Intelligent Hybrid Feature Selection for Healthcare Disease Prediction: A Critical Review of Methods, Applications, and Research Challenges

Ankit Mallick[1*], Stuti Chandrakar[2], Bipradip Gayen[3], Bapi Bhuinya[4]
[1]*Independent Researcher, Department of Information technology, Maulana Abul Kalam Azad University of Technology, Durgapur, West Bengal, India*
[2]*Independent Researcher, Department of Civil Engineering, Bhilai Institute of Technology, Chhattisgarh, India*
[3]*Department of Information technology, Maulana Abul Kalam Azad University of Technology , Durgapur, West Bengal, India*
[4]*Independent Researcher, Maulana Abul Kalam Azad University of Technology, Durgapur, West Bengal, India*

***Abstract*—The use of feature selection can help with the reduction of irrelevant measurements, provide stability to model estimates and improve the performance of disease prediction systems. However, "hybrid feature selection" includes many types of approaches including: filter-wrapper cascades, regularized embedded selection, ensemble voting and metaheuristic subset search. This critical review examines the methodology behind these approaches and their application in diabetes, cardiovascular disease, cancer and chronic kidney disease. A focused search of publisher and bibliographic databases was conducted using publications from 2021-2026 and foundational methodological guidance. The literature indicates a key trade-off: adaptive wrappers can identify interactions that would have been missed by univariate screening, but they significantly increase the computational burden of model fitting and can result in greater selection bias if the validation strategy is weak. In high dimensional omic environments, subset stability is as important as achieving a high level of discrimination; whereas, for routine clinical data, acquisition costs, missingness, calibration and external transportability may be of more importance than small differences in benchmark accuracy. We compare groups of selectors, provide summaries of disease specific studies (without attempting to combine non-comparable metrics), and suggest a transparent evaluation framework that nests feature selection within training folds and evaluates independent patients or sites when possible. A three-part methodological agenda is presented: assess predictive utility and subset stability simultaneously; document computational and clinical acquisition costs; and differentiate between reproducible benchmark improvements and evidence of patient-level clinical utility. There are no new patient-level experiments reported.**

***Index Terms*—Hybrid feature selection; disease prediction; metaheuristic optimization; clinical machine learning; stability; external validation.**

I. INTRODUCTION

Clinical data are often heterogeneous: an EHR will include multiple measurements from various points in time; while a cancer gene expression array may include significantly more features than samples. Missingness, collinearity among predictors and differing sample collection methodologies during the development of models and their intended uses are common issues in both types of clinical data. Therefore, the process of selecting a subset of potential predictor variables is much more than simply a software optimization task. It impacts what a model is measuring, how well a model can be understood, and whether the selected predictors can be measured in a typical clinical setting [1-3].

The use of a feature selection method is advantageous if it reduces the number of measurements that are made during data collection, decreases the time required to train a machine learning model, or reduces

the amount of variance in the predictions of a model. However, each potential advantage must be supported by evidence. It may be easier to collect a small number of biomarkers than to perform a complex search for many biomarkers. Also, a very complex search can make it difficult to reproduce the results. This article is intended to provide distinction among these two competing objectives as well as to assess the quality of evidence provided in previous studies of disease prediction [4]– [6].

1.1. Review questions and contribution

The author is asking: Which methods do they use to combine these mechanisms into an intelligent hybrid selector? What data exists that supports the idea that this approach can provide benefits across multiple diseases? In what way do researchers evaluate each of the five criteria (predictive ability, selection size, stability, efficiency, and transportability) as a whole? Rather than creating a new algorithm or performing a meta-analysis, the authors have developed a shared conceptual framework for evaluating various approaches to selecting subsets of variables. They also offer a validation-based method for evaluating these different approaches.

Performance should be evaluated in terms of what is being tested. Screening sensitivity, diagnostic discrimination, individual risk estimation, and prognosis cannot be used interchangeably. An acceptable accuracy rating on a well-balanced public benchmark is insufficient to demonstrate clinical effectiveness, especially when the prevalence, threshold for intervention, and potential downstream harm vary from those of the test environment [20]-[24].

II. REVIEW METHODOLOGY

This paper used a focused critical narrative review approach based on journal articles that had been published before October 17th, 2026. A reference list was created by conducting search-topic-based searches as well as searching publisher and bibliographic records for combinations of "hybrid feature selection", "filter wrapper", "embedded selection", "meta-heuristic feature selection", "disease prediction", "clinical classification", "stability", and "validation". Articles were included in the reference list if they provided relevant information about feature selection methods, health-care applications, or reporting and validating principles. The data was organized into categories (families) of methods, and clinical outcomes; however, each individual article's results were only compared with other articles' results if there was supporting evidence available for both articles in the form of a source document. This organization method provides a structured thematic synthesis of the data, but it is not a systematic search nor is it a reproducible census of all of the possible articles that could have been selected. Additionally, the complete database-specific query logs and exclusion counts were not retained and therefore can not be claimed.

