Scientific Reports

https://doi.org/10.1038/s41598-026-53812-w

Article in Press

Structured vital sign prediction in hospital environments via an Al-Biruni earth radius optimization–driven unified metaheuristic framework

Received: 1 March 2026

Accepted: 14 May 2026

Published online: 20 May 2026

Cite this article as: Alzakari S.A., Eid M.M., Alhussan A.A. *et al.* Structured vital sign prediction in hospital environments via an Al-Biruni earth radius optimization–driven unified metaheuristic framework. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-53812-w

Sarah A. Alzakari, Marwa M. Eid, Amel Ali Alhussan & S. K. Towfek

Structured Vital Sign Prediction in Hospital Environments via an Al-Biruni Earth Radius Optimization–Driven Unified Metaheuristic Framework

Sarah A. Alzakari[1], Marwa M. Eid[2,3], Amel Ali Alhussan[1], S.K. Towfek[4*]

[1]Department of Computer Sciences, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia.

[2]Faculty of Artificial Intelligence, Delta University for Science and Technology, Mansoura, Egypt.

[3]Jadara Research Center, Jadara University, Irbid, Jordan.

[4*]Computer Science and Intelligent Systems Research Center, Blacksburg 24060, Virginia, USA.

*Corresponding author(s). E-mail(s): sktowfek@jcsis.org;

Contributing authors: saalzakari@pnu.edu.sa; mmm@ieee.org; aalhussan@pnu.edu.sa;

Abstract

Accurate prediction in structured hospital monitoring data is challenging because inpatient datasets are high-dimensional and often contain redundant features and suboptimal hyperparameter settings. This problem is important because unreliable prediction can limit the effectiveness of hospital monitoring and clinical decision support. To address this, this study proposes a unified optimization framework that integrates the Al-Biruni Earth Radius (BER) metaheuristic with the Feature-Transformed Learning Model (FTLM) for both binary feature selection and continuous hyperparameter optimization. BER is first applied in a discrete search space to identify informative subsets of vital-sign, demographic, clinical, and temporal variables from the *Patient Vital Signs and Event Tracking* dataset, and then in a continuous space to tune FTLM hyperparameters under the same computational budget used for competing optimizers, including GWO, PSO, BA, WAO, SBO, SCA, FA, GA, and SAO. At baseline, FTLM achieved a mean squared error (MSE) of 0.012028 and $\boldsymbol{R^2}$ of 0.782413. After BER-based feature selection, performance improved to an MSE of $\mathbf{9.80 \times 10^{-3}}$ and $\boldsymbol{R^2}$ of 0.860654, with correlation of 0.848181 and Nash–Sutcliffe efficiency of 0.879577. Following BER-driven hyperparameter optimization, FTLM attained an MSE of

$\mathbf{7.43 \times 10^{-7}}$, RMSE of $\mathbf{8.62 \times 10^{-4}}$, correlation coefficient of 0.955593543, $\boldsymbol{R^2}$ of 0.961124043, and Willmott Index of 0.963281686, achieving the strongest empirical performance among the evaluated optimizers under the same experimental setting. To further assess generalizability, an external validation experiment was conducted on an independent Human Vital Sign Dataset containing 200,000 samples, where BER + FTLM again achieved the strongest empirical performance among the evaluated optimizers. These findings show that BER provides stable convergence, reduced variance, and strong predictive alignment for structured clinical data modeling.

Keywords: Metaheuristic optimization, Feature selection, Hyperparameter tuning, Vital sign prediction, Clinical data modeling

1 Introduction

1.1 Clinical Context and Technical Background

Constant inpatient monitoring is becoming a standard aspect of hospital care particularly in the setting in which preventable physiological degeneration is typical and timely reaction denotes the dependency on timely acknowledgment of deviant patterns as opposed to separate values and measurements of the patient status s/he varies over time Subbe et al. (2001b); Smith et al. (2014); Churpek et al. (2016). In contemporary wards and high-dependency environments, this emphasis on patterns rather than isolated readings reflects a practical reality: clinical risk often accumulates through gradual changes, intermittent instability, or recurrent excursions beyond acceptable ranges, and the operational challenge is to recognize these developments early enough to trigger proportionate responses. Accordingly, clinicians increasingly interpret monitoring not as a sequence of disconnected measurements but as evolving evidence of physiologic adaptation, compensation, and eventual decompensation. In this framing, the objective of monitoring is not merely documentation, but the creation of actionable situational awareness—a process in which the temporal organization of observations matters as much as their numeric valuesAgrawal et al. (2025).

As a concept, the heart rate (HR), respiratory rate (RR), systolic blood pressure (SBP), body temperature (TEMP), or peripheral oxygen saturation (SPO_2) are repeatedly written on the bedside systems. Although these signals are typically treated as "basic" observations, their clinical meaning is highly contingent on context and clinical intent. For instance, the same HR or RR may be interpreted differently depending on age, baseline functional status, medication exposure, acute pain, fever, or the presence of respiratory support. Clinical meaning of these signals is just that they are observed in context, e.g., demographic variables like age and sex, structured measures of consciousness like AVPU, respiratory support reported (e.g., mask type), a series of timestamped clinical events that delineate changing care processes Brunker and Harris (2015); Escobar et al. (2020). This contextual dependence is not an ancillary consideration but a defining property of inpatient physiology: vital signs act as surface indicators of deeper processes, while structured descriptors and event markers provide the explanatory substrate that links measurement to mechanism, care delivery, and trajectory. That is, the current practice of monitoring vital-signs on a regular basis is already becoming a type of multivariate surveillance system—its clinical usefulness increases when measurements are combined instead of being conducted as an independent snapshot. From a modeling standpoint, this implies that principled

learning approaches should aim to preserve cross-variable interactions and temporal dependencies, rather than compressing the patient state into an overly simplified representationAgrawal and Panda (2025b).

After adding physiological values to the temporal and descriptive contexts, bedside observations made by hand are transformed into a longitudinal clinical observation. Practically, the shift from a single measurement to a longitudinal record changes the inferential target: the question becomes not only "what is the value now?" but also "how did we get here, how rapidly is change occurring, and what is likely to happen next?" The admission process through the use of admission time (`admittime`) and discharge process through discharge time (`dischtime`) boundaries the complete episode of care and the timestamps in the middle (e.g., `charttime`, `eventtime`) would indicate observationAgrawal and Panda (2025a), intervention or alteration of the clinical condition. These time stamps operationalize the episode-of-care narrative, enabling analyses that align physiologic observations with clinical workflow and event timing. The sequential dependence (e.g., being able to state what happens before and after the event), time-dependent dependence (e.g., how much time elapses before first event) are implicitly encoded by variables (e.g., `prev_event`, `next_event`) and work across time—dependencies are normally invisible in cross-sectional summaries but play definitive roles in detecting deterioration Churpek et al. (2016); Roberts et al. (2017). In other words, a patient record that supports "before/after" reasoning and explicit elapsed-time quantification can express clinically meaningful dynamics such as escalation, delayed response, recovery periods, and recurrent instability. This interpretation through the use of trajectories has been corroborated by the more general movement of clinical machine learning towards modeling the progression of the patient condition as a time-varying process rather than a time-static vector, in which prediction performance and clinical utility depend on encoding patient condition as a time-varying process instead of a time-fixed vector as well as on the representation of time-varying processes as time-varying trajectories Lipton et al. (2016); Rajkomar et al. (2018b). Importantly, this trend is motivated not only by accuracy considerations, but also by the clinical logic that decisions are made in time: a model that is sensitive to trend, volatility, and recency is more naturally aligned with the way bedside decisions are justified and documented.

This paper draws on the publicly available data on the Kaggle platform of the dataset entitled the *Patient Vital Signs and Event Tracking*, which gathers structured hospital data on a tabular scale with seventeen variables used in the study ParmaJha (2024). The dataset provides an instructive setting because it combines physiologic measurements with descriptive and event-oriented variables in a way that resembles many operational hospital databases: measurements appear repeatedly across an admission, while event fields encode discrete changes in care processes and patient state documentation. The following variables comprise continuous physiological values (HR, RR, SBP, TEMP, SPO_2), demographic characteristics (age, sex), categorical clinical characteristics (AVPU, mask type), and various time-based fields specifying admission episodes and series of events. Since the dataset contains repeated observations at the episode level during admissions, it enables supervised learning formulations that are physiologically outcome-oriented and event-related, but also contains the practical constraints and peculiarities that are inherent to routinely collected clinical data. Such constraints include measurement noise, heterogeneity in documentation practice, potential missingness or irregular sampling, and the presence of mixed data types that must be reconciled into a coherent representation suitable for learning algorithms.

As a machine learning problem, the given dataset presents the challenge that is inherent to hospital tabular data and cannot be properly dealt with the assumption of independent and identically distributed (i.i.d.) samples. This is not a purely technical

caveat: violations of i.i.d. assumptions often reflect the underlying clinical process (repeated measurement of the same patient, time-varying acuity, interventions that change physiology, and care pathways that introduce structured correlations). To begin with, the heterogeneity of features is especially strong: physiological variables measured on different numeric scales and units; categorical ones need to be coded in ways that allow clinical semantics to be retained; and temporal ones often have to be converted into durations or relative time to form a stable learning Goodfellow et al. (2016). In practice, these transformations are consequential: they can either preserve or distort the relationships that the model must learn, particularly when categories encode ordered clinical severity (as with consciousness scales) or when time is represented in a way that inadvertently introduces leakage or spurious predictive cues. Second, the same admission identifier may have multiple rows, creating within-admission correlation and increasing the likelihood that naive random splitting would spill patient-trajectory information across partitions. This risk is especially salient in longitudinal clinical datasets because random splits can place earlier and later measurements from the same admission into different partitions, artificially inflating apparent generalization. Third, clinical measurement distributions are often skewed, bounded, and irregularly sampled; these characteristics can induce optimization behavior, model calibration and generalization with no action taken against these characteristics LeCun et al. (2015); Goodfellow et al. (2016). Consequently, even when models achieve low error on held-out data under naive validation, their calibration and stability may degrade in settings where sampling frequency, measurement noise, and event prevalence differ.

These structural features put real demands on both data preparation and strategy in data modeling. Numerical normalization is required to fill-in discrepancies in scale without destroying clinically significant magnitudes. In clinical tabular data, normalization must be handled carefully: scaling should support efficient optimization while preserving interpretability and avoiding disproportionate emphasis on high-variance variables that may not be clinically central. Categorical encoding should take into consideration the nominal or ordinal design of clinical tools like AVPU. This point is particularly important because encoding choices can impose implicit geometry on the data: ordinal encodings can reflect graded severity, while one-hot encodings can prevent the model from inventing an artificial ordering among categories. The temporal characteristics should be designed and tested with clear protections against leakage, including validation designs that respect dependence structures, instead of merely depending on random resampling alone Roberts et al. (2017). Temporal characteristics have to be designed and tested with look-ahead safeguards against leakage, and validation designs have to acknowledge dependence structure, not just, again, looking solely to random resampling. In practical terms, this means that any feature derived from event timing must be computed in a manner consistent with what would be known at prediction time, and evaluation must reflect how the model would be deployed (e.g., within-admission forecasting rather than retrospective reconstruction). Within such limits, successful modeling cannot be merely a case of selecting a learning algorithm, but we must have a unified pipeline where representation decisions, feature selection, and hyperparameter selection are made as joint design choices. This pipeline perspective is essential because in mixed-type, longitudinal clinical tabular data, choices that appear “preprocessing-related” can materially affect what patterns are learnable, what correlations are exploited, and how robust the system is to distributional shifts. Such a coupling is especially fateful in clinic, where the slight amelioration in error measures may prove clinically insignificant provided it is accomplished by irregular representations or leakage-sensitive testing. Thus, methodological rigor must prioritize validity and stability alongside performance, particularly when the downstream intent is to support clinical monitoring or risk stratification.

1.2 Problem Statement

Structured clinical data predictive accuracy of machine learning models on structured clinical data depends on both architecture choice and confluence between representation and optimization. In this setting, "architecture" includes not only the selection of a learning family (e.g., tree-based models, neural models, or hybrids) but also the practical structure of the training objective, the regularization regime, and the evaluation protocol. The confluence between representation and optimization is especially central in tabular hospital data because the signal is distributed across heterogeneous variables, many of which are only informative when combined with context or temporal alignment. The relevance of the chosen input features, as well as the model hyperparameters configuration regulating training dynamics and the generalization behavior decide two factors in this interaction. Put differently, the model can only learn what the representation makes available, and optimization can only reliably identify strong solutions if the search space is structured and constrained in a way that avoids unstable or misleading shortcuts.

The issue of feature selection is automatically non-trivial in hospital monitoring datasets. Feature relevance may be conditional, context-dependent, and temporally modulated: for example, the importance of oxygen saturation can vary with respiratory support, and the informativeness of RR can depend on measurement quality and documentation patterns. If physiological variables are correlated, fields of temporal derivation may represent overlap of information on the timing of events and categorical descriptors can have clinically significant stratification where such phenomena appear weakly associated by marginal statistics. This motivates feature selection strategies that consider multivariate interactions rather than relying solely on univariate screening. Redundant features may scale up dimensionality without adding information and therefore raise estimator variance, or pruning can bias the model towards partial patient state capture; also spurious data in a patient can introduce bias that drives the model to other patterns other than the optimal predictions; overfitting may also represent a variant of bias potentially leading to overfitting or underfitting depending on the data conditioning the model and the predictor that acts upon it, and furthermore, the opposite effect is possible due to dimensionality inversion Guyon and Elisseeff (2003a). In this context, feature selection must be framed as a bias–variance management problem under heterogeneity and dependence: removing variables can improve stability and interpretability, but excessive pruning can eliminate weak yet complementary signals that are crucial when combined with other variables. Also, heterogeneous scales of features are a complicating factor: unless scaled appropriately, an optimizer can preferentially update a big-value variable, which will both affect the final parameter estimates and compromise the convergence stability of the process itself Goodfellow et al. (2016). Irregular sampling, measurement noise, and pattern of rare events further complicate undertaking the process of identifying consistent, informative subsets. As a result, feature selection in hospital data is not only about reducing the number of variables; it is about improving the fidelity of the learned mapping under realistic data constraints and evaluation safeguards.

The second layer of complexity is brought up with hyperparameter configuration. Even when a model class is fixed, the effective behavior of the model depends strongly on hyperparameters that control regularization strength, capacity, learning dynamics, and optimization stability. Hyperparameters of learning models, including learning rate, regularization, and capacity, may have a significant impact on convergence, stability, and the bias–variance trade-off. In clinical prediction settings, this sensitivity can be amplified because the data contain correlated samples, mixed feature types, and potentially nonstationary temporal patterns. Gridsense hyperparameter spaces on a large scale are computationally infeasible, and even random (as opposed to

grid) search, although more effective than grid search in most practical scenarios, is unguided and unproductive in large regions of the search space containing near-optimal configurations Bergstra and Bengio (2012a). This inefficiency is particularly problematic when training requires repeated fitting and validation across multiple partitions designed to prevent leakage, since each evaluation becomes expensive. Due to this reason, optimization without derivatives is useful where it is too costly to compute the objective surface, which is nonconvex or which can be formulated in terms of repeated training and validation steps; so-called derivative-free metaheuristic optimization is also a viable alternative in such cases Yang (2010a). Such methods are attractive in practice because they can operate over mixed or constrained search spaces, accommodate noisy objective estimates arising from stochastic training and validation, and provide flexible exploration–exploitation trade-offs.

Importantly, both the feature selection and hyperparameter optimization rely on each other. Feature subset choices change the effective dimensionality, noise profile, and correlation structure of the input, which in turn can alter the hyperparameter settings that yield stable training and good generalization. Conversely, the hyperparameter regime (e.g., regularization and capacity) influences which features appear useful: strongly regularized models may suppress subtle signals, while high-capacity models may exploit idiosyncratic correlations unless constrained by careful validation. The most common hyperparameter choices can be determined by the dimensionality of the input, the distribution of features, and the noise structure and the apparent utility of features might vary depending on the different regularization constraints and model capacities. The separation of such processes into separate stages can consequently produce inefficient pipelines on a global scale despite the effectiveness of all stages in isolation. Therefore, the methodological challenge is to coordinate these choices so that feature selection does not merely optimize an intermediate criterion, and hyperparameter tuning does not implicitly compensate for representational weaknesses or leakage-prone artifacts.

In order to plan the solution of this twofold optimization problem, the current work explores a combined approach, which is the Al-Biruni Earth Radius (BER) optimization algorithm El-kenawy et al. (2023). BER is utilized in two roles and coordinated. It first consists of searching in a binary space to select informative subsets of features in the available variables. This formulation is appropriate for wrapper-style feature selection, where candidate subsets are evaluated through the downstream predictive performance they enable, thereby aligning the selection criterion with the modeling goal. Second, it is scaled to a continuous space of hyperparameters optimization of the FTLM model applied to this paper. This centre of use facilitates a coherent exploration–exploitation scheme on both discrete and continuous spaces without compromising an identical optimization basis. Conceptually, this allows a consistent optimization philosophy to govern both "what information to present to the model" (feature subset) and "how the model should learn from that information" (hyperparameters), while still respecting the structural differences between binary and continuous decision variables.

It should be stressed that neither the BER nor the FTLM is presented as innovative algorithmic contributions. They have each been previously established. The methodological value of this research is that it analyzes their combined use in a systematic clinical modeling pipeline and compares the performance of the resulting system to an established metaheuristic optimizers, controlled in experimental settings. In this sense, the contribution is best understood as an integrated experimental study: it emphasizes disciplined pipeline design, controlled benchmarking, and a transparent decomposition of performance gains across baseline modeling, feature selection, and hyperparameter optimization.

Despite the growing use of machine learning in hospital monitoring, reliable prediction from structured inpatient data remains methodologically difficult. The challenge is not only the heterogeneity of physiological, demographic, clinical, and temporal variables, but also the fact that predictive performance depends jointly on feature representation and hyperparameter configuration. A clear research gap therefore remains: existing studies often examine feature selection and hyperparameter optimization separately, or do not evaluate them within a unified and leakage-aware framework under controlled computational settings for structured hospital data. As a result, it remains unclear whether reported improvements arise from better feature representation, better model calibration, or differences in evaluation design. To address this gap, the present study proposes a unified BER–FTLM framework in which the Al-Biruni Earth Radius optimizer is used in both binary and continuous search spaces to perform feature selection and hyperparameter optimization under controlled benchmarking and a reproducible validation protocol.

BER was selected as the core optimizer in this study because it provides a unified search mechanism that can be adapted to both binary and continuous optimization spaces, which is particularly suitable for the coupled feature-selection and hyperparameter-tuning problem considered here. This makes BER practically attractive for structured clinical data modeling, where the optimization landscape is heterogeneous, noisy, and nonconvex. In addition, BER offers an adaptive balance between exploration and exploitation and includes diversity-preserving search behavior, which can be beneficial when repeated model training and validation make objective evaluation computationally expensive and potentially unstable.

