOTE, compare ten classifiers, tune hyperparameters, and explain the winner with SHAP. XGBoost on their SF-2 subset reached 97.57% accuracy with a 0.98 AUC. This is about as far as the accuracy-first recipe goes, and it is worth being precise about what it does and does not establish. It establishes that a boosted-tree model can separate disease from non-disease on a combined Cleveland-and-private set of 503 records. It does not establish that the same model transfers to Framingham or to a 70,000-patient competition cohort, because those were never in the evaluation. The SHAP layer, likewise, explains the final model after the fact; it plays no part in how the model was built.

B. Attention arrives, but the tree problem does not go away.

Transformers reshaped language and vision before anyone asked them to rank clinical tabular rows. Rahman et al. built a self-attention model specifically for the small Cleveland cohort and reported 96.51% [7], which is, to our reading, the most convincing of the recent single-cohort deep results-not because the number is highest but because the architecture is matched to the data rather than borrowed wholesale from NLP. Even so, it inherits the tradition's blind spot: one cohort, no deployment accounting.

There is a deeper tension underneath. Grinsztajn et al. benchmarked tree ensembles against modern deep networks across forty-five tabular datasets and found the trees still state-of-the-art on medium-sized data, and faster [8]. Their explanation is that neural networks are biased toward smooth functions, are hurt by uninformative features, and do not preserve the orientation of the data-three properties that clinical tables violate constantly. The tabular-transformer designs that dominate the deep leaderboard were built partly in answer to this: TabNet learns sequential attention masks over features [9], Tab Transformer embeds categorical and processes them with stacked self-attention [10], and FT-Transformer generalizes the idea to mixed-type columns [11]. They are elegant and they are heavy-tens of millions of parameters; GPU memory measured in gigabytes. Whether that weight buys anything on a thirteen-feature cardiac table is a question the papers introducing them did not have to answer, because they were never aiming at a phone

C. A parallel track: metaheuristic feature selection

Alongside the deep-learning work runs a separate line that treats feature selection as combinatorial optimization and hands it to a bio-inspired search. Particle swarm, the grey wolf optimizer [12], whale optimization, and their many hybrids have all been wrapped around Cleveland and Statlog. Narasimhan and Victor give a representative recent instance: genetic, swarm, and ant-colony variants each feeding a random forest, with accuracies in the low nineties depending on the pairing [13]. Two things recur across this track. The optimizer almost always feeds a classical classifier-random forest, support vector machine, XGBoost-so whether metaheuristic selection helps a deep model is left open. And the optimizers themselves have been worked over so thoroughly that yet another swarm variant is hard to justify as novel [14].

D. The sustainability blind spot.

The green-AI literature has given us a vocabulary for model cost-parameters, FLOPs per inference, energy per query, the carbon of training [18], [19]. Almost all of it targets large language and vision models.

Key Contribution

Gap Identification.

Gaps in the Accuracy

Gap 1: generalization is asserted, not measured.

Reporting accuracy on one cohort and assuming it would hold elsewhere is the most prevalent practice in this research. When a model that was fine-tuned on 303 Cleveland patients is transferred to Framingham's 10-year prospective design or a 70,000-row competition set, much of the benefit disappears. The model also fits the age distribution, class ratio, and recording procedures of that cohort. Since their evaluation never left the training group, the majority of papers—including the strongest ones—simply lack the evidence for transfer. A model's stated accuracy is a statement about a dataset, not about heart disease, unless it is trained on certain cohorts and evaluated on a held-out one.

Gap 2: The optimizer and the deep model never meet

For years, the deep-learning track and the metaheuristic track have operated independently of one another. Random forests and support vector machines are fed by feature-selection searches; deep tabular models accept all features and use attention to sort them. The obvious hybrid—let a metaheuristic trim the feature set that a transformer then learns from—is uncommon.

Gap 3: Explanation is decoration, not structure.