Eligibility was based on whether the article was a journal publication that addressed either predictive feature selection (PFS), disease-specific evaluation (DSE), or validation/reporting principles (VRP); non-clinical optimization work was only considered if it would be helpful in organizing the method taxonomy. Conference abstracts, unverified claims, duplicate publications, and other image enhancement studies were not used as primary evidence of disease selection. Information was collected on the following aspects: model task, dataset/population (if reported), components of selection, classifier family, description of validation, reported metric type, and limitations. Since this is a narrative review, risk-of-bias judgments are qualitative assessments and the tables provided should not be interpreted as formal systematic reviews.

2.1. Limits of source verification

The bibliography includes 40 journal-paper entries related to the theme. A small number of these records were cross-referenced with their respective publishers' websites or bibliographic databases; the remaining records will be subjected to an individual review of each author's work as well as each article prior to being submitted. A valid DOI alone is insufficient to confirm a reported level of performance. Data-extraction fields included: disease/outcome, dataset, selection method, learning algorithm, validation strategy, key results, and limitations. When a reference did not provide sufficient data, the table clearly indicates that the outcome could not be independently verified.

2.2. Narrative synthesis

Thematic analysis organized the literature into basic family types, hybrid architectures, disease contexts, and validation difficulties. Specifically, the synthesis verifies if selection occurs during resampling folds, if

stability and resource usage have been measured, and if validation is based on new patients versus a holdout from the initial development pool [5], [21], [22].

III. FOUNDATIONS OF FEATURE SELECTION

The goal of feature selection is to find a subset (S) of features from a given input matrix X with d possible features and label y that will optimize the predicted outcome, based on the predicted performance and perhaps cost. Feature selection differs from feature extraction in that it typically retains identifiable original variables. This difference is important when working within a clinical setting. While some blood test results may be ordered individually and used to estimate the value of other blood tests, most latent components produced by feature extraction cannot be ordered in the same manner.

3.1. Filter methods

The filter approach ranks each variable individually without taking into consideration how it would perform when used in conjunction with a predictive estimator. It uses methods such as correlation, mutual information, chi square association, information gain, and redundancy sensitive criteria to assess the relationship between the variable and the target. This method has several advantages over other approaches; one of which being that it is less computationally intensive than the wrapper approach. In addition, the filter approach may be more suitable for initial screening of variables prior to using a wrapper approach. However, there are also some disadvantages to this approach. For example, a variable that appears to have little or no association with the target when evaluated on its own, may actually be able to work together with other variables to improve the performance of a predictive estimator. Additionally, a variable that shows high levels of association with the target may only do so because its measurement is affected by either the target itself or by the treatment received after diagnosis [1], [3].

3.2. Wrapper methods

Candidate subsets can be evaluated through model fitting using wrapper methods such as sequential forward selection (SFS), sequential backward elimination (SBE), or recursive feature elimination (RFE). Wrapper methods evaluate each candidate subset by applying a specific classification method. Therefore, wrapper methods can be aligned with the final end-point. However, wrapper methods have two major disadvantages: first, they require multiple evaluations of the model with different subsets, and second, there may be an over-optimistic selection of the subset if the same hold-out data set was used both for selecting the subset and reporting the performance of the selected subset [4], [5].

3.3. Embedded methods

Embedded approaches include selection during model development. Regularization (such as L1) can reduce coefficient values to zero; Elastic Net approaches to dealing with correlated predictor variables allow for greater flexibility in terms of predictor variable selection; Tree Ensemble methods provide a way to identify conditional feature usage. Embedded Importance does not necessarily indicate causality nor does it guarantee stable identification of biomarkers; highly correlated variables can share the importance of each other and rankings may change when data is resampled [5, 6, 15].

3.4. Method comparison

Table 1 provides some examples of the types of characteristics that are usually found. The baseline does not necessarily have to be "no selection"; many times, a simple regularized linear estimator can provide a better and less expensive comparison against a more complex hybrid search.

Table 1. Conventional feature-selection approaches

Family	Typical techniques	Strength	Principal limitation
Filter	Mutual information; chi-square; mRMR	Fast preliminary reduction	Interactions can be missed
Wrapper	RFE; sequential selection	Classifier-specific objective	Nested fitting is expensive
Embedded	LASSO; elastic net; trees	Fits jointly with model	May be unstable with collinearity
Ensemble	Aggregated rankings	Can improve consensus	Aggregation rule must be justified

IV. INTELLIGENT HYBRID FEATURE SELECTION

The definition of a hybrid framework is the combination of two or more mechanisms; however, it must also provide the information that will cross over each stage. The filter-wrapper cascade eliminates the obviously irrelevant features prior to executing the adaptive subset search. The filter-embedded procedure screens the predictor variables prior to the regularized model fitting. The wrapper-embedded combination uses a previously trained estimator to generate potential subsets, eliminate those that are deemed inappropriate, and/or rank the potential subsets. The reliability of these types of systems depends upon if the composite process was trained using only training data [1]-[5].

4.1. Metaheuristic optimization

The Genetic Algorithms (GA) encode the candidates as a set of possible solutions (candidate subsets), and then the selection, crossover and mutation operators are used to explore alternative solutions. Particle Swarm Optimization (PSO) utilizes individual and collective search signals to update the candidate solutions; whereas, the Ant Colony Optimization (ACO) family and Grey Wolf Optimization (GWO) family utilize other forms of search dynamic. These types of methods can be useful when the relationships between the features make simple ranking insufficient; however, they do not provide any guarantees of finding the globally optimal solution. To make comparisons meaningful among the various types of optimization methods, the definition of fitness function, number of computational resources available (tuning budget), initial seed value for random number generation (stochastic seeds), and stopping criteria are critical [1]– [4], [8].