The main contributions of this work are methodological and experimental rather than algorithmic. First, it presents a unified optimization pipeline for structured hospital vital-sign prediction in which BER is applied in both binary and continuous search spaces to perform feature selection and hyperparameter optimization within a single methodological framework. Second, it establishes a leakage-aware and reproducible evaluation setting for heterogeneous inpatient monitoring data, including controlled preprocessing, validation, and benchmarking conditions. Third, it provides a progressive experimental analysis that separates baseline modeling, BER-based feature selection, and BER-based hyperparameter optimization, thereby making the source of performance gains transparent. Fourth, it offers a controlled comparison against established metaheuristic optimizers under identical computational budgets. Accordingly, the contribution of this study lies not in proposing a new optimizer or a new prediction model, but in the rigorous integration, evaluation, and benchmarking of an optimization-aware BER–FTLM pipeline for structured clinical data.In this sense, the value of the work is to provide a carefully controlled and domain-relevant experimental framework that clarifies when and how combined optimization of representation and learning dynamics can improve structured hospital prediction.

1.3 Research Objectives

The current research is also directed by mutually connected goals that serve to guarantee methodological rigor, reproducibility, and comparative transparency in the framework of structured clinical data, based on continuous inpatient monitoring. These objectives are formulated to address not only model performance, but also the validity of evaluation, the stability of improvements across experimental phases, and the interpretability of pipeline decisions in a clinical context.

To begin with, the paper builds a reproducible modeling system that is specific to the case under investigation, which involves preprocessing stage, transformation of features, training of the model, validation score and assessment within controlled experimental

setting in expressly controlled conditions ParmaJha (2024). Reproducibility here entails that each experimental stage is defined by explicit steps, consistent transformations, and a fixed protocol for training and evaluation so that results can be re-derived and meaningfully compared. Special care is paid to making sure that normalization, categorical encoding, and time processing are consistent in all comparative experiments, and assessment does not use leakage-prone splitting, which can overstate apparent performance Roberts et al. (2017). This emphasis is central because in longitudinal clinical datasets, seemingly minor deviations in splitting strategy or feature derivation can yield overly optimistic estimates that do not translate to deployment conditions.

Second, BER would be utilized in binary search scenario of feature subset selection. In this formulation, every possible solution on the candidate set represents an inclusion or exclusion choice made over the collection of variables, and the optimization will look at constant subsets which enhance the forecasting precision and can restrain dimensionality. Beyond computational efficiency, this objective also supports practical modeling considerations: reduced dimensionality can improve robustness, reduce susceptibility to noise, and promote clearer attribution of performance to clinically meaningful inputs, provided that selection is conducted under a rigorous validation protocol.

Third, BER is generalized in the case of continuous hyperparameter optimization of FTLM. Search space parameters defined over bounded real-valued parameters control model learning dynamics, and BER finds configurations that lead to better generalization with a fixed computational budget El-kenawy et al. (2023). This objective recognizes that in applied clinical modeling, tuning must be both effective and resource-conscious: it must identify strong configurations without exhaustive search, and it must do so under evaluation schemes that avoid leakage and overfitting to a particular split.

Fourth, the predictive performance is determined with the help of a detailed combination of standardized regression measures on various experimental phases, such as a baseline analysis, the performance following the selection of features and the performance following hyperparameter optimization. The multi-metric protocol is embraced to contribute to the prevention of over-interpretation of single-number summaries and to describe the performance based on the degree of errors, bias, and association structure. This is particularly relevant for clinical regression tasks, where different error characteristics can imply different operational risks (e.g., systematic bias versus occasional large deviations), and where clinical acceptability may depend on stability across patient strata rather than average performance alone.

The research of this work is both methodological and experimental. First, it builds a hierarchical preprocessing and validation chain trained on nonhomogenous hospital monitoring data, explicit management of mixed feature types, and time dependence recognition of mixed feature types and time dependence Roberts et al. (2017). This goal emphasizes that the evaluation pipeline must be treated as part of the scientific contribution, since inappropriate validation can invalidate conclusions even when modeling is sophisticated. Second, it tests a single optimization framework that incorporates feature selection and hyperparameter optimization in a single metaheuristic underpinning and is driven by BER management El-kenawy et al. (2023). Third, it uses a progressive benchmarking protocol to isolate the effect of every optimization component and allows the transparent comparison across competing strategies. Fourth, it employs a multi-metric assessment scheme to offer a more detailed description of predictive behavior not just in one accuracy measure. Collectively, these aims align the study with best practices in applied machine learning for health: improvements are interpreted only insofar as they arise under controlled, leakage-aware evaluation, and their magnitude is contextualized across complementary metrics and experimental phases.

The rest of the paper is structured in the following manner. Section 2 is a review of related literature on vital-sign modeling and metaheuristic optimization. Section 3

presents the dataset and the research methodology that is proposed. The experimental set up is described in Section 4. The empirical results are provided in Section 5. In Section 6, methodological implications and limitations are discussed. Section 7 is a conclusion part where the future research directions are stated.

2 Literature Review

Three somewhat overlapping strains of predictive modeling on inpatient monitoring data have been developed: (i) clinically inspired early-warning systems based on physiological thresholds and summary scoring, (ii) data-driven learning on structured EHR and bedside-monitoring variables, and (iii) optimization strategies that serve the two most significant controls governing generalization in heterogeneous tabular contexts, feature subset composition and hyperparameter configuration. Each of these strands has evolved in response to practical clinical demands: the need for interpretability and rapid bedside deployment, the availability of increasingly large and complex digital health records, and the recognition that performance gains in structured data settings are often contingent on disciplined control of representation and learning dynamics. Although the individual trajectories are well-understood in isolation, the literature has a practical gap, namely relatively few studies impose strict computational parity and simultaneously interrogate binary feature selection and continuous tuning of hyperparameters in a single and disciplined pipeline tailored to the structured monitoring of a hospital. This gap is particularly relevant in operational clinical environments, where reproducibility, fairness in benchmarking, and resistance to leakage are as important as nominal performance improvements.

From Early-Warning Scores to Data-Driven Inpatient Risk Modeling

Initial mainstream strategies of identifying inpatient deterioration were based on clinically interpretable scoring principles. An example of such work is the Modified Early Warning Score (MEWS), a combined risk score consisting of frequently measured physiological variables meant to be bedside usable in a concise form of result display and notification to the user Subbe et al. (2001a). These systems operationalized clinical expertise by assigning discrete points to thresholded vital-sign ranges and summing them into a single risk index. Their primary strengths lie in transparency, ease of computation, and immediate interpretability by frontline staff. Such systems have merits consisting of transparency and low computation cost. Their shortcomings, though, are structural. Discretization breaks down continuous physiological dynamics into crudely defined bins, interactions are largely ignored, and the logic associated with scoring is not learned but preprogrammed. Consequently, nuanced nonlinear relationships, patient-specific baselines, and evolving temporal patterns may not be adequately captured. These limitations grow depending on the scale of monitoring systems, and as decision support demanded by hospitals becomes more sensitive to nonlinearity, contextual variables, and time trends. In high-volume, digitally instrumented settings, the rigidity of threshold-based scoring can become a bottleneck to performance refinement.

The growth of EHR infrastructures and amalgamation of large clinical datasets fuelled the transition towards statistical and machine learning models. One of them is the MIMIC-III database that created an open benchmark resource for reproducible work in critical care modeling and facilitated the systematic comparison of learning pipelines on common data Johnson et al. (2016). By providing standardized access to richly annotated ICU data, such resources enabled the community to move from heuristic scoring toward data-driven risk estimation. In these infrastructures, risk modeling

was increasingly reconstructed as a supervised learning problem, where demographic, physiological, and laboratory data that tend to have irregular sampling and missingness need to be projected to clinical outcomes. This reformulation introduced new challenges, including the need to encode heterogeneous features, manage missingness, and design evaluation procedures that respect patient-level and temporal dependencies.

Another thread accentuated risk adjustment and inpatient mortality prediction in terms of routinely accessible administrative and physiologic indicators. Escobar et al. introduced a risk-adjusting methodology based on automated inpatient, outpatient, and laboratory data in which physiology-based variables significantly contributed to predictive discrimination Escobar et al. (2008). Although not designed strictly as a bedside early-warning score, this work demonstrated that structured physiologic signals, when integrated with broader contextual information such as comorbidities and admission characteristics, substantially enhance predictive validity. Even though this is not a bedside early-warning model strictly speaking, this work has methodological implications: it indicates the advantage of using structured information about physiology in concert with broader context (e.g., comorbidities, admission characteristics) in order to achieve better predictive validity and to isolate outcome modeling from post-admission confounding. The integration of multiple information sources foreshadowed later developments in multivariate hospital risk modeling.

Simultaneously, the use of deep learning methods to perform EHR prediction tasks, particularly at scale, gained popularity. Rajkomar et al. showed that models developed using deep learning can perform well in predicting various clinical tasks when trained on large EHR corpora Rajkomar et al. (2018a). Their work illustrated that flexible, high-capacity architectures can extract predictive structure from raw or minimally processed EHR data. This literature pointed to a major tension that remains in hospital settings: high-capacity models may be accurate, but they can be brittle without careful attention to representation, regularization, and evaluation design, especially when the data are structured rather than raw waveforms. In tabular monitoring data, where features are pre-aggregated and often semantically encoded, the advantage of deep architectures is not guaranteed unless pipeline design is meticulous.

It was also noted in the literature that inpatient physiological data usually have temporal structure even when represented in a tabular format. Shickel et al. conducted a review of deep learning techniques for analyzing EHR and identified recurrent problems including irregular sampling, missingness, and the need to represent clinical trajectories rather than single measurements Shickel et al. (2018). Their review emphasized that performance improvements frequently depend on how temporal information is encoded rather than on architecture alone. Similarly, Ghassemi et al. investigated severity-of-illness prediction and forecasting with sparse and heterogeneous ICU data through multivariate time-series modeling, demonstrating that explicit representation of multivariate temporal structure can improve acuity estimation Ghassemi et al. (2015). Together, these papers encourage two practical lessons: structured monitoring datasets require normalized heterogeneous features, encoded representations, and disciplined temporal alignment; and improvements often depend less on nominal model choice and more on refining control of the entire pipeline. Thus, temporal sensitivity and representation discipline emerge as foundational themes in modern inpatient predictive modeling.

Feature Selection in Heterogeneous Clinical Tabular Domains

The importance of structured hospital prediction is mainly centered on feature selection for both statistical and operational reasons. Statistically, correlated physiological variables, redundant derived timestamps, and categorical descriptors sharing clinical

meaning may inflate variance and destabilize optimization. Redundancy can obscure the marginal contribution of individual variables and amplify sensitivity to sampling variability. Operationally, a smaller set of features can simplify deployment, reduce integration burden, and improve interpretability, which is important when models inform clinical decisions. In hospital environments, where data pipelines interact with electronic systems and clinical workflows, parsimony can facilitate maintenance and auditability.

Guyon and Elisseeff provide a clear conceptual basis for feature selection in machine learning, categorizing methods into filter, wrapper, and embedded families and outlining trade-offs in computational cost and task-specific optimality Guyon and Elisseeff (2003b). Their framework clarifies that wrapper methods evaluate feature subsets directly with respect to a predictive model, whereas filter methods rely on generic criteria independent of the learner. Their study remains especially applicable to clinical modeling since wrapper methods, notwithstanding their increased cost, are explicitly geared toward optimizing predictive performance under the target learner. In heterogeneous hospital data, where interactions and conditional dependencies are common, wrapper approaches can better capture multivariate effects.

Swarm and evolutionary computation have become popular in biomedical applications where dimensionality is frequently high and relationships are nonlinear, making wrapper-based feature selection attractive. Saeys et al. surveyed feature selection methods in bioinformatics and highlighted the necessity of approaches resistant to redundancy, noise, and small-sample regimes Saeys et al. (2007). Their conclusions correspond to hospital monitoring data where measurement noise and correlated physiological variables are common. Díaz-Uriarte and Alvarez de Andrés demonstrated, in a high-dimensional biomedical context, that wrapper-based selection with strong learners such as random forests can yield robust predictive accuracy and mitigate overfitting Díaz-Uriarte and Alvarez de Andrés (2006). Though the domain of application varies, the methodological implication holds: careful choice of a compact subset may improve stability when the raw representation contains irrelevant or redundant dimensions. Stability and robustness become particularly critical when models are evaluated under strict cross-validation or temporal splits.

Survey work on evolutionary computation addressed feature selection at scale. Xue et al. reviewed evolutionary computation methods for feature selection and provided an organizing perspective on the impact of search operators and fitness design on convergence and subset quality Xue et al. (2016). Bolón-Canedo et al. provided a systematic review of feature selection on synthetic datasets aimed at exploring redundancy, noise, and interaction effects, emphasizing that performance claims must be interpreted relative to data-generating structure and evaluation protocols Bolón-Canedo et al. (2013). In clinical datasets, this is not an academic point: apparent improvements may disappear under strict leakage control or temporal splits. Therefore, feature selection must be embedded within rigorous validation to ensure that selected subsets generalize beyond a specific partition.

One major stimulus to the practical use of swarm-based feature selection was the development of binary counterparts of continuous optimizers. Emary et al. suggested binary grey wolf optimization (bGWO) variations as a feature selection method and empirically showed competitive performance in classification tasks Emary et al. (2016). These methods are appealing in hospital monitoring since they encode inclusion/exclusion decisions and can incorporate multi-term fitness functions that penalize subset size in addition to predictive error. The binary formulation aligns naturally with the discrete nature of feature selection, enabling direct control over representation sparsity.

Taken together, the feature selection literature encourages a methodological stance: in structured clinical prediction, feature selection is not simply a computational

convenience. It is often necessary to decrease variance, enhance optimization stability, and avoid spurious dependence on redundant covariates, particularly when subsequent hyperparameter optimization amplifies the statistical structure of the representation. Effective feature subset design can therefore serve as a foundational step in constructing reliable hospital monitoring pipelines.

Hyperparameter Optimization: From Random and Bayesian Search to Metaheuristics

The second lever is hyperparameter tuning, which strongly influences model generalization. Convergence stability in hospital monitoring pipelines can be affected by small changes in learning rate, regularization, or architectural width, shifting the bias–variance balance. In practice, hyperparameters mediate the trade-off between fitting complex patterns and maintaining generalization under noise and heterogeneity. Mainstream literature established that naive exhaustive strategies are inefficient in high-dimensional spaces. Bergstra and Bengio demonstrated that random search can outperform grid search since only a subset of hyperparameters typically has significant impact on performance Bergstra and Bengio (2012b). This finding remains practically significant for clinical datasets where evaluation cost is high and exhaustive enumeration yields diminishing returns.

Bayesian optimization offered a more sample-efficient approach by modeling the objective surface and selecting new trials via acquisition functions. Snoek et al. designed a practical Bayesian optimization methodology for machine learning algorithms and demonstrated improved efficiency compared with uninformed strategies Snoek et al. (2012). However, Bayesian procedures can struggle when dimensionality grows, when objective noise dominates, or when the search space contains mixed discrete–continuous variables, conditions that commonly arise when tuning structured learners on clinical tabular data with early stopping and cross-validation. In such settings, noise introduced by resampling and stochastic training can complicate surrogate modeling.

Metaheuristic optimization occupies a pragmatic middle ground. It tolerates nonconvexity and noisy objectives and naturally supports mixed-variable spaces. Particle swarm optimization (PSO) introduced a social–cognitive exploration and exploitation mechanism and became a canonical baseline for continuous hyperparameter tuning Kennedy and Eberhart (1995). Grey wolf optimizer (GWO) introduced a leadership-based search dynamic and has been widely adopted in engineering and machine learning tuning Mirjalili et al. (2014). The whale optimization algorithm (WOA) further diversified bio-inspired search rules by offering a trade-off between exploitation via encircling and exploration via spiral movements Mirjalili and Lewis (2016). These algorithms differ in practice because their update dynamics encode distinct inductive biases about contraction, diversification, and escape from local minima. Such diversity in search behavior motivates comparative benchmarking under controlled computational budgets.

Evolutionary strategies have been extensively applied in deep learning and neural architecture search to reduce the burden of manual configuration. Young et al. described evolutionary hyperparameter optimization for deep learning in high-performance computing contexts Young et al. (2015). Elsken et al. surveyed neural architecture search and conceptualized the broader space of search strategies, including evolutionary designs Elsken et al. (2019). In structured hospital monitoring data, where compute budgets are often limited and reproducibility is paramount, metaheuristics remain attractive because they can operate under strict budget constraints while maintaining robust exploration behavior. Their flexibility and tolerance to objective irregularities make them suitable for noisy clinical regression tasks.

Coupled Optimization of Representation and Learning Dynamics

One common drawback of applied clinical ML is that feature selection and hyperparameter choice are treated as independent and sequential problems. Such separation is convenient but generally suboptimal. Hyperparameter choices effective in a high-dimensional, noisy feature space may become ineffective after redundancy is removed; conversely, the significance of a feature may depend on model capacity and regularization. This coupling is increasingly recognized in biomedical applications where small-sample effects and heterogeneity are common. Integrated optimization strategies seek to mitigate the risk that gains achieved in one stage are neutralized in another.

Hybrid optimization models have therefore been proposed, typically combining a binary search step for feature determination with a continuous search step for learning parameter tuning. Sayed et al. introduced a hybrid wrapper–filter architecture using a binary grey wolf optimizer and demonstrated improved biomedical classification performance Sayed et al. (2020). Too et al. proposed a binary salp swarm algorithm for feature selection and showed that binary swarm methods can provide compact subsets with competitive predictive performance Too et al. (2019). Although these studies emphasize classification, their methodological implications extend to regression contexts found in vital-sign prediction: joint control of representation sparsity and learner configuration often determines whether gains persist under strict validation. By coordinating discrete and continuous search processes, hybrid designs aim to stabilize performance across varying data regimes.

The practical implication is that a unified optimization framework capable of operating coherently in both binary and continuous spaces provides a principled means to address the coupled nature of the problem without violating fairness constraints. Such a framework is particularly valuable in hospital monitoring settings when combined with reproducible evaluation designs (fixed splits, controlled seeds, repeated runs), since clinical deployment requires stability rather than one-time best-case performance. Robustness, comparability, and transparency become decisive evaluation criteria.

Synthesis and Implications for Optimization-Aware Hospital Monitoring Pipelines

Three consolidated claims are supported by the reviewed literature. First, structured inpatient monitoring data are predictive yet methodologically challenging: heterogeneity of feature types, scale variations, and latent temporal dependencies require careful preprocessing and representation control Johnson et al. (2016); Shickel et al. (2018); Ghassemi et al. (2015). Second, feature selection is often decisive in clinical tabular tasks due to variance reduction, improved stability, and alignment with operational constraints Guyon and Elisseeff (2003b); Saeys et al. (2007); Xue et al. (2016); Emary et al. (2016). Third, hyperparameter optimization significantly affects generalization, and derivative-free metaheuristics are attractive in noisy and mixed-variable regimes under limited computational resources Bergstra and Bengio (2012b); Snoek et al. (2012); Kennedy and Eberhart (1995); Mirjalili et al. (2014); Mirjalili and Lewis (2016). These themes collectively suggest that predictive success in hospital monitoring contexts depends less on isolated algorithmic novelty and more on disciplined orchestration of representation and optimization.