Most contemporary articles finish with SHAP as a bar chart. It cannot affect the design because it is computed after the model is fixed, and no one can determine whether the explanation is stable because it is displayed for a single cohort. A single attribution plot is insufficient for a doctor.

Gap 4: the deployment cost nobody prices

This is the gap that the whole framing turns on. A tabular transformer with thirty-eight million parameters and a hundred-and-fifty-megabyte footprint is a fine object on a cloud GPU and a useless one on an entry-level Android phone with intermittent power. Yet parameter count, inference FLOPs, quantized model size, on-device latency, and energy per prediction are absent from essentially the entire cardiac-prediction literature.

Gap 5: soft validation lets weak results look strong.

Headline accuracy on a single split, with no significance testing and sometimes with feature selection performed on the test fold, recurs often enough to be a structural problem. Because class imbalance inflates accuracy while sensitivity lags, and because MCC is rarely reported [17], a model can look strong on the metric that is printed and weak on the metric that matters clinically. Without held-out test folds, paired significance tests, and imbalance-aware metrics, the ranking of methods in this field is less settled than the numbers suggest.

Method, Experiments and Results

Reading the Gaps Against SFO-LiteTransMLP

We now test one recent design against these five gaps. SFO-LiteTransMLP pairs Synergistic Fibroblast Optimization for binary feature selection with a two-layer transformer-MLP back end that adds channel-attention recalibration and a residual feature path, and it carries a SHAP layer through the pipeline. It is evaluated on four cohorts-UCI Cleveland, Statlog, Framingham, and the Kaggle cardiovascular set, more than 74,000 records in total. The question is not whether it wins on accuracy. It is whether it answers the gaps, and whether it does so without quietly creating new ones.

Generalization (Gap 1). The design is evaluated with a leave-one-dataset-out protocol: train on three cohorts, test on the fourth, with each cohort reweighted so size does not dominate the gradient. The mean transfer accuracy is 86.21%, only about five points below the in-cohort figures, and the drop is largest exactly where one would predict-Framingham, whose ten-year horizon is unlike the other three. This is the measurement the accuracy-first papers omit, and reporting it is more informative than any single in-cohort number, because it estimates behaviour on a population the model never saw.

Optimizer meets deep model (Gap 2). SFO here selects the feature mask that the transformer then learns from, which is, as far as we can tell, the first time this particular metaheuristic has driven a deep cardiac model rather than a classical classifier. The ablation is the useful evidence: removing SFO and feeding all eighteen harmonized features costs 1.34 accuracy points on Cleveland, the second-largest single drop after removing the transformer encoder itself. The two tracks that ran in parallel for years are, in this design, doing complementary work-the search prunes, the attention learns interactions on what survives.

Explanation as structure (Gap 3). SHAP is computed inside the pipeline for every prediction, and the global rankings are compared across all four cohorts rather than shown once. The recovered ordering-ST-depression, number of major vessels, chest-pain type, maximal heart rate, age-matches the risk factors clinicians already weight, and the stable features stay stable while cohort-specific ones move in the expected direction. That cross-cohort consistency is the property Gap 3 asks for, and it is absent from the post-hoc SHAP plots of the accuracy-first papers, including the base paper [2].

The accuracy-first program could not make these measurements, not because its authors lacked the tools, but because a thirty-eight-million-parameter model on one cohort leaves nothing to measure them with.

Conclusions

The heart-disease prediction literature has spent a decade improving a number that was never sufficient. Single-cohort accuracy climbed from the mid-eighties to nearly 98%, and along the way the three properties that decide whether a model reaches a patient-cross-population reliability, native interpretability, and on-device feasibility-went unmeasured. This is not a failure of any one

paper. The base paper we anchored is careful and well-executed within its frame; the frame is the problem.

Read against the five gaps that frame produces, SFO-LiteTransMLP is the first design in this line that can be judged on every axis at once and clears the bar on all of them-competitive accuracy, measured cross-cohort transfer, cohort-stable SHAP explanations, corrected significance testing, and a deployment footprint two orders of magnitude below its nearest deep competitor.