4.2. Ensemble and multi-objective designs

An ensemble can use ranking methods on either resampled data or across different selector families, to identify potential candidate biomarkers. Rather than using one overall accuracy measure to evaluate multi-objective procedures, they allow for a trade-off. This is significant because a four-variable panel with slightly lower discriminatory power might be easier to replicate and obtain than a larger, less stable group of variables [5], [6].

A general scalarized objective function can be represented as $J(S) = \alpha L(S) + \beta|S|/d + \gamma C(S)$. In this representation, L represents the validation loss of a model, $|S|$ represents the number of selected variables in a subset S, d represents the total number of available candidate variables, C represents an explicit cost or instability penalty, and α, β, and γ are non-negative weights that are specified prior to evaluating the performance of the model. It should be noted that the above representation is merely a conceptual framework and is not an implementation of any of the algorithms reviewed in this article.

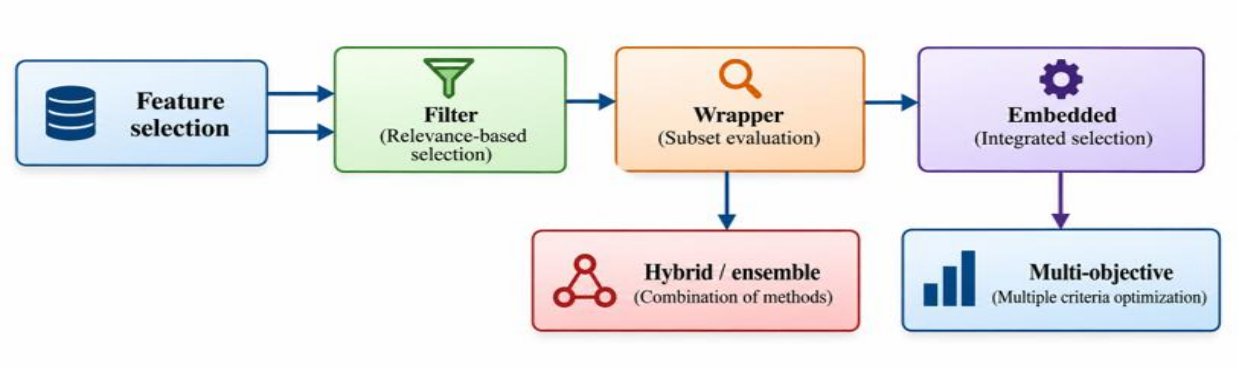

Figure 1. Taxonomy of feature-selection families and their hybrid extensions (conceptual synthesis).

4.3. Leakage-free evaluation architecture

Data preparation and data cleaning do not take place in an objective vacuum. When you apply imputation, normalization, supervised feature selection, class balance adjustment, hyperparameter tuning, and decision-making on threshold settings (which are all forms of preprocessing) based on data that has been used in a validation step of a cross-validation process, you may introduce bias into your model. Nested cross validation is one way to manage these issues. It

involves two steps: an outer split that assesses the performance of your entire modeling process, and an inner split that directs how you will select features and tune the parameters of your model. You cannot base the selection of optimal weights or number of observations to include in the evaluation of the outer test result [21], [22].

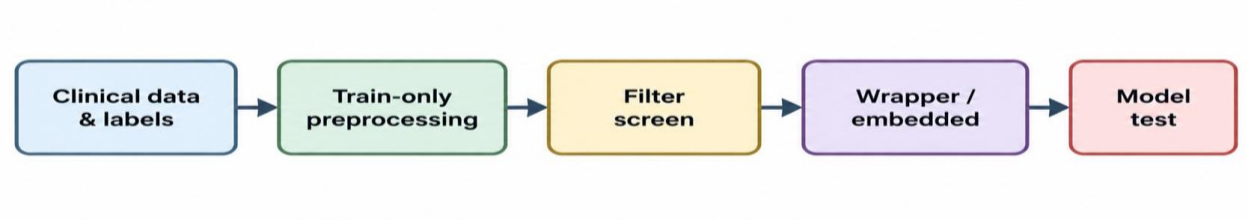

Figure 2. Training-only selection and held-out evaluation workflow (conceptual; actual studies may require nested resampling).

4.4. Measuring stability and cost

If you're resampling over and over again, you can use Jaccard Overlap and Adjusted Stability Indices (Adjusted for Chance) to describe the stability of subsets. You'll need to report on: the average number of features selected; the distribution of how many times each feature was included; the amount of computer time used; and the "inference" (or acquisition) costs. It's important to remember that high stability might be a false positive if your selector is always picking out the same trivial one-feature option. Therefore, you must look at both the stability of your selections as well as the ability of those selections to predict things accurately and clinically relevant [5], [6].