What remains comparatively scarce are studies that: (i) enforce strict computational parity among multiple optimizers, (ii) evaluate both discrete (feature selection) and continuous (hyperparameter) optimization within a shared methodological infrastructure, and (iii) quantify improvements using a broad suite of regression metrics

that capture magnitude error, correlation structure, and agreement. Addressing this gap is less about inventing new algorithms and more about executing optimization-aware modeling with methodological discipline. In structured inpatient monitoring, such discipline is essential for translating empirical gains into robust and reproducible clinical insight.

3 Materials and Methods

Overview of the proposed method. Figure 1 is an overview of the proposed unified optimization pipeline for structured hospital vital-sign prediction. The framework is designed as a sequential yet internally coupled process in which representation learning and optimization are treated as coordinated components rather than isolated steps. The workflow begins with data preprocessing (cleaning, encoding, scaling, and dataset partitioning), ensuring that heterogeneous clinical variables are transformed into a numerically stable and semantically consistent feature matrix. This stage establishes the structural integrity of the modeling space and prevents downstream optimization from amplifying artifacts or leakage.

Following preprocessing, the Al-Biruni Earth Radius (BER) optimizer operates in a binary search space to perform wrapper feature selection (bBER). In this stage, each candidate solution corresponds to a binary vector representing inclusion or exclusion decisions over the available attributes. The fitness of each solution is evaluated using the downstream predictive performance of the learning model under cross-validation, thereby directly aligning subset quality with regression accuracy. The output of this stage is a compact, performance-oriented subset of informative variables that balances predictive capacity and dimensionality control.

The Feature-Transformed Learning Model (FTLM) is then trained and validated using K-fold cross-validation with the selected features. This intermediate training stage serves two purposes: it provides an initial estimate of generalization performance using the reduced representation, and it defines the objective function for the subsequent hyperparameter optimization phase. Afterward, BER is utilized again, this time in a continuous bounded space, to optimize the FTLM hyperparameters. In this second stage, each candidate solution represents a real-valued configuration of learning parameters constrained within predefined bounds. The optimizer explores and exploits this space using the same core search logic as in the binary stage, but adapted to continuous dynamics.

The resulting configuration—comprising both the selected feature subset and the optimized hyperparameters—is finally evaluated on the held-out test set using regression performance metrics. The inset in Figure 1 highlights the shared BER search loop (initialization, exploration, exploitation, diversity/mutation, and best-solution update) underlying both the discrete and continuous optimization stages. This unified design ensures methodological coherence and computational parity across stages, while maintaining strict separation between training/validation and final testing.

3.1 Dataset Description

The empirical analysis conducted in this study is based on the *Patient Vital Signs and Event Tracking* dataset publicly available through Kaggle. The dataset is structured in tabular form and contains seventeen attributes representing clinically or operationally meaningful variables recorded during hospital admission. The data are organized around a unique hospital admission identifier (`hadm_id`), which enables the association of multiple physiological observations and clinical events with a single episode of

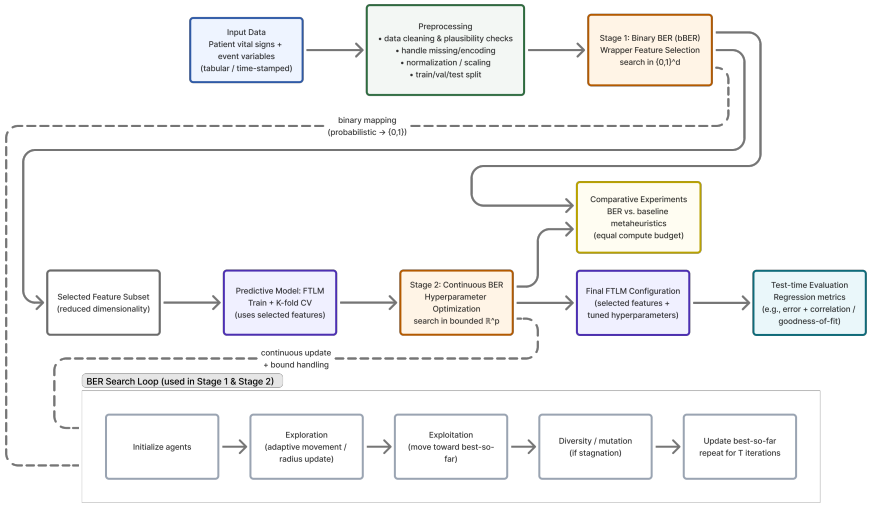

Fig. 1 Proposed unified BER–FTLM framework for structured hospital vital-sign prediction. The pipeline consists of: (i) preprocessing and data partitioning; (ii) Stage 1 binary BER (bBER) wrapper feature selection to obtain a reduced feature subset; (iii) FTLM training with K-fold validation; (iv) Stage 2 continuous BER hyperparameter optimization over a bounded real-valued search space; and (v) final test-set evaluation with regression metrics. The inset summarizes the shared BER optimizer loop used in both stages (initialization, exploration, exploitation, diversity/mutation, and best-solution update).

care. This admission-centric organization reflects real-world hospital data structures, where repeated measurements and events are temporally nested within a defined hospitalization window.

The dataset integrates physiological, demographic, clinical state, and temporal variables into a unified representation. Physiological measurements include heart rate (HR), respiratory rate (RR), systolic blood pressure (SBP), body temperature (TEMP), and peripheral oxygen saturation (SPO_2). These variables constitute the core vital-sign indicators routinely monitored in inpatient settings and collectively represent cardiovascular, respiratory, and thermoregulatory functions. HR and RR are expressed in beats per minute and breaths per minute, respectively; SBP is measured in millimeters of mercury (mmHg); TEMP reflects body temperature in standardized clinical units; and SPO_2 represents the percentage of hemoglobin saturated with oxygen. Variations across these signals encode dynamic physiological responses to illness, treatment, and environmental factors. The joint modeling of these measurements allows construction of a multidimensional physiological state space for each observation.

Demographic attributes consist of patient age and sex. Age is recorded as a continuous variable reflecting age in years at the time of admission, while sex is encoded as a categorical variable indicating male or female. Although limited in number, these attributes provide essential contextual modifiers. Age influences baseline physiological ranges, comorbidity burden, and risk stratification, whereas sex may be associated with differential physiological baselines and response patterns. Inclusion of these variables supports adjustment for demographic heterogeneity during modeling.

Clinical state variables include `avpu`, `masktype`, `prev_event`, and `next_event`. The `avpu` variable encodes the patient's level of consciousness according to the Alert–Voice–Pain–Unresponsive scale, capturing neurological responsiveness in an ordinal

format. The `masktype` variable indicates the type of respiratory support device in use (e.g., nasal cannula or face mask), reflecting the intensity of respiratory assistance. The variables `prev_event` and `next_event` capture sequential clinical events or interventions associated with hospitalization, implicitly encoding transitions in care status and treatment phases. These categorical descriptors introduce discrete state transitions that complement continuous vital signs and allow the model to capture interactions between physiologic state and care context.

Temporal variables include `admittime`, `dischtime`, `eventtime`, `charttime`, and `hrs_to_firstevent`. The `admittime` and `dischtime` variables define the temporal boundaries of each hospital admission episode. The `eventtime` variable records the timestamp of a specific clinical event, whereas `charttime` represents the time at which a medical observation or documentation entry was recorded. The derived variable `hrs_to_firstevent` quantifies the duration between admission and the first recorded event, providing a continuous measure of early-event timing. These temporal fields allow reconstruction of within-admission trajectories and enable derivation of interval-based features that reflect progression dynamics.

The dataset spans the period from 2017 to 2019, covering multiple calendar years of hospital activity. This multi-year coverage introduces natural variability related to seasonal patterns, operational practices, and patient population shifts. While this enhances representativeness and ecological validity, it also necessitates careful partitioning strategies to prevent inadvertent temporal leakage between training and evaluation subsets.

Distributional inspection of the primary physiological variables reveals bounded and clinically plausible ranges. HR values predominantly fall within typical adult physiological intervals, with tails representing bradycardic and tachycardic states. RR exhibits dispersion corresponding to both normal ventilation and increased respiratory effort. SBP spans hypotensive and hypertensive regimes, indicating heterogeneity in cardiovascular status. TEMP measurements cluster around normothermic values with dispersion toward febrile and hypothermic ranges. SPO_2 values are concentrated within high-percentage intervals consistent with adequate oxygenation, while lower values indicate compromised respiratory function. These patterns confirm physiological coherence while preserving variability necessary for predictive modeling.

Overall, the *Patient Vital Signs and Event Tracking* dataset provides a heterogeneous yet structured representation of hospital admissions, integrating continuous physiological measurements, categorical clinical states, demographic attributes, and temporally indexed events within a single tabular framework. This integrated structure supports both feature selection and hyperparameter optimization under a unified modeling paradigm.

3.2 Data Preprocessing

Before feature selection and model optimization, preprocessing was conducted to ensure numerical stability, semantic consistency, and methodological reproducibility. Given the heterogeneous structure of the dataset, preprocessing procedures were implemented sequentially and in a controlled manner to avoid information leakage while preserving clinically meaningful relationships among variables. Each step was executed using training-data-derived parameters and subsequently applied to validation and testing subsets.

Numerical Consistency and Plausibility Validation.

Systematic numerical consistency checks were performed on all continuous physiological variables, including HR, RR, SBP, TEMP, and SPO_2. These checks verified the absence of invalid numeric encodings, missing placeholders embedded as strings, and non-physical values. Physiological plausibility validation was conducted by comparing observed ranges against established clinical limits. Values clearly outside medically reasonable boundaries were flagged and manually reviewed to determine whether they reflected measurement artifacts or extreme but possible clinical states. Rather than applying indiscriminate truncation, a conservative retention strategy was adopted to preserve clinically meaningful extremes. This ensures that rare but informative cases are retained while preventing model distortion due to erroneous entries.

Categorical Variable Encoding.

Categorical variables, including `sex`, `avpu`, `masktype`, `prev_event`, and `next_event`, were transformed into numerical representations suitable for supervised learning. Nominal variables without inherent ordering were encoded in a manner that prevents artificial ordinal relationships. For ordinal variables with clinically meaningful structure, such as the AVPU scale, encoding preserved the intrinsic order to reflect graded neurological states. All encoding mappings were derived from the training subset and consistently applied to validation and testing partitions to ensure alignment and prevent category mismatch.

Temporal Variable Handling and Derived Intervals.

Timestamp variables (`admittime`, `dischtime`, `eventtime`, and `charttime`) were converted into standardized datetime objects to allow arithmetic operations. Derived temporal intervals were computed where clinically relevant, including total admission duration and time differences relative to key events. The variable `hrs_to_firstevent` was cross-validated against admission timestamps for internal consistency. Temporal transformations were performed strictly within each `hadm_id` to prevent cross-admission contamination. All derived features were constructed using only information available up to the defined prediction reference, thereby avoiding look-ahead bias.

Z-Score Normalization of Continuous Features.

To address scale heterogeneity across physiological measurements, Z-score normalization was applied to all continuous input variables. For each feature x, the standardized value z was computed as

$$z = \frac{x - \mu}{\sigma}, \tag{1}$$

where μ and σ denote the mean and standard deviation calculated exclusively on the training subset. This standardization ensures balanced contribution of variables during optimization and prevents dominance of high-magnitude features. The learned normalization parameters were then applied unchanged to validation and testing data, preserving strict separation between training and evaluation phases.

Outlier Inspection Within Physiological Limits.

Outlier detection was performed using both statistical dispersion indicators and domain-informed plausibility thresholds. Observations located in extreme distribution tails were evaluated for measurement realism rather than automatically removed. This approach acknowledges that extreme physiological values may correspond to clinically

significant events. By integrating domain knowledge with statistical inspection, the preprocessing pipeline mitigates noise while preserving predictive signal.

LLM-Assisted Methodological Support.

High-level methodological guidance during preprocessing was informed by consultation with the open-weight large language model DeepSeek-R1. The model was used exclusively to suggest general diagnostic considerations, validation checkpoints, and preprocessing alternatives at a conceptual level. No patient-level data, intermediate results, or quantitative outputs were shared with the model, and no automated transformation decisions were delegated to it. All preprocessing operations were implemented deterministically by the authors. This controlled use ensured methodological transparency while retaining full analytical responsibility.

Collectively, these preprocessing procedures established a validated, leakage-aware, and numerically stable feature matrix suitable for the subsequent feature selection and hyperparameter optimization stages within the unified BER–FTLM framework.

3.3 Exploratory Data Analysis

A comprehensive exploratory data analysis (EDA) was conducted to characterize distributional properties, inter-variable relationships, event-conditioned differences, and potential nonlinear structures within the *Patient Vital Signs and Event Tracking* dataset. The analysis integrates univariate density inspection, temporal aggregation, subgroup comparison, multivariate visualization, and density-based relational mapping. This step serves two purposes: (i) to verify statistical plausibility prior to optimization stages and (ii) to identify structural patterns that may influence feature selection and hyperparameter search landscapes. In practical terms, EDA functions as a diagnostic checkpoint: it reveals whether preprocessing has produced stable feature behavior, whether any variable exhibits pathological distributions likely to destabilize training, and whether separability appears to rely on subtle multivariate interactions rather than strong univariate thresholds. Because the subsequent stages include wrapper-based selection and repeated model fitting during hyperparameter search, identifying potentially problematic distributions and weakly separable marginal patterns is essential to avoid misinterpreting optimization outcomes.

The histogram and kernel density estimate of heart rate (HR) are presented in Figure 2. The distribution exhibits an approximately unimodal and symmetric structure centered near 60 beats per minute, with tails extending toward lower (45 bpm) and higher (75 bpm) values. Mild dispersion is observed without extreme skewness, suggesting suitability for Z-score normalization. The relatively compact variance indicates stable physiological recording conditions across admissions. From a modeling perspective, the near-symmetric shape implies that linear scaling will not excessively compress one tail or amplify the other, and that optimization procedures relying on gradient-like updates in the learner are less likely to be driven by heavy-tailed HR extremes. Clinically, the observed range also appears coherent with routine inpatient observations in relatively stable populations, while still including sufficient variation to represent physiologic stress.

Temporal aggregation of clinical events by hour is illustrated in Figure 3. The distribution shows moderate variability across the 24-hour cycle, with observable fluctuations during late afternoon and evening hours. No extreme concentration in a single hour is detected, indicating temporal coverage without severe recording bias. This uniformity reduces the likelihood of time-induced confounding during modeling. More broadly, the pattern suggests that the dataset does not simply reflect a narrow

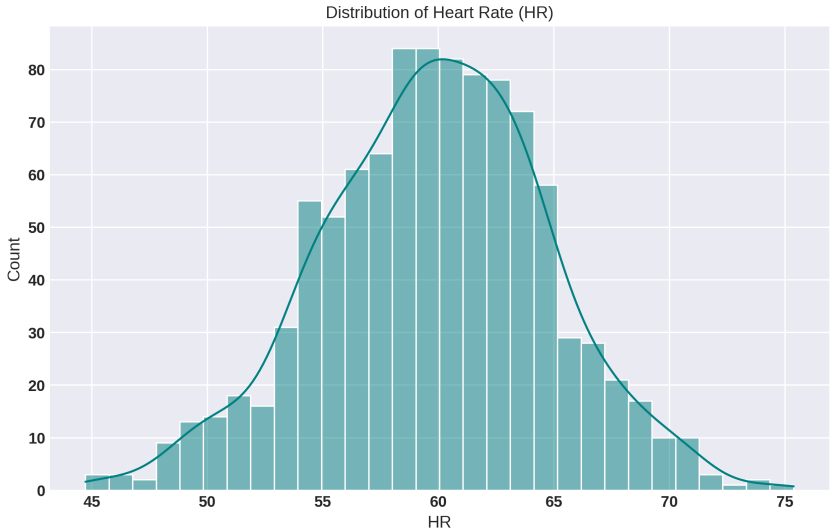

Fig. 2 Distribution of heart rate (HR) with kernel density estimation.

operational window (e.g., only morning rounds or only daytime documentation), which would otherwise risk coupling event occurrence with staffing schedules or routine workflows. While moderate diurnal variation is expected in hospitals due to shift changes and scheduled processes, the absence of severe spikes supports the view that events are captured across the day in a manner compatible with trajectory-based modeling.

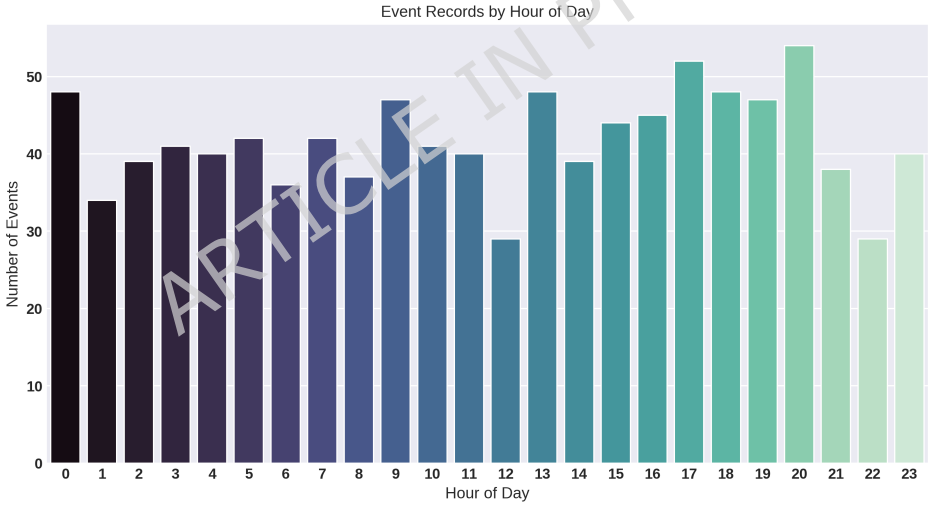

Fig. 3 Number of recorded events by hour of day.

Figure 4 presents a swarm visualization of HR stratified by sex and next event status. The distributions for male and female patients display overlapping ranges with comparable central tendencies. Event-conditioned color encoding reveals substantial overlap between event categories (0 and 1), suggesting that HR alone may not be a strong univariate discriminator. However, subtle dispersion differences justify multivariate evaluation. In particular, even when medians appear similar, differences in density around the central region or in tail thickness can contribute to discrimination when combined with other physiological variables or contextual covariates. This observation

is consistent with clinical intuition: HR is rarely interpreted as a stand-alone predictor of deterioration, but becomes informative when contextualized by respiratory status, blood pressure, oxygenation, and neurological responsiveness.

Fig. 4 Swarm plot of heart rate by sex and next event classification.

Figure 5 presents kernel density estimates for HR, respiratory rate (RR), systolic blood pressure (SBP), temperature (TEMP), oxygen saturation (SPO2), age, hours to first event, and previous event status, stratified by next event outcome. Most physiological variables exhibit high overlap between classes, indicating moderate class separability in marginal space. This degree of overlap implies that any decision boundary relying on a single variable would either be insensitive (missing many event-associated cases) or overly sensitive (producing many false alerts), which motivates multivariate learning and interaction-aware modeling. Age demonstrates multimodal structure, while hours to first event shows asymmetric dispersion. The multimodality in age is consistent with heterogeneous admission cohorts (e.g., distinct patient age groups) rather than a single homogeneous population, and such heterogeneity can induce interaction effects whereby physiological values may carry different meaning depending on age strata. The asymmetric dispersion in hours to first event suggests that early-event timing is not normally distributed and may encode workflow- or acuity-related effects that are concentrated near admission for some admissions and delayed for others. These observations highlight the necessity of multivariate modeling rather than reliance on single-variable thresholds.