No benchmark can provide the next step. The sustainability estimates imply hardware that may change, and all of the results below are based on publicly available datasets compiled for research rather than prospective clinical data. A deployment study in an operational primary-care context with a fairness audit across the demographic groups these public cohorts underrepresent would resolve the issue. Until that exists, the case made here is that the field has been measuring the wrong thing, and that a model built small enough to deploy is what finally lets it measure the right ones.

References

1. World Health Organization, "Cardiovascular diseases (CVDs)," WHO Fact Sheets, Geneva, Switzerland, 2023.
2. H. El-Sofany, B. Bouallegue, and Y. M. Abd El-Latif, "A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method," Scientific Reports, vol. 14, art. 23277, 2024, doi: 10.1038/s41598-024-74656-2
3. C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," Informatics in Medicine Unlocked, vol. 16, art. 100203, 2019, doi: 10.1016/j.imu.2019.100203.
4. S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," IEEE Access, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.
5. Z. Arabasadi, R. Alizadehsani, M. Roshanzamir, H. Moosaei, and A. A. Yarifard, "Computer aided decision making for heart disease detection using hybrid neural network–genetic algorithm," Computer Methods and Programs in Biomedicine, vol. 141, pp. 19–26, 2017, doi: 10.1016/j.cmpb.2017.01.004.
6. R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, "Prediction of heart disease using a combination of machine learning and deep learning," Computational Intelligence and Neuroscience, vol. 2021, art. 8387680, 2021, doi: 10.1155/2021/8387680.
7. A. U. Rahman, Y. Alsenani, A. Zafar, K. Ullah, K. Rabie, and T. Shongwe, "Enhancing heart disease prediction using a self-attention-based transformer model," Scientific Reports, vol. 14, art. 514, 2024, doi: 10.1038/s41598-024-51184-7.
8. L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on tabular data?," in Advances in Neural Information Processing Systems (NeurIPS), vol. 35, 2022, pp. 507–520.
9. S. Ö. Arik and T. Pfister, "TabNet: Attentive interpretable tabular learning," in Proc. AAAI Conf. Artificial Intelligence, vol. 35, no. 8, 2021, pp. 6679–6687, doi: 10.1609/aaai.v35i8.16826.

10. X. Huang, A. Khetan, M. Cvitkovic, and Z. Karnin, "TabTransformer: Tabular data modeling using contextual embeddings," arXiv:2012.06678, 2020.
11. Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," in Advances in Neural Information Processing Systems (NeurIPS), vol. 34, 2021, pp. 18932–18943.
12. S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey wolf optimizer," Advances in Engineering Software, vol. 69, pp. 46–61, 2014, doi: 10.1016/j.advengsoft.2013.12.007.
13. . Narasimhan and A. Victor, "A hybrid approach with metaheuristic optimization and random forest in improving heart disease prediction," Scientific Reports, vol. 15, art. 10971, 2025, doi: 10.1038/s41598-024-73867-x.
14. T. T. Dhivyaprabha, P. Subashini, and M. Krishnaveni, "Synergistic fibroblast optimization: A novel nature-inspired computing algorithm," Frontiers of Information Technology & Electronic Engineering, vol. 19, no. 7, pp. 815–833, 2018, doi: 10.1631/FITEE.1601553.
15. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017, pp. 4765–4774.
16. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.
17. D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," BMC Genomics, vol. 21, art. 6, 2020, doi: 10.1186/s12864-019-6413-7.
18. R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green AI," Communications of the ACM, vol. 63, no. 12, pp. 54–63, 2020, doi: 10.1145/3381831.
19. E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in NLP," in Proc. 57th Annual Meeting of the Association for Computational Linguistics (ACL), 2019, pp. 3645–3650, doi: 10.18653/v1/P19-1355.