4.5. Why interpretation is not the same as selection

Post hoc explanation methods do not necessarily recover the features which were used to choose the model prior to training. Post-hoc attribution can be affected by the presence of highly correlated variables, and the properties of the background data. The use of features in making predictions (feature selection) should be assessed by repeatedly fitting the model, and then testing on independent validation sets; assessing how useful an explanation is for clinicians will probably need to be done separately [15], [19].

V. APPLICATIONS IN HEALTHCARE DISEASE PREDICTION

5.1. Diabetes

In addition to providing some background on the use of compact clinical tables in diabetes studies (which typically contain glycemic measurements, anthropometric indicators, demographic information and other relevant health factors), we also provide an example of such a compact table in the form of a hybrid selection-and-classification study conducted by Talari et al., as well as a metaheuristic feature selection review of diabetes prediction studies published in 2025 by Sirmayanti et al. [7], [11]. While both studies illustrate the practical value of selecting smaller sets of variables, the performance of classifiers used in these types of studies will vary depending upon the number of individuals involved in the study, how missing/physiologically implausible entries are dealt with and whether or not resampling is performed prior to or after data splitting.

The first major methodological concern is that the performance of a model may appear to be overly high due to the repeated use of a single public benchmark during the development phase of a method. In addition to creating artificially inflated results, using the same dataset repeatedly throughout multiple published studies reduces the ability to claim independent validation. A better way to develop models would be to define a "held out" cohort based upon either time (e.g., all patients admitted after a certain date) or by hospital (e.g., all patients from one hospital), then evaluate both calibration and discrimination on those patients.

5.2. Cardiovascular disease

The cardiovascular literature contains examples of metaheuristic approaches used in conjunction with random forest models, either as part of a single model (ensemble) or as separate components (embedded), and in some cases, both approaches are combined into a multi-step process for selecting features [8,9,12-17]. This body of research illustrates many different design

options, rather than a common standard for competing against each other: the study populations, the outcomes measured, how data was processed, and how the results were validated may vary. Therefore, the results obtained from analyzing a publicly available tabular dataset should not be viewed as providing a reliable estimate of the potential performance of such models when applied to screen large numbers of people.

The possible relevant input measurements can include clinical exam, metabolic marker, ECG indicator, and symptom variable. The selector needs to consider if the input measurement is available at the time of the prediction, whether the input measurement has been affected by treatment decisions, and if there are costs associated with obtaining the input measurement.

5.3. Cancer and molecular classification

The number of genes measured in a gene expression classification study may be many times larger than the number of subjects participating in that same study. There have been several reviews published about different approaches to selecting features in microarray studies, as well as about using ensemble techniques to find markers that will be consistently identified across multiple studies [6, 28, 29]. When selecting features from the complete set of data prior to validation, it is extremely dangerous since feature selection could occur based on chance associations within the small cohort. Both the biological relevance of the selected genes and whether the techniques used to identify those genes are reproducible across laboratories are separate issues from how accurate the genes are at classifying subjects.

The use of imaging for predicting cancer causes more unit of analysis issues. There can be high dependency between multiple scans, slices, or patches from a single patient. To say that your model generalizes independently, you will need to split patients, have other people review your work, and validate that your model works well on the same type of sites as the ones you used to develop it. The medical imaging reporting guidelines emphasize being clear about what units you sampled, what reference standards you used to evaluate performance, and how you designed your study to evaluate performance [20], [23], [24].

5.4. Chronic kidney disease

Yogesh et al. developed a signature-based hybrid selection using a recurrent deep learning classifier for chronic kidney disease (CKD) [10]. A separate article describes a hybrid feature selection method in CKD classification [18]. These articles support the use of common laboratory tests to identify CKD, however, neither article provides evidence that supports their use prospectively. Missing data may result from the way clinical workflows operate and differences in patient demographics in community versus specialty care settings can affect the distribution of predictors.

5.5. Cross-disease synthesis

Different types of clinical tables, molecular measurements, and imaging representation require different design choices for selecting features. Cancer omics data with high dimensions focuses more on being stable and controlling for risks associated with multiple testing. Cardiovascular disease tables or diabetes data sets that have fewer dimensions can utilize simple, pre-existing models (baselines) as an alternative. Missing data and reproducing results are more important in kidney disease studies. Regardless of which domain is used, when feature selection methods are evaluated, the improvements should be based upon evaluations designed so that they are directly comparable.

VI. COMPARATIVE ANALYSIS OF PUBLISHED STUDIES

Table 2 records representative empirical and evidence-synthesis articles. The role column separates original comparative research from reviews, avoiding a misleading appearance that all papers report directly comparable new datasets.