Robust distribution summaries using boxen plots are shown in Figure 6. Median shifts between event classes remain subtle across HR, RR, SBP, TEMP, SPO2, and age. However, distributional spread differs in certain variables, particularly SBP and age, where extended tails suggest heterogeneity among patients with different event outcomes. This type of spread difference is important: even when central tendencies match, variability changes may reflect instability or broader physiologic dispersion in one subgroup, which can be predictive when integrated with other signals. The boxen representation also highlights how differences may be more prominent in the upper or lower quantiles than at the median, implying that outcome associations might be expressed through tail behavior (e.g., hypotensive or hypertensive extremes) rather than through average shifts. The absence of extreme outlier clustering confirms that prior plausibility filtering was effective. This is relevant for subsequent optimization, since extreme erroneous values can dominate loss functions, bias feature selection toward artifact-driven predictors, and destabilize hyperparameter search by making objective evaluations noisier.

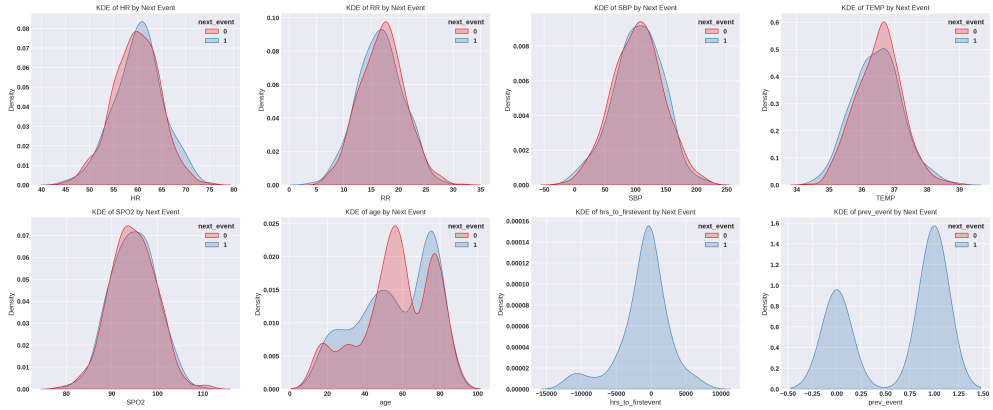

Fig. 5 Kernel density estimates of key variables stratified by next event outcome.

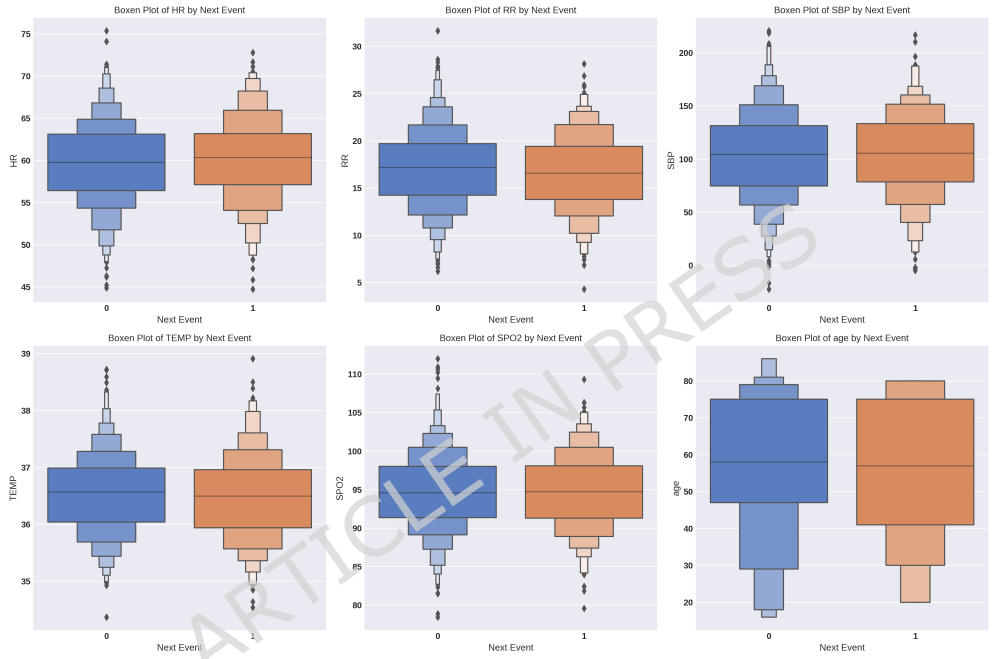

Fig. 6 Boxen plots of physiological variables by next event category.

Figure 7 displays pairwise scatter relationships with KDE diagonals stratified by next event. Linear associations appear weak to moderate among certain pairs (e.g., HR and RR), while SBP exhibits broad dispersion relative to other variables. The observed HR–RR association is clinically plausible, as tachycardia and tachypnea can co-occur during physiological stress; however, the relationship is not tight enough to indicate redundancy, suggesting that both variables may still contribute complementary information. The broad SBP dispersion indicates that blood pressure spans a wide range across observations and may interact with other variables (e.g., HR compensation patterns) in nonlinear ways. Overlapping class distributions confirm that discriminative boundaries are unlikely to be linearly separable in low-dimensional projections. This reinforces the need for nonlinear learning architectures and optimized hyperparameters. In optimization terms, if separability is weak in simple projections, improvements are

more likely to arise from better representation and tuning rather than from any single feature exhibiting strong marginal effect.

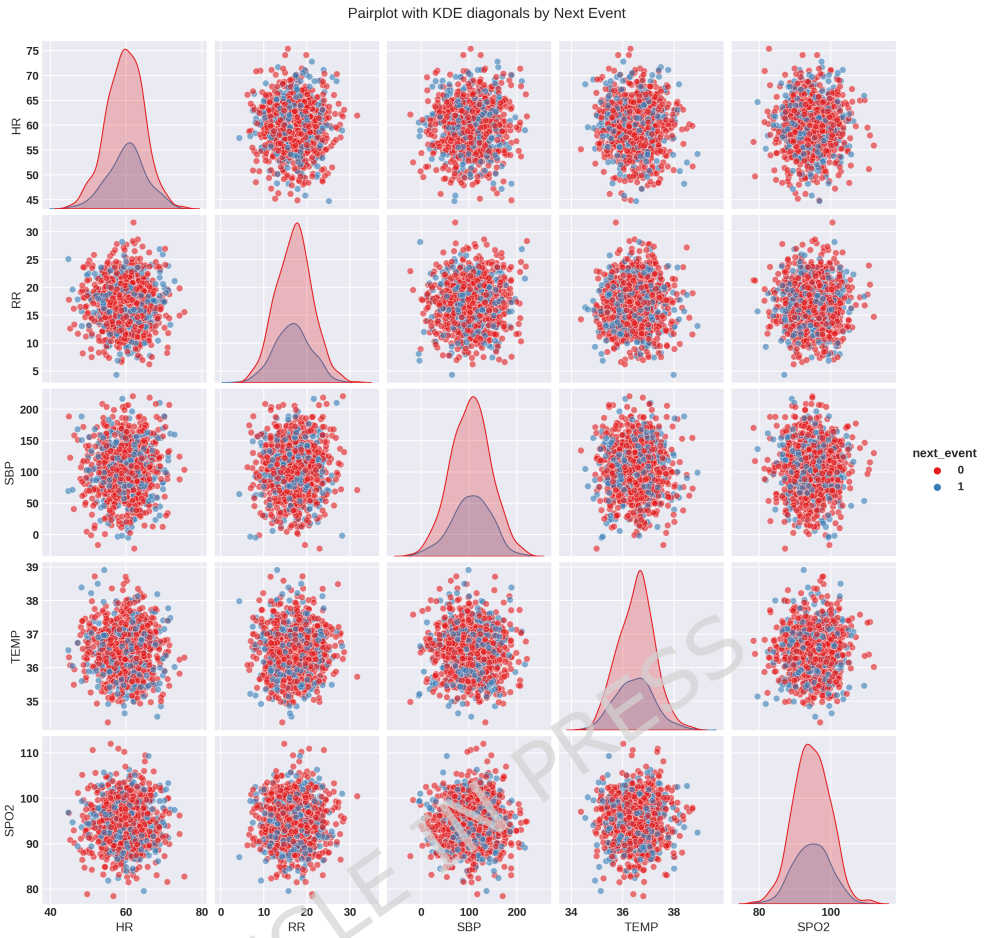

Fig. 7 Pairplot with KDE diagonals illustrating pairwise relationships by next event.

The hexbin density plot in Figure 8 illustrates the joint distribution of HR and SPO2. The highest density region is centered near HR 60 bpm and SPO2 95%, with no strong nonlinear curvature evident. The absence of pronounced negative or positive monotonic trends suggests limited direct dependence between these two variables. This observation supports the inclusion of interaction-aware modeling rather than simple bivariate regression assumptions. Importantly, the lack of strong bivariate structure does not imply irrelevance; instead, it suggests that any predictive utility of HR and SPO2 may be conditional on additional variables such as respiratory support type, RR, or event timing. Thus, the hexbin plot provides empirical justification for modeling approaches capable of capturing higher-order interactions and context dependence.

Figure 9 provides a refined multivariate visualization using KDE on the diagonal and scatter plots off-diagonal. Compared to the previous pairplot, this representation emphasizes density overlap between event categories. The substantial overlap confirms that class discrimination requires integrated feature combinations rather than isolated thresholds. At the same time, the plots suggest that the joint feature space is structured rather than random: clusters and dense regions exist, but event categories occupy overlapping portions of these regions. This configuration is typical of clinical monitoring

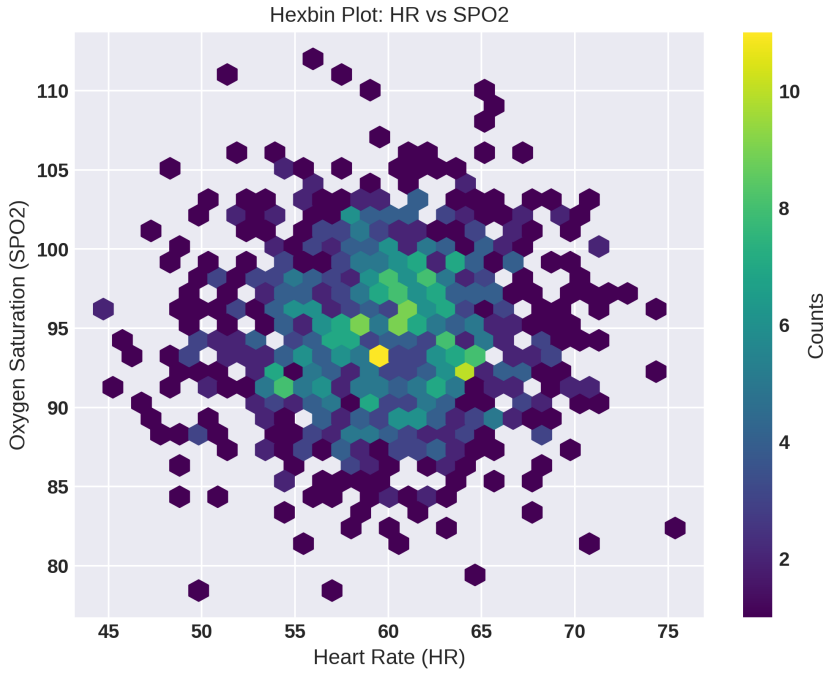

Fig. 8 Hexbin density plot of heart rate versus oxygen saturation.

tasks, where outcomes are influenced by multifactorial processes and where discriminative signal is distributed across correlated physiological and contextual variables. The structured yet overlapping clusters justify the subsequent feature selection stage to identify informative variable subsets. Specifically, feature selection can help isolate variables that contribute incremental predictive information within these overlapped regions, thereby improving generalization and reducing the burden on the learner.

The exploratory analyses indicate that most physiological variables exhibit approximately symmetric distributions, moderate inter-variable correlation, and significant class overlap in marginal projections. No critical outliers or pathological distributions were detected. From an optimization perspective, this combination implies that improvements are unlikely to be obtained through simplistic thresholding or linear separation, but rather through careful selection of complementary features and appropriate tuning of model capacity and regularization. These properties imply that predictive separation requires multivariate modeling supported by systematic feature selection and hyperparameter optimization, rather than reliance on simple threshold-based rules. The insights obtained from EDA informed both the binary encoding strategy for feature selection and the continuous search bounds for hyperparameter optimization described in subsequent sections.

3.4 Baseline Learning Models

To establish a controlled and interpretable reference point prior to any optimization procedure, a diverse suite of baseline learning models was evaluated on the fully preprocessed *Patient Vital Signs and Event Tracking* dataset. The intent of this stage was not to exhaustively tune architectures, but to quantify how different inductive biases behave under an identical preprocessing pipeline, identical data partitions, and a unified evaluation protocol. This baseline benchmarking serves as a methodological "floor" against which any subsequent improvement can be meaningfully attributed to the proposed optimization stages rather than to uncontrolled variation in data

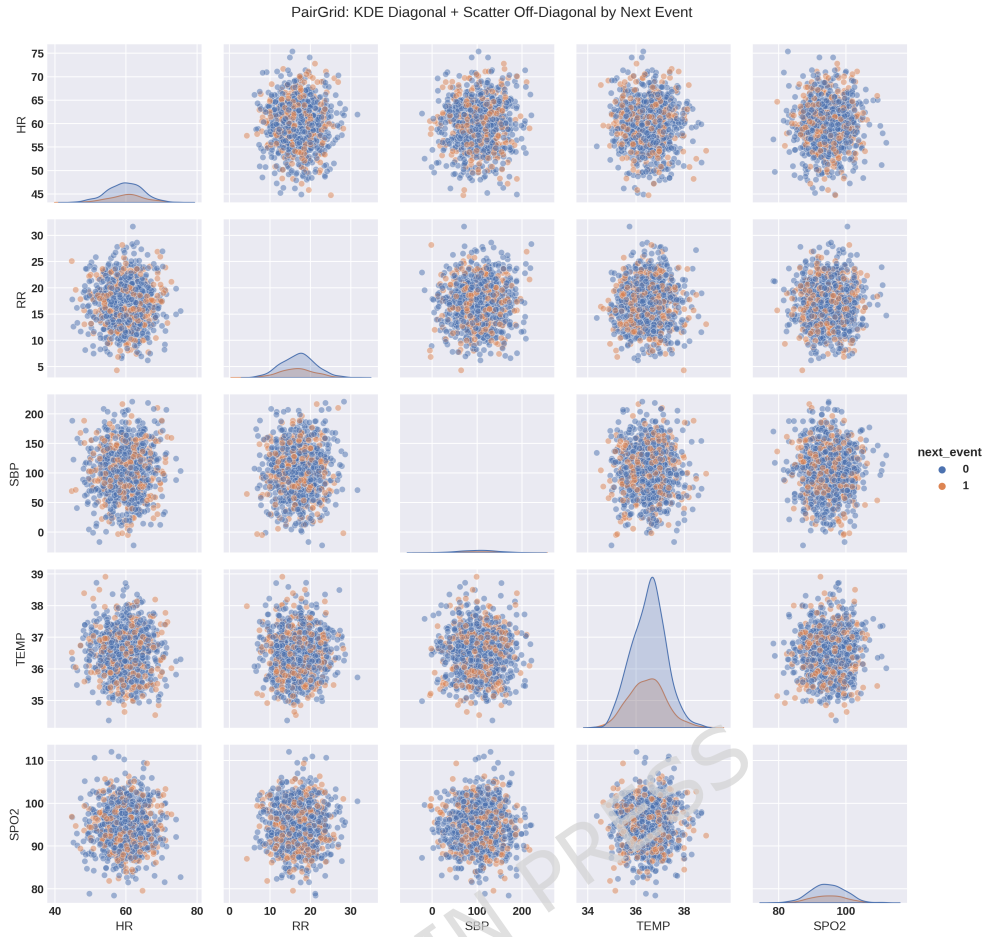

Fig. 9 PairGrid visualization with KDE diagonals and scatter off-diagonals stratified by next event.

handling or opportunistic tuning. In structured clinical prediction, such baselines are especially important because model performance can be artificially inflated by subtle design choices (e.g., encoding conventions, leakage through temporal variables, or accidental partition overlap). By holding these factors constant, baseline evaluation isolates the role of the learning algorithm itself.

This is important in structured hospital monitoring data because the feature space is inherently heterogeneous: continuous physiological streams (e.g., HR, RR, SBP, TEMP, SPO_2), categorical clinical descriptors (e.g., AVPU, mask type, event labels), and time-derived variables encode distinct statistical regularities and interact nonlinearly. Continuous measurements reflect bounded, noisy biological processes with differing measurement scales; categorical descriptors encode discrete clinical state, device usage, or care transitions; and time-based features implicitly embed progression, recency, and workflow patterns. These feature families do not merely contribute additively: their predictive meaning is often conditional (e.g., oxygen saturation depends on respiratory support type; the implication of a given heart rate depends on blood pressure and consciousness status). A model that learns cross-feature interactions well in tabular form may not be the one that best leverages within-admission temporal structure, and a representation-learning model can appear stable in correlation while still drifting in bias. Moreover, the same model may exhibit different failure modes across patient strata or event contexts, underscoring the need for multiple baselines to identify whether

errors arise from inadequate interaction modeling, inadequate temporal modeling, or unstable representation learning.

By evaluating a carefully selected range of tabular, temporal, and latent-variable baselines, we obtain a calibrated empirical reference that isolates the incremental contribution of feature selection and hyperparameter optimization in later stages. Specifically, the baseline stage answers the question: given the same preprocessed inputs and the same evaluation protocol, what performance level is achievable by representative modeling paradigms without specialized optimization? This is critical because the proposed pipeline aims to improve performance through (i) representation refinement via feature subset selection and (ii) training-dynamics refinement via hyperparameter optimization; both of these improvements should be interpreted relative to a stable and transparent baseline anchor.

FTLM.

FTLM serves as the primary modeling framework in this study and is treated as a structured tabular learner designed to capture nonlinear interactions among heterogeneous variables while retaining computational tractability. Architecturally, FTLM is implemented in the spirit of modern deep tabular baselines that adapt Transformer-style attention to feature-tokenized inputs, explicitly modeling cross-feature dependencies rather than relying on handcrafted interactions Gorishniy et al. (2021). The attention-based formulation is particularly appropriate for mixed-type clinical tabular data because it enables the model to learn conditional dependencies among variables (e.g., how oxygenation patterns interact with respiratory rate and support device type) without requiring manual construction of interaction terms. In clinical tabular settings, interaction effects between physiology and context are rarely additive, and a feature-interaction learner provides a reasonable default hypothesis class before any optimization is introduced.

In the baseline configuration, FTLM is trained using the default hyperparameter values specified in the experimental setup, producing a stable anchor for interpreting improvements obtained after feature selection and hyperparameter tuning. This choice is deliberate: by avoiding aggressive baseline tuning, the framework ensures that any subsequent improvements can be traced to the proposed optimization pipeline rather than to extensive manual calibration. Consequently, baseline FTLM performance functions as the principal reference point for both the feature-selection stage and the hyperparameter-optimization stage in later sections.

CTSM.

CTSM is included as a competitive structured model that emphasizes contextualized representations for tabular attributes. Operationally, CTSM follows the principle that categorical and encoded clinical descriptors should not be embedded independently; rather, their representation should be conditioned on surrounding feature context to reflect clinically meaningful dependence (e.g., the predictive meaning of a consciousness scale depends on concurrent vitals). This design intuition aligns naturally with inpatient monitoring, where discrete descriptors (such as AVPU and mask type) act as context signals that modulate the interpretation of continuous measurements rather than functioning as independent predictors.

This aligns with contextual embedding strategies in tabular Transformers, where self-attention transforms categorical embeddings into context-aware representations that can be more robust to noise and missingness Huang et al. (2020). Contextualization is especially relevant in clinical datasets because categorical fields may be sparsely

represented, inconsistently documented, or strongly confounded by care pathway. By conditioning embeddings on other features, CTSM aims to mitigate brittleness arising from isolated categorical encodings. As with FTLM, CTSM is trained under identical preprocessing outputs and partitioning rules to ensure that the comparison reflects modeling bias rather than differences in data handling. Thus, CTSM provides a baseline for the hypothesis that contextual tabular embedding improves robustness in heterogeneous monitoring data.

VAST.

VAST is evaluated as a tabular representation model that emphasizes attention-based feature interaction while additionally encouraging representation quality through instance-level structure during training. Concretely, the baseline VAST implementation follows the general design logic of attention-driven tabular learners that incorporate row/instance attention and contrastive-style regularization to stabilize learning on heterogeneous tabular distributions Somepalli et al. (2021). This class of models is motivated by the observation that structured hospital datasets can contain recurring physiological "motifs" across admissions—for example, patterns reflecting stable states, acute deterioration, or recovery trajectories—and that explicitly shaping instance representations can improve generalization by discouraging over-reliance on idiosyncratic covariates.

This model family is included because structured hospital datasets can exhibit recurring physiological motifs across admissions, and instance-aware representation learning can reduce sensitivity to redundant attributes or spurious correlations that arise in observational data. In hospital records, correlations can emerge due to documentation practice, device usage, or care routines, which may not be stable across wards or time. Regularization mechanisms that emphasize representation consistency can therefore be beneficial. The baseline VAST configuration is intentionally not aggressively tuned at this stage so that subsequent improvements can be attributed to the proposed optimization pipeline rather than to extensive manual architectural refinement. In this role, VAST serves as a baseline for the hypothesis that instance-level representation shaping provides robustness benefits beyond standard attention-based tabular modeling.

LSTM.

Long Short-Term Memory (LSTM) networks are canonical recurrent architectures designed to model sequential dependencies through gated memory mechanisms that mitigate vanishing gradients and enable learning over long contexts Hochreiter and Schmidhuber (1997). Although the dataset is organized as structured records, the presence of admission-anchored timestamps and event-related fields introduces temporal structure that can be exploited when observations are arranged into within-admission sequences. In hospital monitoring, temporal order often encodes causality-relevant information: deterioration is typically reflected in trends, accelerations, or volatility rather than in isolated values.

The LSTM baseline is therefore included to test whether recurrent dynamics provide an advantage in capturing temporal correlations implicitly encoded by admission trajectories and event timing. Inputs are formatted to preserve within-admission ordering while maintaining the leakage-prevention constraints described in preprocessing and validation. This formulation is intended to reflect a realistic scenario in which the learner observes a sequence of temporally ordered measurements and produces predictions conditioned on recent and longer-range context. As a baseline, the LSTM serves as a comparator for whether the task benefits meaningfully from explicit temporal modeling rather than solely from feature interaction modeling in a static tabular view.

DTCN.

DTCN represents a deep temporal convolutional baseline intended to extract hierarchical time-dependent patterns using convolutional operators rather than recurrence. Temporal convolutional designs are attractive because they offer stable optimization, parallelizable computation, and multiscale receptive fields that capture both short- and longer-range dependencies without explicit recurrent state Bai et al. (2018). By stacking convolutional layers, temporal convolutions can summarize local dynamics (short windows) and progressively integrate longer context, often with more stable training characteristics than recurrent networks.

In this study, the DTCN baseline is instantiated as a deep temporal convolution network in the sense that temporal context is propagated across layers to support discrimination at deeper representations, consistent with prior formulations that explicitly refer to this family as a DTCN Koh et al. (2021). Its inclusion enables a direct comparison between recurrent temporal modeling (LSTM) and convolutional temporal modeling (DTCN) under identical preprocessing constraints and evaluation metrics. This comparison is practically relevant because, in real deployments, convolutional temporal models may provide computational advantages due to parallelism, but their effectiveness depends on whether the relevant predictive information is sufficiently local or multiscale in a way that convolutional receptive fields can capture.

TST.

TST is evaluated as a Transformer-based temporal baseline that leverages attention mechanisms to model dependencies across time steps without strict recurrence. The architectural rationale traces to the Transformer formulation, which replaces recurrence with attention-based dependency modeling and has become a standard reference point for sequence learning Vaswani et al. (2017). Attention mechanisms can, in principle, connect distant time steps directly, enabling the model to capture long-range dependencies, irregular temporal relevance, and nonlocal interactions across the sequence.

For time-series and multivariate temporal data, Transformer encoders have also been shown to provide effective representations for downstream regression and classification when temporal structure is meaningfully presented to the model Zerveas et al. (2021). In admission-structured hospital data, this can be advantageous when clinically meaningful changes occur over variable time horizons, and when the importance of a past observation is not strictly a function of its distance in time. In the present framework, TST is included to test whether attention-driven temporal dependency modeling provides advantages over recurrent and convolutional temporal baselines when applied to admission-structured hospital monitoring data. As a baseline, TST helps determine whether the additional flexibility of attention across time yields improvements under the same preprocessing and evaluation safeguards.

VAE.

The Variational Autoencoder (VAE) is included as a representation-learning baseline grounded in variational inference, where an encoder maps observations to a latent distribution and a decoder reconstructs inputs from sampled latent codes Kingma and Welling (2014). Although VAEs are typically introduced as generative models, their latent embeddings can be leveraged for prediction as a compact representation of heterogeneous inputs. This is particularly relevant in structured hospital data, where the raw feature matrix can contain redundancy, noise, and mixed-scale variables;

compressing the data into a latent space can provide a regularized representation that emphasizes dominant factors of variation.

Latent compression can function as an implicit regularizer by dampening the effect of redundant variables, reducing sensitivity to measurement noise, and stabilizing downstream regression when raw feature scales and distributions differ substantially. In this study, the VAE is trained to encode the preprocessed feature matrix into a lower-dimensional latent space, and the resulting embeddings are used for supervised regression under the same evaluation protocol as all other baselines. The VAE baseline therefore probes whether unsupervised representation learning, followed by supervised prediction, provides a competitive alternative to directly supervised tabular or temporal learners in this setting.

Evaluation Protocol for Baselines.

All baseline models were trained and evaluated using identical preprocessing outputs, data partitions, and performance metrics. This design ensures that differences in results reflect learning biases rather than preprocessing artifacts or partition inconsistencies. The evaluation criteria include mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), mean bias error (MBE), correlation coefficient (r), coefficient of determination (R^2), relative RMSE (RRMSE), Nash–Sutcliffe efficiency (NSE), and Willmott index (WI). This metric suite is intentionally multi-perspective: in clinical regression, a model can appear competitive under correlation while still exhibiting clinically consequential magnitude error or systematic bias, and agreement indices help expose such mismatches. In particular, magnitude-based errors (MSE/RMSE/MAE) quantify absolute deviation, bias metrics (MBE) capture systematic over- or under-estimation, correlation (r) reflects association strength, R^2 captures explained variance, and agreement-style indices (NSE and WI) provide complementary perspectives on predictive adequacy under variability.

The aggregated baseline results are reported in Table 2, providing the empirical reference against which the effects of feature selection and hyperparameter optimization are subsequently assessed. By anchoring later improvements to this baseline set, the study ensures that the contribution of the proposed BER-driven unified optimization pipeline is interpretable, attributable, and methodologically grounded.

3.5 BER-Based Feature Selection

Feature selection in the present study is formulated as a binary combinatorial optimization problem defined over the d-dimensional feature space, where $d = 17$ corresponds to the total number of available attributes in the *Patient Vital Signs and Event Tracking* dataset. This formulation reflects the practical objective of identifying a compact representation that preserves predictive signal while removing redundant or weakly informative variables that can inflate variance, destabilize learning, and increase the computational burden of repeated model training during optimization. Let the complete feature vector be denoted as

$$\mathbf{X} = \{x_1, x_2, \ldots, x_d\}. \tag{2}$$

A candidate feature subset is represented by a binary vector

$$\mathbf{S} = \{s_1, s_2, \ldots, s_d\}, \quad s_i \in \{0, 1\}, \tag{3}$$

where $s_i = 1$ indicates inclusion of the i-th feature and $s_i = 0$ indicates exclusion. The optimization objective is to identify a subset $\mathbf{S}^*$ that minimizes prediction error

while controlling model complexity. In the wrapper setting adopted here, "complexity" is operationalized primarily through subset cardinality, which directly influences the dimensionality of the input to the learner and indirectly impacts regularization requirements, sensitivity to noise, and the stability of optimization trajectories.

To address this discrete search problem, the Al-Biruni Earth Radius (BER) optimization algorithm was adapted to operate in a binary domain. The original BER formulation, as introduced in El-kenawy et al. (2023), models cooperative swarm behavior inspired by Al-Biruni's geometric method for estimating the Earth's radius. The value of employing BER in this study is methodological: it provides a single metaheuristic backbone that can operate consistently across both discrete (feature selection) and continuous (hyperparameter optimization) spaces. The geometric principle underlying BER relies on the relationship

$$R = \frac{h\cos(\alpha)}{1-\cos(\alpha)}, \tag{4}$$

where h denotes the measured height and α represents the angular dip toward the horizon. In the optimization context, this formulation is abstracted to control adaptive search radius around candidate solutions. Conceptually, the radius mechanism induces a tunable balance between diversification and intensification: larger effective radii support broad exploratory moves over the search space, whereas smaller radii focus search locally around promising solutions. This interpretation motivates the algorithm's staged design in which population members transition between exploration, exploitation, and mutation-driven diversity restoration.

Population Representation.

Each search agent in BER is represented by a binary vector $\mathbf{S}^{(t)} \in \{0,1\}^d$ at iteration t. Under this encoding, an agent corresponds to a complete hypothesis about which clinical and temporal variables should be included in the predictive representation. The initial population of size N is randomly generated within the binary hypercube. Random initialization is used to avoid injecting handcrafted assumptions about feature relevance and to provide broad initial coverage of the combinatorial space, which grows exponentially with d even for modest feature counts.

To enable continuous update equations while maintaining binary feasibility, an intermediate real-valued vector $\mathbf{V}^{(t)} \in \mathbb{R}^d$ is maintained, and a transfer function $\mathcal{T}(\cdot)$ is applied to map continuous updates into binary space:

$$s_i^{(t+1)} = \begin{cases} 1, & \text{if } \sigma(v_i^{(t+1)}) > \tau, \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

where $\sigma(\cdot)$ denotes the sigmoid activation function and $\tau \sim \mathcal{U}(0,1)$ is a random threshold. This probabilistic binarization mechanism can be interpreted as a stochastic rounding operation: larger positive values of $v_i^{(t+1)}$ correspond to higher probability of selecting feature i, whereas strongly negative values push selection probability toward zero. The random threshold τ introduces controlled stochasticity, which is beneficial in binary combinatorial optimization because it prevents the entire population from deterministically collapsing into a narrow region of the search space early in the run. As a result, the algorithm can maintain diversity and continue to explore alternative subsets, even when early leaders appear promising.

Exploration Phase.

In BER, the population is dynamically divided into exploration and exploitation subgroups. During exploration, candidate solutions are perturbed using a radius-inspired update mechanism derived from

$$r = \frac{h\cos(x)}{1-\cos(x)}, \tag{6}$$

where $h \in [0, 2]$ and $x \in (0, \pi)$ are randomly sampled parameters controlling search amplitude. Within the feature selection context, this phase can be viewed as attempting coarse-grained modifications of subsets: adding or removing multiple features in a coordinated manner, rather than making only small local edits. Such behavior is desirable in early optimization because the algorithm has not yet identified reliable relevance structure; broad exploration reduces the chance of prematurely committing to a locally good but globally suboptimal subset.

The exploration displacement is computed as

$$\mathbf{D}^{(t)} = \mathbf{r}_1 \odot (\mathbf{S}^{(t)} - \mathbf{1}), \tag{7}$$

and the position update follows

$$\mathbf{V}^{(t+1)} = \mathbf{S}^{(t)} + \mathbf{D}^{(t)} \odot (2\mathbf{r}_2 - \mathbf{1}), \tag{8}$$

where $\odot$ denotes element-wise multiplication and $\mathbf{r}_1$, $\mathbf{r}_2$ are random coefficient vectors derived from the radius equation. This mechanism enables adaptive expansion and contraction of the search region, thereby mitigating premature convergence. In practice, the interaction of $\mathbf{r}_1$ and $\mathbf{r}_2$ allows different dimensions (features) to undergo different perturbation magnitudes within the same iteration, supporting heterogeneous search behavior across the feature index set. This is relevant in clinical tabular data because some variables can be consistently informative across admissions, while others may only contribute in interaction with specific contexts; exploration must remain flexible enough to discover such conditional utility.

Exploitation Phase.

Let $\mathbf{L}^{(t)}$ denote the current best solution (leader) at iteration t. Once promising subsets begin to emerge, BER shifts more emphasis toward exploitation, which aims to refine solutions near the leader while still allowing controlled deviations. Two exploitation strategies are employed:

(i) Directed Movement Toward the Leader:

$$\mathbf{D}^{(t)} = \mathbf{r}_3 \odot (\mathbf{L}^{(t)} - \mathbf{S}^{(t)}), \tag{9}$$

$$\mathbf{V}^{(t+1)} = \frac{r}{2} \odot (\mathbf{S}^{(t)} + \mathbf{D}^{(t)}). \tag{10}$$

This operator can be interpreted as an attraction mechanism, where each agent is pulled toward the leader subset in proportion to a randomized weighting $\mathbf{r}_3$, thereby enabling selective alignment on certain feature dimensions while allowing other dimensions to remain distinct. Such selective attraction is useful because it can propagate high-quality structural components of the leader (e.g., a core set of consistently valuable vital signs) throughout the population without forcing all agents to become identical.

(ii) Local Search Around the Leader:

$$\mathbf{V}^{(t+1)} = r \odot \left(\mathbf{L}^{(t)} + \mathbf{k}\right), \tag{11}$$

where

$$\mathbf{k} = z + \frac{2t^2}{T^2}, \tag{12}$$

$z \sim \mathcal{U}(0,1)$, and T is the maximum number of iterations. This mechanism intensifies search in the vicinity of the current best subset while maintaining adaptive diversity. The time-dependent term $\frac{2t^2}{T^2}$ increases as iterations progress, meaning that the local-search component evolves across the run rather than remaining static. In the feature selection setting, this can be understood as progressively tightening or reshaping the neighborhood exploration around the leader, enabling finer-grained adjustments later in optimization after broad structure has been discovered.

Mutation Mechanism.

To prevent stagnation in local optima, BER incorporates a mutation operator triggered when no improvement is observed for three consecutive iterations. This stagnation criterion functions as a practical diagnostic: if the leader does not improve for several steps, the population may have collapsed around a basin of attraction that does not contain better solutions, especially in a rugged combinatorial landscape. The mutation update is defined as

$$\mathbf{V}^{(t+1)} = \mathbf{k} \odot z - \frac{h\cos(x)}{1-\cos(x)}. \tag{13}$$

This stochastic perturbation enhances population diversity and supports renewed exploration. In wrapper-based feature selection, mutation is particularly important because the objective function (validation error plus subset penalty) can be noisy: small changes in a subset may yield variable effects depending on training stochasticity and partition composition. Mutation helps prevent the optimizer from interpreting short-term plateaus as convergence and encourages continued search for alternative feature combinations that may yield better generalization.

Fitness Function Formulation.

Feature selection is cast as a bi-objective optimization problem combining predictive error minimization and subset size reduction. The bi-objective framing is necessary because a pure error-minimization objective tends to favor larger subsets in observational tabular data, where weakly informative variables can still improve training fit but may harm generalization and interpretability. Conversely, aggressively minimizing subset size without regard to predictive error risks discarding complementary variables whose joint interactions are necessary for accurate prediction.

Let $\mathcal{E}(\mathbf{S})$ denote the regression error obtained by training FTLM using the selected feature subset $\mathbf{S}$. The fitness function is defined as

$$F(\mathbf{S}) = \alpha\,\mathcal{E}(\mathbf{S}) + \beta\frac{\|\mathbf{S}\|_0}{d}, \tag{14}$$

where $\|\mathbf{S}\|_0 = \sum_{i=1}^{d} s_i$ denotes the number of selected features, and $\alpha, \beta \in [0,1]$ satisfy $\alpha + \beta = 1$. This formulation enforces a trade-off between predictive accuracy and dimensionality reduction. The normalization by d ensures that the subset-size term remains scale-consistent and comparable across potential applications with different dimensionalities, and it bounds the penalty contribution to the unit interval. By tuning

α and β, the optimization can be biased toward predictive fidelity or toward sparsity; in this study, the formulation is used to ensure that subset compactness is explicitly rewarded while preserving the primary objective of predictive accuracy.

Comparative Benchmarking.

The performance of binary BER (bBER), derived from the Al-Biruni Earth Radius (BER) metaheuristic El-kenawy et al. (2023), was benchmarked against a set of widely used binary metaheuristic optimizers under strictly identical experimental conditions. The comparator suite was chosen to cover complementary search paradigms (swarm intelligence, evolutionary computation, and physics-inspired population search) and to ensure that each competing method corresponds to a well-defined and citable canonical source. This benchmarking design is essential for interpretability: feature selection outcomes can vary substantially across optimizers due to differences in exploration–exploitation scheduling, binarization schemes, and diversity control; therefore, controlled parity is required to attribute observed differences to the algorithmic dynamics rather than to unequal computational resources or inconsistent evaluation.

Concretely, the competing binary algorithms include: bGWO (binary Grey Wolf Optimizer), adapted from the Grey Wolf Optimizer proposed by Mirjalili *et al.* Mirjalili et al. (2014); bPSO (binary Particle Swarm Optimization), derived from the canonical PSO framework introduced by Kennedy and Eberhart Kennedy and Eberhart (1995); bBA (binary Bat Algorithm), based on Yang's Bat Algorithm Yang (2010b); bWAO (binary Whale Optimization Algorithm), derived from the Whale Optimization Algorithm introduced by Mirjalili and Lewis Mirjalili and Lewis (2016); bSBO (binary Satin Bowerbird Optimizer), based on the Satin Bowerbird Optimizer proposed by Moosavi and Bardsiri Moosavi and Bardsiri (2017); bSCA (binary Sine Cosine Algorithm), derived from Mirjalili's SCA formulation Mirjalili (2016); bFA (binary Firefly Algorithm), grounded in Yang's Firefly Algorithm Yang (2009); bGA (binary Genetic Algorithm), grounded in Holland's foundational evolutionary formulation Holland (1975); and bSAO (binary Secretary Bird Optimization Algorithm), derived from the Secretary Bird Optimization Algorithm introduced by Fu *et al.* Fu et al. (2024).