Table 2. Published studies relevant to hybrid healthcare feature selection (part I)

Source	Clinical dataset / endpoint	Selection and classifier	Result / evidence limitation
[7] Talari (2024)	Diabetes	Hybrid selection + classifier	Disease-specific empirical study; compare protocols carefully
[8] Narasimhan (2025)	Heart disease	Metaheuristics + random forest	Reported optimized predictor; external generalization not established here

[9] Praveen (2024)	Cardiovascular	Hybrid selection + boosted fusion	Multiple pipeline components complicate attribution
[10] Yogesh (2024)	Kidney disease	Signature selection + LSTM/GRU	Clinical transportability needs separate confirmation
[11] Sirmayanti (2025)	Diabetes	Review of metaheuristic selectors	Secondary evidence, not independent benchmark
[12] Gupta (2025)	Heart disease	Hybrid selection + ensemble	Dataset-specific model evaluation
[13] Shilpa (2025)	Cardiovascular	Hybrid filter design	Outcome/sampling differences limit pooling
[14] Lonang (2026)	Heart disease	Chi-square + backward elimination	Published single-task result; leakage controls need full-text audit
[16] Tan (2025)	Heart disease	Optimized selection + DNN	Higher capacity needs controlled baseline
[17] Dissanayake (2024)	Heart disease	Meta-learning selection	Training and tuning details require inspection

The table deliberately does not reproduce isolated accuracy percentages without verifying identical endpoint definitions, sample separation, and train–test procedures. Such percentages can create a false ranking across heterogeneous benchmark tasks.

Table 2. Published studies relevant to hybrid healthcare feature selection (part II)

Source	Clinical dataset / endpoint	Selection and classifier	Result / evidence limitation
[1] Kaur (2023)	Disease diagnosis	Systematic metaheuristic review	Heterogeneous primary evaluations
[2] Houssein (2024)	Healthcare	Optimization and AI survey	Broad clinical tasks; limited comparability
[3] Barrera-García (2024)	Cross-domain	Metaheuristic objective taxonomy	Healthcare is a subset of applications
[4] Nssibi (2023)	Cross-domain	Nature-inspired FS survey	Algorithm count is not validation quality
[5] Barbieri (2024)	Bioinformatics / ML	Selection performance and stability	Stability needs choice of resampling unit
[6] Łukaszuk (2024)	Multi-omics	Empirical selection stability	High-dimensional but not routine clinical data
[15] Takefuji (2025)	Clinical prediction	Critical note on SHAP and selection	Explanations do not establish subset necessity
[18] Yogesh (2025)	CKD	Hybrid FS + ensemble DL	Potential overlap with related report
[19] Agrawal (2025)	Healthcare XAI	Evaluation of explanation methods	Not a feature-selection superiority experiment
[25] Kumar (2025)	Medical imaging	Bio-inspired hybrid methods review	Imaging benchmarks differ from tables

6.1. Reading claims of improvement

The way that performance can be improved must be separated. If you are changing the preprocessing, balancing, selection, classifier, and tuning all at the same time, then you will not know if your improvement is due to feature selection alone. A controlled comparison of performance changes the same split and classifier, but removes or replaces one component. Comparing the predictions for the same group of independent patients allows us to estimate the difference in performance more precisely.

6.2. A minimum reporting set

For each study there must be reporting on outcome definitions, cohort provenance, sample units, missingness, how the data was split (training vs. validation), the complete nested pipeline, the number of selected features, the frequency of subsets used during repeat runs of the model, the computational costs incurred, discrimination and calibration results and, when possible, an independent external test of the model [20-24].

Table 3. Qualitative methodological comparison

Strategy	Computational demand	Stability considerations	Best-use scenario
Filter–wrapper	Medium to high	Check filter and wrapper within folds	Many initial features; expensive model
Filter–embedded	Low to medium	Regularization may vary across folds	Clinical tabular variables
Metaheuristic wrapper	High	Report seeds and subset inclusion frequency	Strong nonlinear interactions suspected
Ensemble selection	Medium to high	Aggregation and subsampling matter	Small-sample biomarker exploration
Multi-objective selection	Variable, often high	Show Pareto sensitivity and cost	Measurement cost is important

6.3. Objectives beyond classification score

Literature often reports higher discrimination in studies reporting on improved discrimination. However, the operational value of a predictive model will depend upon the endpoint selected for that model. For clinical risk models, both calibration and decision thresholds are important. On the other hand, for screening models, the number of false negatives (missed cases) and the number of unnecessary follow-ups must be considered. While a feature selector that deletes essential variables could potentially produce an acceptable average accuracy when applied to an "easy" test distribution [21]-[24], there are several reasons why this might not always happen.

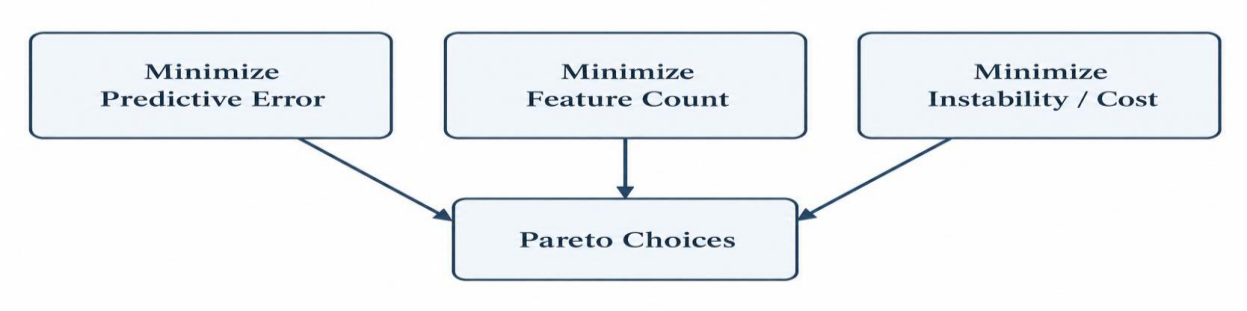

Figure 3. Competing objectives in selection: loss, number of features, and cost or instability (conceptual).