All algorithms were executed under identical population sizes, iteration limits, and evaluation budgets, ensuring that performance differences reflect optimization dynamics rather than unequal computational exposure. Statistical indicators—including average error, average selected feature ratio, average fitness, best fitness, worst fitness, and fitness standard deviation—were computed over repeated runs under controlled seed variation. In the context of wrapper-based selection, these stability indicators are essential: a method that occasionally finds a very good subset but exhibits high variance may be less reliable for clinical modeling than a method that produces consistently strong subsets across runs. The aggregated comparative results of these feature selection experiments are presented in Table 3, which provides a structured evaluation of optimizer stability, subset compactness, and predictive effectiveness across all competing binary methods.

3.6 Hyperparameter Optimization Framework

Hyperparameter optimization of FTLM is formulated as a continuous nonlinear optimization problem defined over a bounded real-valued search space. Unlike feature selection, which operates over a discrete combinatorial domain, hyperparameter tuning involves continuous control variables that govern the internal training dynamics, representational capacity, and regularization behavior of the model. Let the hyperparameter vector be denoted as

$$\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_p]^\top \in \mathbb{R}^p, \tag{15}$$

where p represents the number of tunable hyperparameters governing the internal learning dynamics of FTLM. These hyperparameters may include, for example, learning rate, regularization coefficients, embedding dimensions, depth-related parameters, or dropout ratios, depending on the architectural specification. Each hyperparameter θ_i is constrained within predefined lower and upper bounds,

$$\theta_i^{\min} \leq \theta_i \leq \theta_i^{\max}, \quad i = 1, 2, \ldots, p, \tag{16}$$

thereby defining a compact feasible domain $\Omega \subset \mathbb{R}^p$. The bounded nature of Ω is essential for metaheuristic optimization, as it ensures numerical stability, prevents divergence toward infeasible or non-meaningful configurations, and encodes prior practical knowledge about plausible parameter ranges.

Optimization Objective.

The hyperparameter optimization problem seeks a configuration $\boldsymbol{\theta}^*$ that minimizes predictive regression error on the training-validation partition. This formulation treats model training as an inner-loop process and hyperparameter search as an outer-loop optimization, yielding a nested structure in which each candidate $\boldsymbol{\theta}$ defines a complete model instantiation. Let $\mathcal{L}(\boldsymbol{\theta})$ denote the regression loss obtained by training FTLM with hyperparameters $\boldsymbol{\theta}$. The objective is expressed as

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta} \in \Omega} \mathcal{L}(\boldsymbol{\theta}). \tag{17}$$

In this study, $\mathcal{L}(\boldsymbol{\theta})$ is defined using mean squared error (MSE),

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{y}_i(\boldsymbol{\theta})\right)^2, \tag{18}$$

where y_i and $\hat{y}_i(\boldsymbol{\theta})$ denote observed and predicted values, respectively, and n is the number of validation samples. The choice of MSE as the primary optimization criterion ensures smoothness with respect to prediction outputs and penalizes larger deviations more strongly than absolute-error formulations. Importantly, although MSE is differentiable with respect to model parameters, the hyperparameter optimization layer remains derivative-free: BER operates solely on objective evaluations and does not require gradient information with respect to $\boldsymbol{\theta}$. This separation is advantageous because the objective surface induced by repeated model training can be highly nonconvex, noisy, and irregular, particularly when early stopping, stochastic optimization, and cross-validation are employed.

Continuous BER Formulation.

The Al-Biruni Earth Radius (BER) algorithm, originally proposed in El-kenawy et al. (2023), is adapted here to operate in a continuous hyperparameter space. In this formulation, each candidate solution corresponds directly to a real-valued hyperparameter vector. Each individual in the population is represented as

$$\mathbf{S}^{(t)} = [\theta_1^{(t)}, \theta_2^{(t)}, \ldots, \theta_p^{(t)}]^\top. \tag{19}$$

Initialization is performed uniformly within bounds:

$$\theta_i^{(0)} = \theta_i^{\min} + r_i(\theta_i^{\max} - \theta_i^{\min}), \quad r_i \sim \mathcal{U}(0, 1). \tag{20}$$

Uniform initialization ensures unbiased coverage of Ω and avoids clustering candidates near arbitrary default values. Since hyperparameter sensitivity can vary across dimensions, broad initial dispersion increases the likelihood of discovering diverse basins of attraction in early iterations.

Exploration Mechanism.

During exploration, BER computes a radius-inspired adaptive scaling factor based on the geometric relation

$$r = \frac{h\cos(x)}{1-\cos(x)}, \tag{21}$$

where $h \in [0, 2]$ and $x \in (0, \pi)$ are randomly generated. In the hyperparameter context, r acts as a dynamic step-size controller that modulates how far candidate vectors move within the continuous search space. Larger effective values of r encourage global exploration by permitting wide displacements across Ω, while smaller values promote more conservative refinement.

The displacement vector is defined as

$$\mathbf{D}^{(t)} = \mathbf{r}_1 \odot \left(\mathbf{S}^{(t)} - \mathbf{1}\right), \tag{22}$$

and the update rule becomes

$$\mathbf{S}^{(t+1)} = \mathbf{S}^{(t)} + \mathbf{D}^{(t)} \odot (2\mathbf{r}_2 - \mathbf{1}), \tag{23}$$

where $\odot$ denotes element-wise multiplication and $\mathbf{r}_1, \mathbf{r}_2$ are random vectors scaled by r. This formulation enables adaptive expansion of the search region and prevents early contraction around suboptimal hyperparameter configurations. Because each hyperparameter dimension may influence model behavior differently (e.g., learning rate versus embedding dimension), the element-wise scaling supports heterogeneous perturbation magnitudes across components. Such flexibility is important in high-dimensional continuous spaces, where uniform step sizes may either overshoot narrow optima or fail to escape flat regions.

Exploitation Mechanisms.

Let $\mathbf{L}^{(t)}$ denote the best-performing hyperparameter vector at iteration t. Once promising regions of Ω have been identified, BER increases exploitation pressure to refine solutions near $\mathbf{L}^{(t)}$. Two exploitation strategies are employed:

(i) Directed Convergence Toward the Leader:

$$\mathbf{D}^{(t)} = \mathbf{r}_3 \odot \left(\mathbf{L}^{(t)} - \mathbf{S}^{(t)}\right), \tag{24}$$

$$\mathbf{S}^{(t+1)} = \frac{r}{2}\left(\mathbf{S}^{(t)} + \mathbf{D}^{(t)}\right). \tag{25}$$

This mechanism pulls candidate vectors toward the current leader, encouraging population contraction around high-performing regions. The random coefficient vector $\mathbf{r}_3$ introduces dimension-wise variability, ensuring that convergence is not overly rigid and that individual hyperparameters can adjust at different rates.

(ii) Local Search Around the Leader:

$$\mathbf{S}^{(t+1)} = r\left(\mathbf{L}^{(t)} + \mathbf{k}\right), \tag{26}$$

with

$$\mathbf{k} = z + \frac{2t^2}{T^2}, \tag{27}$$

where $z \sim \mathcal{U}(0,1)$ and T is the maximum iteration count. This quadratic schedule increases local intensification as iterations progress, thereby enhancing convergence precision. The time-dependent term $\frac{2t^2}{T^2}$ ensures that as t approaches T, exploitation becomes more pronounced. In practical terms, early iterations emphasize diversity and coarse adjustment, whereas later iterations focus on fine-tuning around the most promising hyperparameter configurations identified thus far.

Mutation Strategy.

To avoid stagnation, BER triggers mutation when no fitness improvement is observed for three consecutive iterations:

$$\mathbf{S}^{(t+1)} = \mathbf{k} \odot z - \frac{h\cos(x)}{1-\cos(x)}. \tag{28}$$

This probabilistic perturbation preserves diversity and facilitates escape from local minima. In hyperparameter optimization, local minima can arise due to complex interactions between learning rate, regularization, and architectural depth, especially when validation performance is influenced by stochastic training dynamics. The mutation operator injects renewed variability into the population, enabling exploration of alternative regions that may not be reachable through incremental exploitation alone.

Comparative Optimization Framework.

For continuous hyperparameter optimization, BER-based tuning was compared against the corresponding continuous-domain metaheuristics under identical search space definitions and computational budgets. The comparator set comprises GWO Mirjalili et al. (2014), PSO Kennedy and Eberhart (1995), BA Yang (2010b), WAO Mirjalili and Lewis (2016), SBO Moosavi and Bardsiri (2017), SCA Mirjalili (2016), FA Yang (2009), GA Holland (1975), and SAO (Secretary Bird Optimization Algorithm) Fu et al. (2024), alongside BER El-kenawy et al. (2023). This selection ensures that each competing optimizer is anchored to an established primary reference and that the benchmark reflects methodological breadth rather than variations of a single search family. The included algorithms represent distinct search philosophies—social sharing (PSO), leadership hierarchy (GWO), bio-inspired echolocation (BA), spiral encircling (WAO), mating-display competition (SBO), trigonometric position updates (SCA), attraction-based movement (FA), evolutionary recombination (GA), and avian predation-inspired search (SAO)—thereby providing a comprehensive comparative landscape.

All algorithms were configured with equal population sizes and maximum iteration counts to preserve strict fairness. The hyperparameter bounds and dimensionality were identical across methods, ensuring that each optimizer operated within the same feasible domain Ω Ibrahim et al. (2026); Hussein et al. (2025); Qaraad et al. (2022b). For each candidate hyperparameter vector, FTLM was trained and evaluated using the same validation protocol described earlier, with performance assessed using MSE, RMSE, MAE, MBE, Pearson correlation coefficient (r), coefficient of determination (R^2), RRMSE, Nash–Sutcliffe efficiency (NSE), and Willmott index (WI). This multi-metric evaluation guards against over-reliance on a single summary statistic and provides complementary perspectives on magnitude error, bias, correlation structure, and predictive agreementQaraad et al. (2021, 2022a, 2026).

The aggregated hyperparameter optimization results for BER and all competitors are presented in Table 5, providing a multi-metric comparison of optimization effectiveness within a shared continuous search space. This table serves as the empirical basis for assessing whether BER's adaptive radius-based exploration–exploitation balance translates into consistent improvements in predictive performance and stability when applied to structured hospital vital-sign prediction.

3.7 Performance Metrics

A comprehensive evaluation framework was adopted to assess predictive performance across all experimental stages. The selected metrics capture complementary aspects of regression behavior, including average error magnitude, bias, dispersion, explained variance, relative error scaling, and agreement between observed and predicted values. In structured hospital monitoring, relying on a single metric can lead to incomplete or even misleading conclusions: a model may achieve strong correlation while exhibiting clinically meaningful magnitude error, or it may display low average error while systematically overestimating high-risk observations. The multi-metric design therefore ensures that improvements obtained through feature selection or hyperparameter optimization are robust across multiple statistical perspectives.

Let $\{y_i\}_{i=1}^n$ denote the observed target values and $\{\hat{y}_i\}_{i=1}^n$ the corresponding model predictions for a dataset of size n. Let $\bar{y}$ denote the sample mean of the observed values:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i. \tag{29}$$

All metrics are computed on the held-out test set unless explicitly stated otherwise.

Mean Squared Error (MSE).

Mean Squared Error quantifies the average of squared residuals and is defined as

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 . \tag{30}$$

The quadratic term in MSE places greater emphasis on larger deviations, making it sensitive to extreme prediction errors. In clinical regression contexts, this sensitivity is desirable when large deviations correspond to clinically unacceptable misestimations. MSE is considered the primary optimization objective during the hyperparameter tuning stage because of its differentiable structure with respect to model outputs and its strong penalization of variance. However, because of its scale dependence and sensitivity to outliers, it is complemented by additional metrics for comprehensive evaluation.

Root Mean Squared Error (RMSE).

Root Mean Squared Error is the square root of MSE:

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}}. \tag{31}$$

RMSE restores the error metric to the same scale as the target variable, thereby improving interpretability while preserving sensitivity to large deviations. In practice, RMSE can be directly interpreted in the physical units of the predicted clinical variable, facilitating domain-specific interpretation of predictive dispersion.

Mean Absolute Error (MAE).

Mean Absolute Error measures the average magnitude of residuals:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \,. \tag{32}$$

Unlike MSE, MAE does not disproportionately penalize outliers, providing a more robust assessment of central predictive tendency. MAE is particularly informative when the target distribution contains occasional extreme values; it reflects the typical absolute deviation rather than amplifying rare extremes.

Mean Bias Error (MBE).

Mean Bias Error evaluates systematic overestimation or underestimation:

$$\mathrm{MBE} = \frac{1}{n}\sum_{i=1}^{n} (\hat{y}_i - y_i) \,. \tag{33}$$

A positive MBE indicates overprediction, whereas a negative value indicates underprediction. This metric is particularly important in clinical contexts where systematic bias may have operational consequences. For example, consistent overestimation could lead to unnecessary interventions, whereas underestimation may delay escalation of care. MBE therefore complements magnitude-based metrics by explicitly quantifying directional error.

Pearson Correlation Coefficient (r).

The Pearson correlation coefficient measures the linear association between observed and predicted values:

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}}, \tag{34}$$

where $\bar{\hat{y}}$ denotes the mean of predicted values. The coefficient $r \in [-1, 1]$ captures directional linear dependence independent of scale. A high r indicates that the model tracks the variability pattern of the observed values, even if magnitude discrepancies exist. In clinical modeling, strong correlation can signal that the model captures relative risk ordering, but it must be interpreted alongside magnitude and bias metrics.

Coefficient of Determination (R^2).

The coefficient of determination quantifies the proportion of variance in the observed data explained by the model:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{35}$$

An R^2 value approaching unity indicates strong explanatory power, whereas negative values imply performance worse than a mean-based predictor. R^2 provides a variance-normalized measure of predictive strength and is useful for comparing performance across models operating on the same target variable.

Relative Root Mean Squared Error (RRMSE).

Relative RMSE normalizes RMSE with respect to the mean of observed values:

$$\text{RRMSE} = \frac{\text{RMSE}}{\bar{y}} \times 100. \tag{36}$$

This normalization expresses predictive dispersion as a percentage of the observed mean, facilitating scale-independent comparison. In settings where the magnitude of the target variable may vary across datasets or subgroups, RRMSE provides a relative perspective on error magnitude.

Nash–Sutcliffe Efficiency (NSE).

Nash–Sutcliffe Efficiency evaluates predictive skill relative to the mean benchmark:

$$\text{NSE} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{37}$$

Although algebraically equivalent to R^2 under certain formulations, NSE is interpreted as a predictive efficiency measure, where values greater than zero indicate performance superior to mean prediction. NSE emphasizes how much better the model performs compared to a naive baseline that predicts $\bar{y}$ for all samples.

Willmott Index (WI).

The Willmott Index of agreement evaluates the degree of predictive conformity:

$$\text{WI} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}\left(|\hat{y}_i - \bar{y}| + |y_i - \bar{y}|\right)^2}. \tag{38}$$

The WI metric ranges between 0 and 1, with values closer to unity indicating stronger agreement. Unlike correlation-based measures, WI accounts for both proportional and additive discrepancies. It is sensitive to the relative alignment of predictions and observations across the full range of variability.

Rationale for Multi-Metric Evaluation.

The joint use of error-based (MSE, RMSE, MAE, MBE), correlation-based (r), variance-based (R^2, NSE), and agreement-based (WI) metrics ensures that predictive assessment is not biased toward a single statistical property. In clinical modeling scenarios involving heterogeneous physiological signals, models may exhibit low absolute error yet retain systematic bias, or achieve high correlation while underperforming in magnitude accuracy. The adopted evaluation framework therefore provides a multidimensional characterization of predictive performance across all experimental configurations, enabling robust interpretation of improvements obtained through feature selection and hyperparameter optimization.

All evaluation metrics used in this study were selected to match the regression nature of the prediction task and to provide complementary views of model performance. Magnitude-based errors (MSE, RMSE, and MAE) assess absolute predictive accuracy, MBE quantifies directional bias, r and R^2 describe association and explained variance, RRMSE provides scale-normalized error, and NSE and WI measure predictive efficiency and agreement. This combination reduces over-reliance on a single metric and is appropriate for structured clinical prediction, where both numerical accuracy and stability are important.

4 Experimental Setup

This section describes the computational environment, model search space, and optimizer-specific configurations used to ensure reproducibility and fair comparison across all experiments. All compared methods were executed under the same hardware environment and were evaluated using the same preprocessing, validation, and optimization pipeline.

4.1 Computational Environment

All experiments were conducted on a workstation with the following specifications: Intel Core i7-10750H processor at 2.60 GHz with 6 cores and 12 threads, NVIDIA GeForce RTX 2060 GPU with 6 GB GDDR6 memory, 16 GB DDR4 RAM at 3200 MHz, and a 512 GB NVMe SSD. The operating system was Windows 11 Pro 64-bit. These specifications provided a consistent execution platform for all baseline, feature-selection, and hyperparameter-optimization experiments.

4.2 Hyperparameter Search Space and Optimizer Settings

To ensure a controlled and transparent comparison, all optimizers searched within the same FTLM hyperparameter space. The selected search ranges were chosen to cover practically relevant training configurations while avoiding excessively unstable settings. Table 1 summarizes the search space and the final configurations obtained by BER and the competing optimizers.

Table 1 Hyperparameter search space and final configurations obtained by the compared optimizers.

Hyperparameter	Search Space	BER	GWO	PSO	BA	WAO	SBO	SCA	FA	GA	SAO
Learning Rate	[0.0001–0.1]	0.0012	0.0031	0.0028	0.0035	0.0039	0.0042	0.0044	0.0051	0.0058	0.0063
Batch Size	{16, 32, 64, 128}	32	64	64	64	64	64	128	128	128	128
Hidden Layers	[1–5]	3	2	2	3	3	3	2	2	2	2
Neurons/Layer	[16–256]	128	96	96	112	112	108	96	88	84	80
Dropout Rate	[0.0–0.5]	0.12	0.22	0.24	0.25	0.27	0.28	0.31	0.33	0.35	0.37
L2 Regularization	[1e-5–1e-2]	0.00018	0.00041	0.00038	0.00045	0.00052	0.00061	0.00074	0.00089	0.00094	0.00102
Optimizer	{Adam, RMSprop, SGD}	Adam	Adam	RMSprop	Adam	Adam	RMSprop	RMSprop	SGD	SGD	SGD
Activation	{ReLU, Tanh, Sigmoid}	ReLU	ReLU	ReLU	Tanh	Tanh	Tanh	ReLU	Tanh	Sigmoid	Sigmoid
Epochs	[50–300]	180	140	145	155	160	165	170	175	182	190

The final configurations indicate that BER converged to a comparatively conservative and stable FTLM setting, characterized by a lower learning rate, moderate depth, higher neurons per layer, lower dropout, and weaker regularization than most competing methods. In contrast, several alternative optimizers converged to progressively higher learning rates, stronger regularization, and larger dropout values, which may partially explain the differences observed in convergence behavior and predictive performance.