6.4. Implications for benchmarking

Researchers should benchmark against all-features regularized logistic regression, a well-tuned tree ensemble, and a simple filter baseline, not only against weaker or unoptimized comparators. A “hybrid” method is scientifically useful when it improves a prespecified objective within a robust protocol; it need not be best on every metric.

VII. RESEARCH CHALLENGES AND IDENTIFIED GAPS

Small biomedical benchmarks are especially vulnerable to this type of bias. This is due to the fact that when you perform repeated model selection (and tune parameters based on the results), you may unintentionally use the test set as part of your development process. To produce publication-ready work, you must document how and in what order each of your transformations took place and ensure that you do not tune any parameters once you have evaluated the performance of your models on the test set [21], [22].

The first challenge with this type of analysis is unstable selection. Resampling can lead to different gene lists or panel of clinical variables being identified as "significant" (even if they result in similar predictive ability). This poses a problem for interpreting the results since we need to consider how well-chosen the selected variables are; whether those variables will be available for use at the time of deployment; and if the variables used are ethical. Another challenge is called "external transportability." Institutions may have different population prevalences, different measurement protocols, and different ways of organizing and delivering health care [5], [6], [20].

Table 4. Research gaps, consequences and corrective designs

Gap	Consequence	Recommended improvement
Feature-selection leakage	Optimistic test metrics	Nested pipelines with locked outer holdouts
Small or dependent samples	Overconfident uncertainty	Patient/group disjoint resampling
Unstable subsets	Unreliable biomarker lists	Repeated fits and selection-frequency reports
Expensive wrappers	Weak reproducibility	Time/energy budget and fixed tuning policy
Limited calibration	Unsafe risk interpretation	Calibration plots and proper scoring rules
No external source	Unknown transportability	Independent temporal/site-level test
Missing acquisition cost	Impractical selected panel	Incorporate test cost and availability

7.1. Evidence quality versus apparent novelty

Metaheuristic renaming is not equivalent to a methodological advance. Novelty must be demonstrated through a distinct search mechanism or objective and evaluated under comparable computational budgets. Clinical applicability is an additional evidence layer, not a consequence of high internal AUROC [1]– [5], [21], [22].

VIII. FUTURE RESEARCH DIRECTIONS

Future research needs to focus on creating stable acquisition-aware panel designs, rather than simply maximizing the accuracy of binary feature vectors. Multi-objective optimization techniques can produce a Pareto optimal design frontier (set) that considers multiple objectives including predictive loss, the number of features used, stability of the selected features, and the financial and/or temporal costs associated with collecting data on those features. When evaluating future research, evaluations should utilize objective selection criteria and pre-specified decision-making criteria, as opposed to selecting the best-performing design from among all evaluated designs after evaluating the test set.

A second area for further exploration is the application of feature selection under dataset shift. Evaluation of the utility of a compact panel through repeated testing in different settings (i.e. hospitals), times (i.e. over time), assays (i.e. platform), and demographic groups will help determine if a compact panel remains relevant outside of the development setting. Additionally, domain knowledge and missing data mechanisms may guide the process of selecting features for interpretation. However, the fidelity of the explanation and the clinical utility of the compact panel should be assessed independently [15], [19]–[22].

8.1. Evidence from recent studies

Recent research primary studies provide evidence of a much broader range of applications of hybrid selection than simply the application to small tabular benchmarks. Two of these primary studies are the two-stage breast cancer optimization study [33], which is focused on optimizing the breast cancer detection process, and the clinical breast cancer survivability selection study [35], which is focused on identifying the most relevant features to predict patient survival from breast cancer. Because each of these primary studies has a very different focus, they cannot be directly compared regarding their claimed accuracies. The Harris Hawk framework [30] is designed to help formalize the trade-offs involved in selecting subsets of features. Cardiovascular feature engineering [28] is also designed to address the issue of whether or not to transform the measured feature space. A sequential population transfer study in 2026 [37] found that although it was possible to create smaller subsets of features while maintaining the majority of the original accuracy, there was no statistically significant overall accuracy improvement when considering multiple comparisons. This example contradicts the idea that creating a hybrid model will always result in improved prediction.

The type of analytical unit can differ across domains as well, with mammogram images [34], histology images [29], breast cancer survival data [35] and EHR histories [32] having different avenues for information loss (leakage paths) and varying types of decision-making environments. For high-dimensional/low-sample biomedical data [39], it is especially important to select features so they do not fall out of the training fold. A reliable comparison would keep the following factors consistent: endpoint, method of dividing up data, family of classification algorithm used, amount of resources allocated for searching, and outcome measure. Comparisons based on rankings of models

developed on un-matched data are unable to demonstrate that one model is superior to another.

IX. DISCUSSION AND CONCLUSION

Hybrid feature selection is most accurately seen as a design family rather than a single superior algorithm. The main reasons for using hybrid feature selection are to use an inexpensive screening process followed by a model sensitive search, or to find a balance between the predictive ability of the model and its stability, as well as the amount of effort required to obtain the measurements. Many published examples of studies in diabetes, cardiovascular disease, cancer, and chronic kidney disease show how wide-ranging the work being done with hybrid feature selection is today. However, these examples cannot be used to rank studies across different research areas based on their overall level of accuracy due to the differences in cohorts, labels, preprocessing methods, and evaluation metrics [7]-[14].