The use of a common search space ensures that the comparison reflects the effectiveness of the optimization strategies rather than differences in allowable model configurations. In this way, all optimizers were given equal opportunity to explore the same FTLM design space under the same experimental conditions.

To ensure fair comparison with relevant competitive methods, all optimizers were evaluated under identical experimental conditions. Specifically, the same dataset, preprocessing pipeline, train–test partition, cross-validation protocol, FTLM search space, population size, iteration limit, and stopping policy were used for BER, GWO, PSO, BA, WAO, SBO, SCA, FA, GA, and SAO. In this way, performance differences reflect the effectiveness of the optimization strategies rather than variations in data handling, search budget, or model configuration.

5 Results

5.1 Baseline Model Performance

The baseline evaluation establishes the intrinsic predictive behavior of all considered learning models prior to any feature selection or hyperparameter optimization. Table 2 reports the complete set of performance metrics, including error-based, correlation-based, and agreement-based indicators. These results provide a structured reference for interpreting the incremental contribution of subsequent optimization stages. Because all models were trained under identical preprocessing pipelines, data partitions, and computational budgets, differences in performance can be attributed to differences in inductive bias and representational capacity rather than to experimental variability.

Table 2 Baseline performance comparison across learning models.

Model	MSE	RMSE	MAE	MBE	r	R^2	RRMSE	NSE	WI
FTLM	0.012028	0.10967	0.075866	0.052781	0.771073	0.782413	5.157115	0.799616	0.841358
CTSM	0.067973	0.260717	0.088968	0.076258	0.773284	0.782875	6.702019	0.796856	0.802959
VAST	0.086502	0.294112	0.105263	0.090896	0.743547	0.756913	6.874417	0.794965	0.795793
LSTM	0.0958	0.309514	0.122044	0.15027	0.758371	0.771089	7.015949	0.773679	0.754339
DTCN	0.208043	0.456119	0.175259	0.176334	0.704451	0.723396	7.527113	0.714233	0.709692
TST	0.469931	0.685516	0.254484	0.349321	0.70032	0.714393	8.016678	0.708144	0.685327
VAE	0.793112	0.890625	0.304678	0.477633	0.667284	0.685537	8.588071	0.680661	0.675797

From an error-based perspective, FTLM exhibits the lowest mean squared error (0.012028) and root mean squared error (0.10967), indicating superior magnitude accuracy relative to all other evaluated models. The mean absolute error (0.075866) further confirms its consistency in minimizing average residual deviation. These values collectively indicate that, under default hyperparameter settings, FTLM achieves the most precise numerical approximation of the target variable. In contrast, models such as VAE and TST demonstrate substantially higher error values, reflecting diminished predictive precision under baseline configurations. The progressive increase in MSE and RMSE from tabular attention-based models (FTLM, CTSM) to temporal and generative architectures (LSTM, DTCN, TST, VAE) suggests that, without specialized tuning, structured tabular learners are inherently better aligned with the statistical properties of the dataset.

The mean bias error (MBE) provides insight into systematic prediction tendencies. FTLM reports an MBE of 0.052781, indicating moderate positive bias. Although CTSM and VAST also exhibit positive bias, their larger MSE and RMSE values suggest that the increased dispersion outweighs any marginal advantage in bias magnitude. Models such as LSTM, DTCN, and TST show progressively larger MBE values, reflecting greater systematic overestimation under default parameter settings. VAE displays the largest MBE (0.477633), indicating pronounced upward bias, which may stem from latent compression effects that distort scale relationships when not carefully regularized. In clinical modeling, systematic bias is particularly important, as consistent overprediction or underprediction can have operational consequences even when correlation remains moderate.

Correlation-based evaluation reveals that CTSM achieves the highest Pearson correlation coefficient ($r = 0.773284$), followed closely by FTLM ($r = 0.771073$). However, correlation alone does not fully capture predictive reliability, as it measures linear association rather than magnitude accuracy. This is reflected in the fact that CTSM, despite a marginally higher correlation, exhibits substantially larger MSE and RMSE compared to FTLM. Thus, CTSM may preserve rank ordering effectively while

sacrificing absolute precision. This distinction illustrates why multi-metric evaluation is essential in regression tasks: two models can display similar correlation yet differ markedly in residual dispersion.

The coefficient of determination (R^2) follows a similar pattern. CTSM achieves $R^2 = 0.782875$, closely aligned with FTLM ($R^2 = 0.782413$). Nevertheless, the lower absolute error metrics of FTLM indicate stronger predictive stability. Lower-performing models such as VAE ($R^2 = 0.685537$) and TST ($R^2 = 0.714393$) explain a smaller proportion of variance, consistent with their elevated residual magnitudes. The near equivalence between r and R^2 trends further confirms that explanatory strength and correlation are strongly aligned for these baselines.

Relative RMSE (RRMSE) further illustrates scale-normalized dispersion. FTLM achieves the lowest RRMSE (5.157115), indicating reduced relative prediction spread when normalized by the mean of observed values. Competing models display progressively higher RRMSE values, with VAE reaching 8.588071, reflecting comparatively weaker scale-adjusted performance. The RRMSE metric is particularly informative when assessing models across different operational ranges, as it contextualizes error magnitude relative to baseline variability.

Efficiency and agreement measures provide additional perspective. FTLM reports the highest Nash–Sutcliffe efficiency (0.799616) and Willmott index (0.841358), indicating strong agreement between observed and predicted values across both variance-based and normalized agreement formulations. Models such as DTCN, TST, and VAE demonstrate lower NSE and WI values, signaling reduced predictive consistency and agreement. Notably, while CTSM approaches FTLM in R^2 and r, its WI value (0.802959) remains lower than that of FTLM, indicating comparatively weaker overall agreement structure.

Overall, the baseline analysis reveals that FTLM provides the most favorable balance between magnitude accuracy, correlation strength, and agreement consistency under default conditions. While certain models approach FTLM in individual metrics—such as CTSM in correlation—the aggregate performance profile consistently favors FTLM. These findings establish FTLM as the primary candidate for subsequent feature selection and hyperparameter optimization stages, where its predictive capacity is further examined under enhanced search strategies.

Following the quantitative reporting of baseline models, a series of complementary visual analytics were conducted to examine dispersion, distributional characteristics, inter-metric dependencies, and relative ranking consistency across models. These graphical analyses provide deeper statistical insight beyond tabulated mean values, allowing assessment of robustness, variance structure, and metric interrelationships.

Figure 10 combines violin plots with embedded boxplots for all evaluation metrics. The violin component illustrates the kernel-estimated distribution of metric values across models, while the boxplot summarizes median, interquartile range, and dispersion. This combined representation reveals both central tendency and distributional density.

Error-based metrics (MSE, RMSE, MAE, MBE, RRMSE) exhibit right-skewed distributions, reflecting larger deviations for weaker baseline models. In contrast, performance metrics such as r, R^2, NSE, and WI show concentrated density in higher-value regions, indicating moderate-to-strong predictive capability for selected models. The relatively narrow interquartile range for the FTLM baseline suggests greater stability compared to architectures exhibiting wider spread. This visual evidence supports the interpretation that FTLM's superiority is not limited to mean performance but also extends to dispersion characteristics.

Figure 11 presents a faceted bar visualization for each metric separately. This representation facilitates direct comparison of model ranking consistency across metrics.

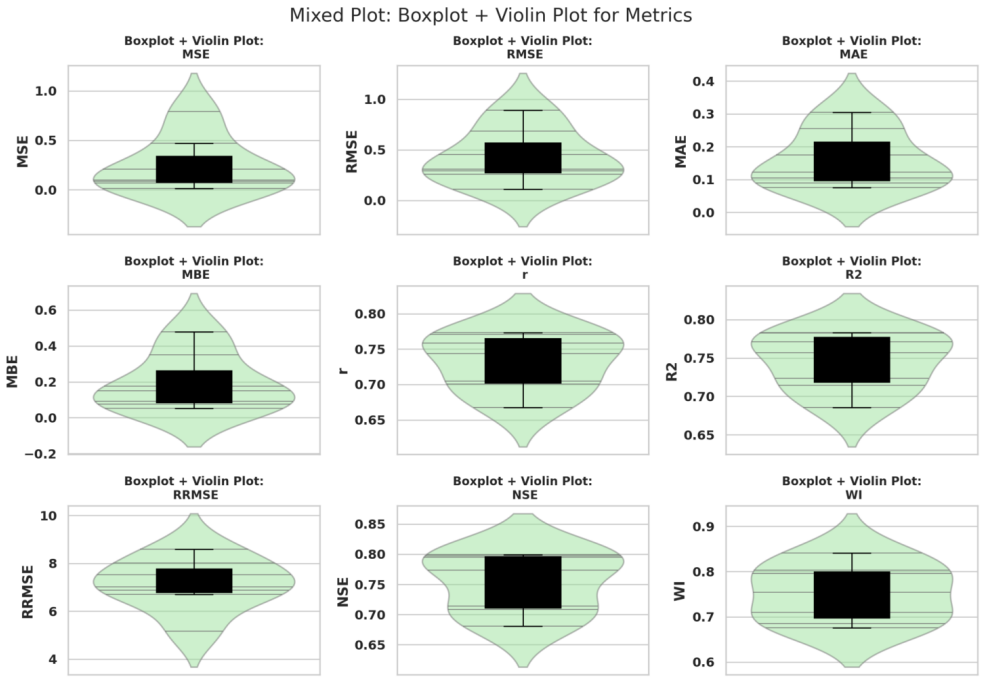

Fig. 10 Mixed violin and boxplot representation of performance metrics across baseline models.

Each facet isolates a single performance measure, allowing inspection of how model ordering changes—or remains stable—across evaluation dimensions.

FTLM consistently demonstrates lower error magnitudes and higher correlation-based scores relative to most baselines. Conversely, generative or transformer-based architectures exhibit increased error dispersion. The alignment of ranking patterns across MSE, RMSE, and MAE confirms internal consistency among scale-dependent error metrics. Deviations between correlation-based and magnitude-based rankings are minimal for leading models, reinforcing the robustness of the overall ordering.

Figure 12 depicts Q–Q plots evaluating normality assumptions for each metric. Points generally follow the reference line with mild deviations at distribution tails, indicating approximate Gaussian behavior for aggregated metric values across models.

Deviation at upper quantiles for MSE and RMSE reflects heterogeneity among weaker models, particularly VAE and TST. Correlation-based metrics show stronger linear conformity, implying reduced skewness and greater homogeneity across models in those measures. These observations justify the use of mean-based aggregation while acknowledging moderate tail variability in error-based metrics.

Figure 13 integrates mean values and standard deviations for MSE, RMSE, and MAE. This visualization emphasizes the trade-off between central performance and variability. Stability across repeated runs is particularly relevant for optimization frameworks, as highly variable baselines may exaggerate apparent gains after tuning.

FTLM not only achieves lower mean error values but also maintains comparatively controlled dispersion, indicating robust convergence behavior. Models with higher means often exhibit amplified standard deviations, suggesting sensitivity to initialization or data partitioning. Stability across repeated runs is therefore an important discriminative factor in model selection.

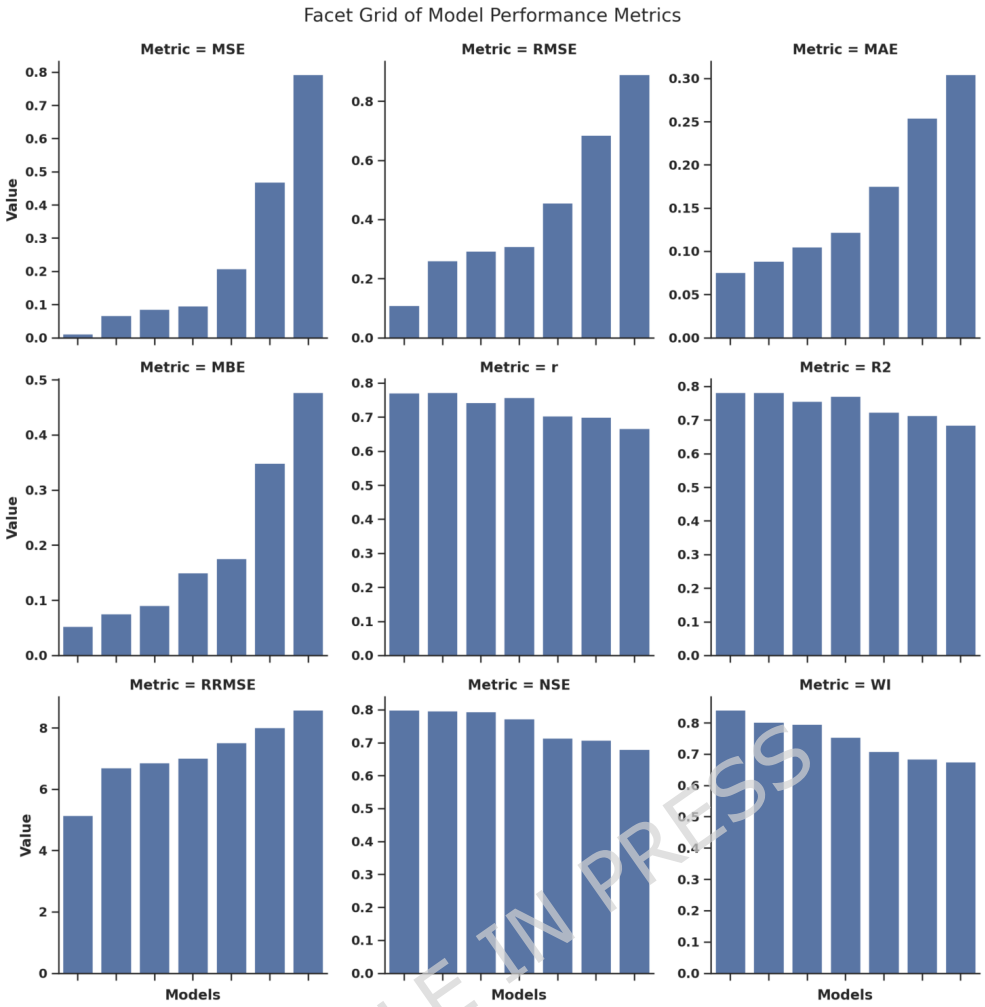

Fig. 11 Facet grid comparison of model performance across individual metrics.

Figure 14 presents a parallel coordinates plot summarizing all metrics simultaneously. Each polyline represents a model across standardized metric axes, providing a holistic, multidimensional view of performance trade-offs.

FTLM exhibits a favorable trajectory characterized by minimal error metrics and elevated correlation-based indicators. Less competitive models display divergent trajectories with simultaneous error amplification and reduction in explanatory power. The multidimensional consistency of FTLM across axes demonstrates balanced performance rather than metric-specific optimization. This visualization confirms that FTLM does not simply excel in one dimension but maintains coherent superiority across the majority of evaluation measures.

Figure 15 illustrates hierarchical clustering of metric correlations. Strong positive correlations are observed among MSE, RMSE, MAE, and MBE, reflecting shared scale dependence. Conversely, these metrics exhibit strong negative correlations with r, R^2, NSE, and WI, which measure explanatory strength.

The clustering structure confirms two principal metric groups: error-based and goodness-of-fit-based. This separation validates the necessity of reporting complementary metric families to avoid redundancy. The clear dichotomy also indicates that

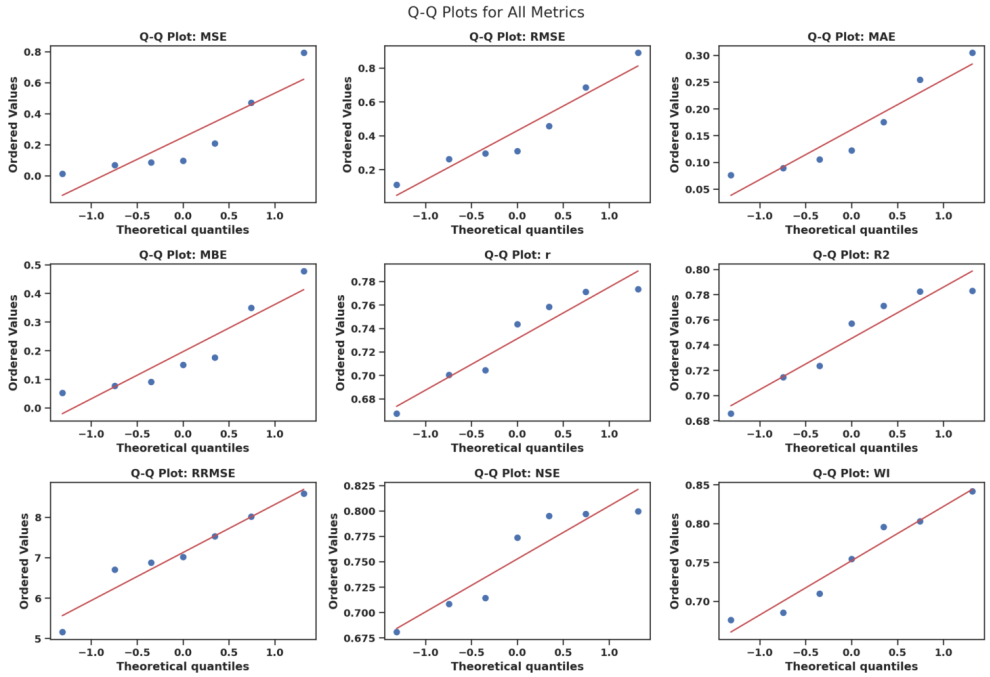

Fig. 12 Q–Q plots assessing distributional characteristics of performance metrics.

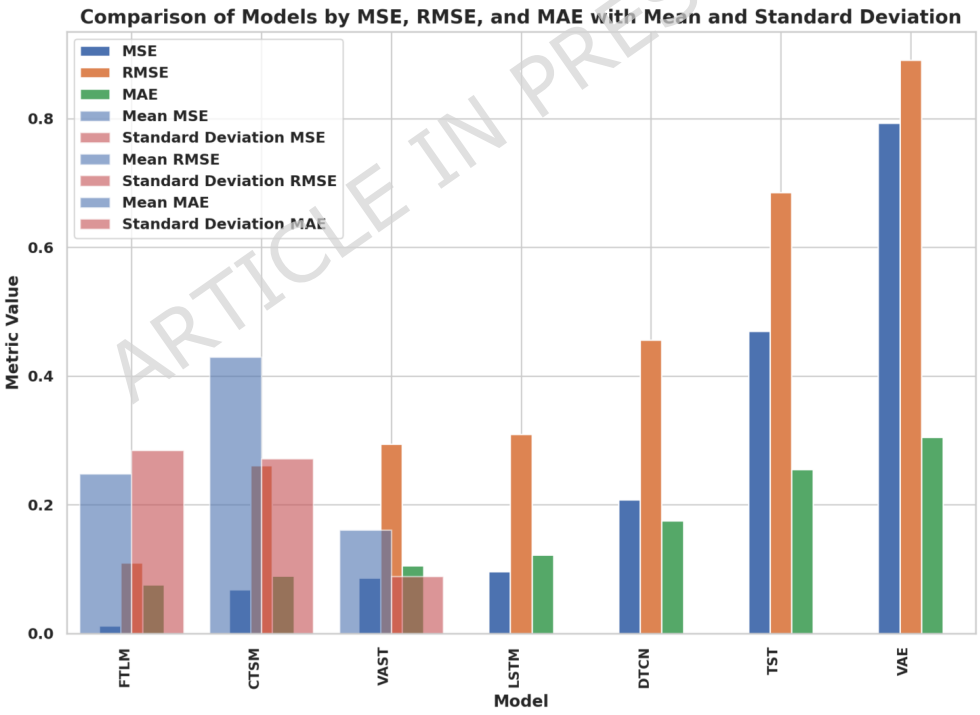

Fig. 13 Comparison of baseline models using mean and standard deviation for primary error metrics.

improvements in one cluster are predictably associated with changes in the other, reinforcing the internal consistency of the evaluation framework.