The goal of this critical review is to develop minimal evidence standards for algorithms used in clinical research; I am not proposing a novel validated algorithm. I propose that authors document where their data came from; clearly state what they are attempting to select (i.e. pre-define the selection objective); perform all data transformations within their training fold(s); use a simple regularized baseline; report subset stability and run times; and quantify uncertainty regarding performance on "held-out" patients or institutions if possible. Since my literature search was limited to articles that were relevant to my research questions and I could not extract complete full-text information from every article, I am unable to provide pooled effects estimates or a complete evidence-based ranking. Therefore, the findings of my review represent promising approaches to clinical research, but do not establish the clinical efficacy of any particular approach.

Author and submission declarations

The manuscript is a narrative literature review; no new patients were recruited or clinical data collected. Author affiliations, funding declaration, conflict-of-interest statement, and any use-of-assistance disclosure must be completed accurately by the submitting author before journal submission.

REFERENCES

[1] Kaur *et al*., "A systematic review on metaheuristic optimization techniques for feature selections in disease diagnosis: Open issues and challenges," *Arch. Comput. Methods Eng.*, 2023, doi: 10.1007/s11831-022-09853-1.

[2] Houssein *et al*., "Integrating metaheuristics and artificial intelligence for healthcare: Basics, challenging and future directions," *Artif. Intell. Rev.*, 2024, doi: 10.1007/s10462-024-10822-2.

[3] Barrera-García *et al*., "Feature selection problem and metaheuristics: A systematic literature review about its formulation, evaluation and applications," *Biomimetics*, vol. 9, no. 1, 2024, doi: 10.3390/biomimetics9010009.

[4] Nssibi *et al*., "Advances in nature-inspired metaheuristic optimization for feature selection problem: A comprehensive survey," *Comput. Sci. Rev.*, 2023, doi: 10.1016/j.cosrev.2023.100559.

[5] Barbieri *et al*., "Analysis and comparison of feature selection methods towards performance and stability," *Expert Syst. Appl.*, 2024, doi: 10.1016/j.eswa.2024.123667.

[6] Łukaszuk *et al*., "Stability of feature selection in multi-omics data analysis," *Appl. Sci.*, vol. 14, no. 23, Art. no. 11103, 2024, doi: 10.3390/app142311103.

[7] Talari *et al*., "Hybrid feature selection and classification technique for early prediction and severity of diabetes type 2," *PLOS ONE*, 2024, doi: 10.1371/journal.pone.0292100. Expression of concern: doi: 10.1371/journal.pone.0324149, 2025.

[8] Narasimhan and Victor, "A hybrid approach with metaheuristic optimization and random forest in improving heart disease prediction," *Sci. Rep.*, 2025, doi: 10.1038/s41598-024-73867-x.

[9] Praveen *et al*., "Enhanced feature selection and ensemble learning for cardiovascular disease prediction: Hybrid GOL2-2 T and adaptive boosted decision fusion with babysitting refinement," *Front. Med.*, 2024, doi: 10.3389/fmed.2024.1407376.

[10] Yogesh *et al*., "Novel statistically equivalent signature-based hybrid feature selection and

ensemble deep learning LSTM and GRU for chronic kidney disease classification," *PeerJ Comput. Sci.*, 2024, doi: 10.7717/peerj-cs.2467.

[11] Sirmayanti *et al*., "A systematic literature review of diabetes prediction using metaheuristic algorithm-based feature selection: Algorithms and challenges method," *Appl. Comput. Sci.*, 2025, doi: 10.35784/acs_6849.

[12] Gupta *et al*., "heart disease prediction using a hybrid feature selection and ensemble learning approach," *IEEE Access*, 2025, doi: 10.1109/ACCESS.2025.3583757.

[13] Shilpa *et al*., "A hybrid feature selection with data-driven approach for cardiovascular disease prediction using machine learning," *IAES Int. J. Artif. Intell.*, vol. 14, no. 2, pp. 1192–1200, 2025, doi: 10.11591/ijai.v14.i2.pp1192-1200.

[14] Lonang *et al*., "Hybrid feature selection for effective heart disease detection: A multi-algorithm machine learning approach," *Sci. J. Informatics*, vol. 13, no. 1, 2026, doi: 10.15294/sji.v13i1.38815.

[15] Y. Takefuji, "Beyond SHAP: Reliable feature selection methods for clinical prediction models," *Arch. Gerontol. Geriatr.*, 2025, doi: 10.1016/j.archger.2025.105873.

[16] Tan *et al*., "Optimized feature selection and deep neural networks to improve heart disease prediction," *J. Digit. Imaging*, 2025, doi: 10.1007/s10278-025-01435-4.

[17] Dissanayake *et al*., "heart disease diagnostics using meta-learning-based hybrid feature selection," *Appl. Comput. Intell. Soft Comput.*, 2024, Art. no. 8800497, doi: 10.1155/2024/8800497.

[18] Yogesh *et al*., "Chronic kidney disease classification through hybrid feature selection and ensemble deep learning," *Int. J. Stat. Med. Res.*, vol. 14, 2025, doi: 10.6000/1929-6029.2025.14.11.

[19] Agrawal *et al*., "eXplainable artificial intelligence-Eval: A framework for comparative evaluation of explanation methods in healthcare," *DIGITAL HEALTH*, 2025, doi: 10.1177/20552076251368045.

[20] Tejani *et al*., "Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 update," *Radiol. Artif. Intell.*, 2024, doi: 10.1148/ryai.240300.

[21] T. P. A. Collins *et al*., "TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods," *BMJ*, 2024, doi: 10.1136/bmj-2023-078378.

[22] K. G. M. Moons *et al*., "PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods," *BMJ*, 2025, doi: 10.1136/bmj-2024-082505.

[23] M. A. Maier-Hein *et al*., "Metrics reloaded: Recommendations for image analysis validation," *Nat. Methods*, 2024, doi: 10.1038/s41592-023-02151-z.

[24] J. P. Reinke *et al*., "Understanding metric-related pitfalls in image analysis validation," *Nat. Methods*, 2024, doi: 10.1038/s41592-023-02150-0.

[25] Kumar *et al*., "Hybrid bio-inspired computing in medical image data analysis: A review," *Inf. Discov. Deliv.*, 2025, doi: 10.3233/IDT-241023.

[26] Gochhait, "Feature selection with metaheuristic algorithms: A review of recent developments (2020–2025)," *Metaheuristic Algorithms Appl.*, 2025, doi: 10.48313/maa.v2i1.34.

[27] M. A. Islam, M. Z. H. Majumder, M. S. Miah, and S. Jannaty, "Precision healthcare: A deep dive into machine learning algorithms and feature selection strategies for accurate heart disease prediction," *Comput. Biol. Med.*, vol. 176, Art. no. 108432, 2024, doi: 10.1016/j.compbiomed.2024.108432.

[28] D. Yadav, D. Rani, and O. P. Verma, "Impact of feature selection and feature engineering in prediction of cardiovascular diseases," *Comput. Biol. Med.*, vol. 197, Art. no. 111027, 2025, doi: 10.1016/j.compbiomed.2025.111027.

[29] M. Alhassan, "An improved breast cancer classification with hybrid chaotic sand cat and Remora Optimization feature selection algorithm," *PLOS ONE*, vol. 19, no. 4, Art. no.

e0300622, 2024, doi: 10.1371/journal.pone.0300622.

[30] "A recommender system with multi-objective hybrid Harris Hawk optimization for feature selection and disease diagnosis," *Healthcare Analytics*, 2025, doi: 10.1016/j.health.2025.100384.

[31] "Federated learning-based disease prediction: A fusion approach with feature selection and extraction," *Biomed. Signal Process. Control*, 2025, doi: 10.1016/j.bspc.2024.106961.

[32] R. Zhang *et al*., "DeepSelective: Interpretable prognosis prediction via feature selection and compression in EHR data," *Pattern Recognit.*, vol. 174, Art. no. 112970, 2026, doi: 10.1016/j.patcog.2025.112970.

[33] T. Karpagalingam and J. J. J. Britto, "A two-stage feature selection approach using hybrid quasi-opposition self-adaptive coati optimization algorithm for breast cancer classification," *Appl. Soft Comput.*, vol. 146, Art. no. 110704, 2023, doi: 10.1016/j.asoc.2023.110704.

[34] W. M. Shaban, "Insight into breast cancer detection: New hybrid feature selection method," *Neural Comput. Appl.*, vol. 35, pp. 6831–6853, 2023, doi: 10.1007/s00521-022-08062-y.

[35] E. J. Sweetlin and S. Saudia, "A new hybrid feature selection sequence for predicting breast cancer survivability using clinical datasets," *Intell. Autom. Soft Comput.*, vol. 37, no. 1, pp. 343–367, 2023, doi: 10.32604/iasc.2023.036742.

[36] "Explainable ensemble-based machine learning model for polycystic ovary syndrome detection using hybrid feature selection," *Int. J. Inf. Technol.*, 2025, doi: 10.1007/s41870-025-03044-4.

[37] C. Lazrak, A. Bouayad, and A. Talha, "A sequential hybrid framework for medical feature selection using population transfer," *Discover Artif. Intell.*, vol. 6, Art. no. 1429, 2026, doi: 10.1007/s44163-026-02217-9.

[38] "Optimizing medical data classification: Integrating hybrid fuzzy joint mutual information with binary Cheetah optimizer algorithm," *Cluster Comput.*, 2025, doi: 10.1007/s10586-025-05102-9.

[39] H. Shin and S. Oh, "An effective heuristic for developing hybrid feature selection in high dimensional and low sample size datasets," *BMC Bioinformatics*, vol. 25, Art. no. 390, 2024, doi: 10.1186/s12859-024-06017-9.

[40] R. Lamba *et al*., "An explainable machine learning framework integrating hybrid feature selection and interaction modeling for hepatitis C stage prediction," *Discover Computing*, vol. 29, Art. no. 355, 2026, doi: 10.1007/s10791-026-10292-w.