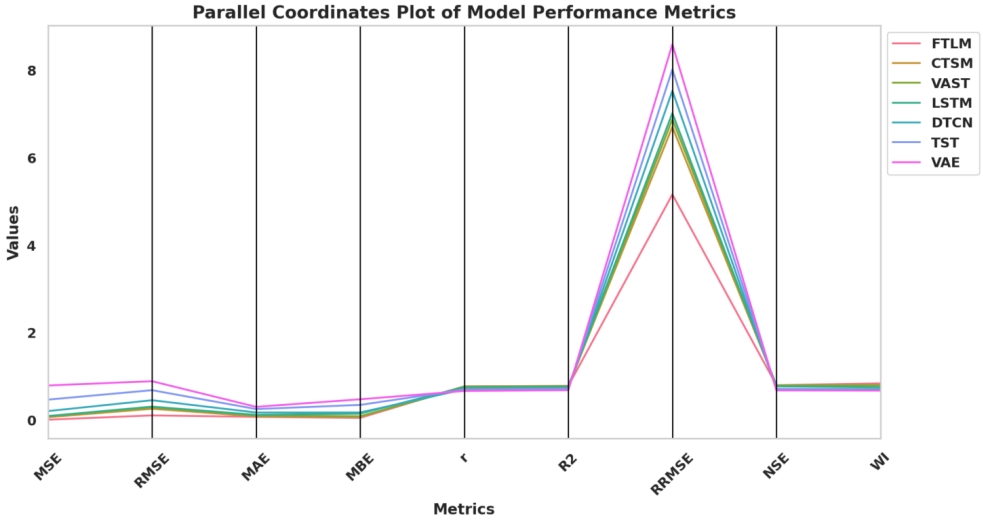

Fig. 14 Parallel coordinates plot illustrating multidimensional metric comparison across models.

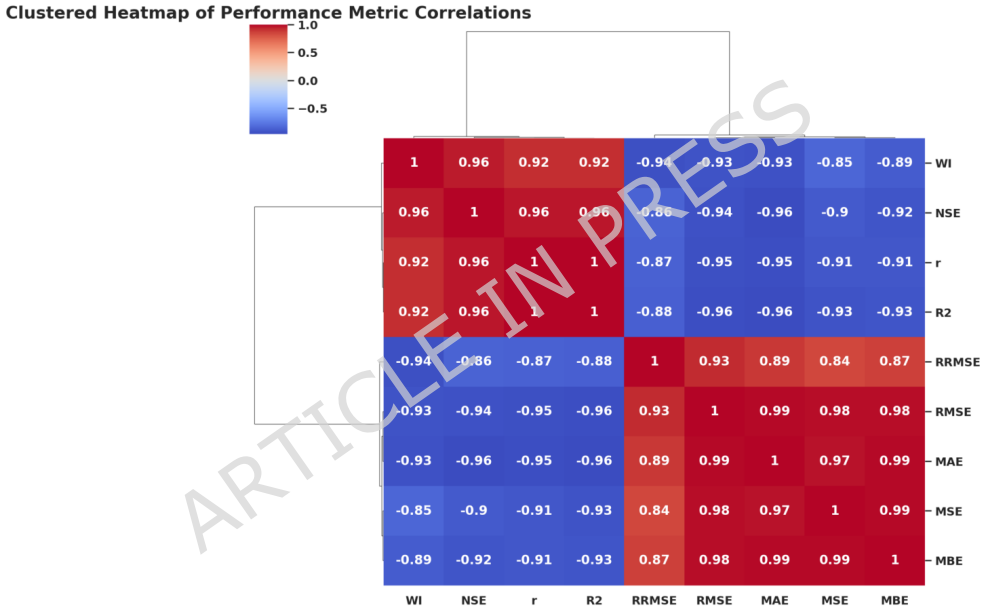

Fig. 15 Clustered heatmap of correlations among performance metrics.

Figure 16 displays combined density and KDE plots for each metric. The unimodal yet mildly skewed shapes of error metrics confirm moderate variability among models. Correlation-based metrics demonstrate tighter clustering near upper bounds, indicating relative performance saturation among leading baselines.

The distributional shapes reinforce earlier findings: FTLM contributes to the left tail (lower error) and right tail (higher goodness-of-fit), shifting the aggregate density toward favorable regions. Models such as VAE extend the right tail of error distributions, accounting for skewness.

Figure 17 presents the full correlation matrix of evaluation metrics. Extremely high positive correlations among MSE, RMSE, MAE, and MBE (> 0.97) confirm their

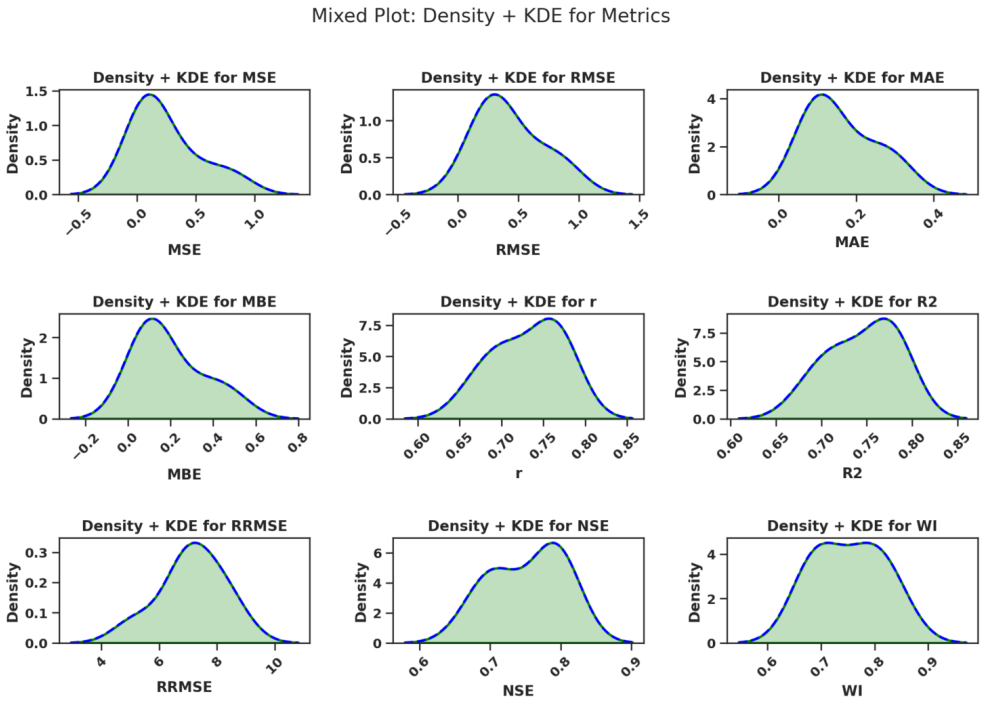

Fig. 16 Density and kernel distribution plots for all evaluation metrics.

shared sensitivity to error magnitude. Strong negative correlations with r, R^2, NSE, and WI (approximately -0.90 or lower) illustrate inverse proportionality between error and explanatory strength.

The near-perfect correlation between r and R^2 reflects their mathematical linkage, whereas NSE and WI provide closely aligned but not identical goodness-of-fit perspectives. These results demonstrate metric redundancy within clusters and support the use of representative metrics in optimization objectives.

The integrated visual analyses confirm that FTLM provides consistently favorable performance across error-based and goodness-of-fit metrics, while maintaining comparatively stable variance. Metric clustering analysis demonstrates redundancy among scale-dependent errors and strong inverse association with explanatory metrics. These observations motivate the subsequent BER-assisted optimization framework, aiming to further reduce error magnitude while preserving stability and generalization capacity.

5.2 Feature Selection Benchmarking

The effectiveness of binary feature selection strategies was evaluated using a comprehensive benchmarking protocol. Table 3 summarizes the aggregated results for all considered binary metaheuristic optimizers, including bBER, bGWO, bPSO, bBA, bWAO, bSBO, bSCA, bFA, bGA, and bSAO. The reported indicators include average prediction error, average selected feature ratio, average fitness, best fitness, worst fitness, and standard deviation of fitness across repeated runs. All algorithms were executed under identical population sizes, iteration limits, and evaluation budgets, thereby ensuring strict computational parity. Consequently, observed differences reflect intrinsic search dynamics rather than unequal exposure to function evaluations.

From the perspective of predictive error minimization, bBER achieves the lowest average error (0.43707) among all compared algorithms. The next closest competitor

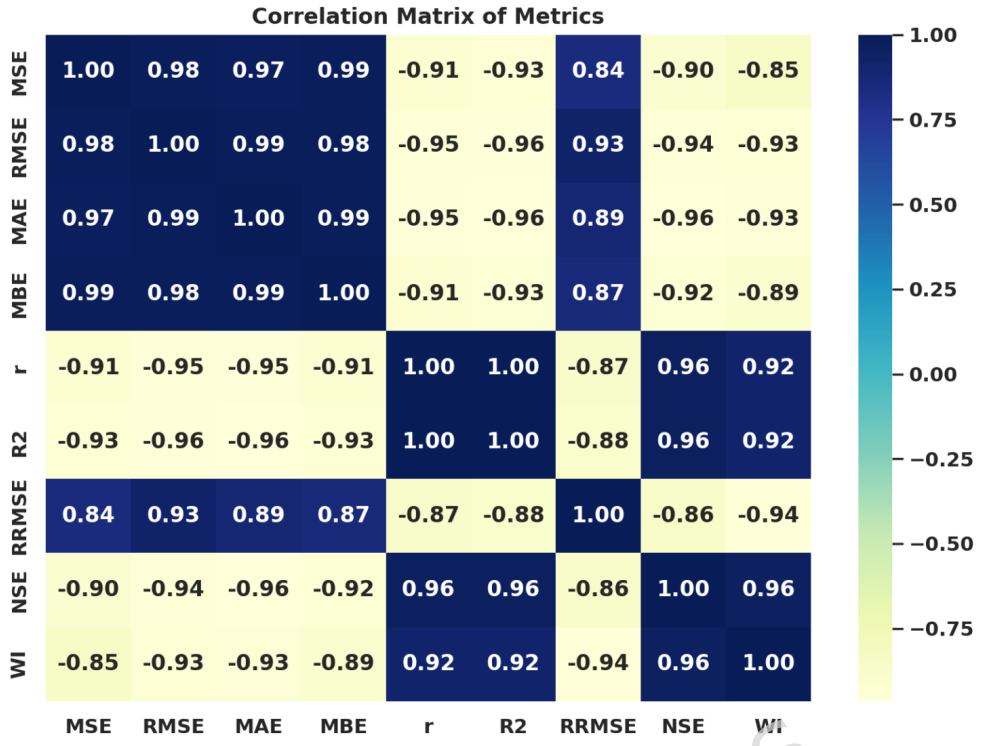

Fig. 17 Correlation matrix of evaluation metrics across baseline models.

Table 3 Comparative benchmarking of binary feature selection optimizers.

Metric	bBER	bGWO	bPSO	bBA	bWAO	bSBO	bSCA	bFA	bGA	bSAO
Average error	0.43707	0.50117	0.59467	0.60427	0.59447	0.60297	0.48227	0.59307	0.57447	0.50767
Average Select size	0.38987	0.73077	0.69647	0.83587	0.85987	0.86677	0.52057	0.73097	0.63887	0.73927
Average Fitness	0.50027	0.53237	0.62147	0.64437	0.62927	0.66117	0.53047	0.67337	0.63447	0.54777
Best Fitness	0.40207	0.48587	0.60177	0.53407	0.59337	0.60427	0.50827	0.59207	0.53777	0.43747
Worst Fitness	0.50057	0.50587	0.66947	0.63567	0.66947	0.68397	0.58447	0.68967	0.65287	0.53907
Std. Fitness	0.32257	0.35307	0.43327	0.44317	0.43547	0.49417	0.33577	0.47007	0.43547	0.34657

is bSCA with an average error of 0.48227, followed by bGWO (0.50117) and bSAO (0.50767). Algorithms such as bBA, bSBO, and bFA exhibit higher average errors exceeding 0.59, indicating reduced predictive effectiveness when selecting feature subsets under equivalent computational budgets. The gap between bBER and the next-best alternative (bSCA) exceeds 0.045 in absolute error units, which represents a meaningful reduction given that all models were evaluated under identical cross-validation protocols. This improvement indicates that the search trajectory induced by BER more effectively navigates the combinatorial feature space.

Feature subset compactness is evaluated through the average selected size ratio. bBER selects, on average, 0.38987 of the available features, representing the smallest subset among all optimizers. In contrast, several competing algorithms select substantially larger subsets, including bSBO (0.86677), bWAO (0.85987), and bBA (0.83587). Even mid-performing optimizers such as bPSO (0.69647) and bGA (0.63887) retain a significantly higher proportion of input variables. The ability of bBER to achieve lower predictive error while simultaneously selecting a smaller subset demonstrates a favorable trade-off between dimensionality reduction and accuracy preservation. This indicates that bBER is not merely achieving improved performance through inclusion

of more variables, but rather by identifying compact and informative subsets that reduce redundancy and noise.

The average fitness metric, which integrates prediction error and subset size through the composite objective function defined in Section 3.4, further reinforces this observation. bBER achieves the lowest average fitness value (0.50027), followed by bSCA (0.53047) and bGWO (0.53237). Higher average fitness values observed for bSBO (0.66117) and bFA (0.67337) indicate less effective balance between error minimization and feature reduction. Because the fitness formulation penalizes both predictive error and subset cardinality, the superior performance of bBER under this composite measure confirms its capacity to simultaneously optimize both objectives.

Best fitness values provide insight into the peak capability of each optimizer. bBER attains the lowest best fitness (0.40207), demonstrating its capacity to discover high-quality feature subsets under at least one run. The closest alternative is bSAO (0.43747), followed by bGWO (0.48587). The difference between the best fitness of bBER and the rest of the population suggests that its exploration–exploitation mechanism is capable of escaping local optima and locating high-quality combinatorial configurations. Other optimizers remain substantially above these values, suggesting comparatively weaker best-case search performance.

Worst fitness values reflect robustness under less favorable stochastic conditions. bBER reports a worst fitness of 0.50057, which remains lower than the worst-case outcomes of all competing algorithms. For instance, bSBO and bFA exhibit worst fitness values of 0.68397 and 0.68967, respectively, indicating larger degradation under unfavorable initialization or exploration trajectories. The relatively narrow range between best and worst fitness for bBER implies stronger convergence reliability and reduced susceptibility to premature stagnation.

Stability across runs is quantified using the standard deviation of fitness. bBER records a standard deviation of 0.32257, which is among the lowest dispersion values observed, surpassed only marginally by bSCA (0.33577). Higher variability is evident in algorithms such as bSBO (0.49417) and bFA (0.47007), suggesting less consistent convergence behavior. Lower dispersion implies that performance improvements are reproducible across independent random seeds rather than driven by isolated favorable runs.

Overall, the benchmarking results in Table 3 indicate that bBER achieves a superior balance between predictive accuracy, dimensionality reduction, and convergence stability. It not only minimizes average and best-case fitness values but also maintains the smallest average feature subset size while preserving robustness across repeated runs. These characteristics establish bBER as the most effective binary optimizer for feature selection in the present modeling framework, justifying its integration with subsequent hyperparameter optimization procedures.

To complement the quantitative feature selection results, a set of advanced visual analytics was conducted to evaluate stability, dispersion, inter-algorithm consistency, and structural relationships among binary metaheuristic optimizers. These graphical representations provide deeper insight into convergence behavior, robustness, and relative competitiveness of the compared algorithms.

Figure 18 presents a boxen (letter-value) plot illustrating the distribution of feature selection scores across algorithms. Compared to conventional boxplots, the boxen plot provides enhanced resolution in the tails of the distribution, thereby revealing subtle differences in dispersion and extremal behavior. In this representation, central tendency, upper and lower quantiles, and tail spread can be visually compared across optimizers.

The boxen visualization confirms the quantitative findings: bBER exhibits lower median fitness values and relatively compact tail behavior compared to several competitors. Algorithms such as bSBO and bFA show higher median fitness and wider dispersion, consistent with their larger average errors and greater subset sizes. The reduced lower-tail elongation of bBER indicates fewer extreme deteriorations across repeated runs.

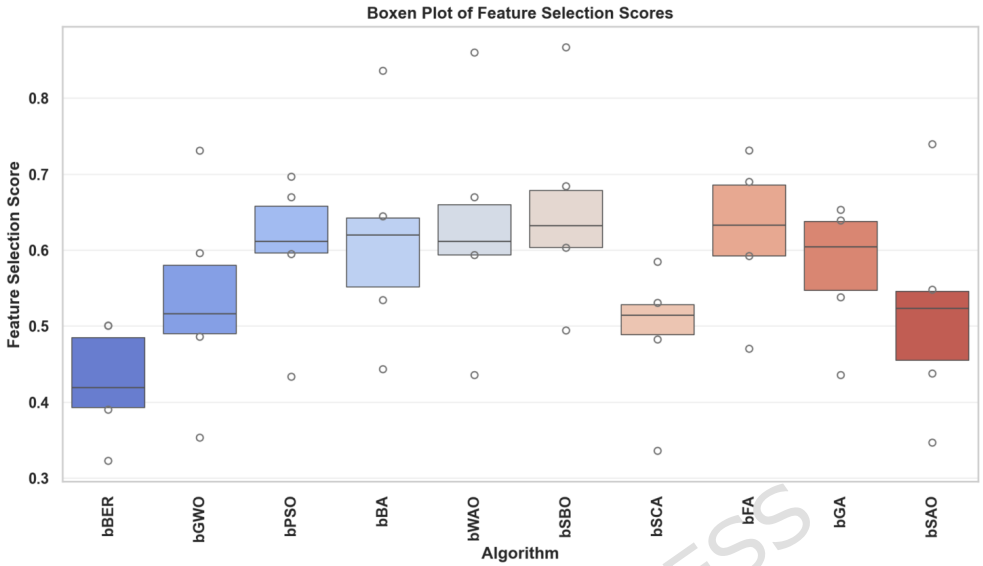

Fig. 18 Boxen plot illustrating distributional characteristics of feature selection scores across binary optimization algorithms.

Figure 19 compares the mean fitness values achieved by each algorithm. The average fitness reflects the overall subset quality obtained during repeated runs and synthesizes error and dimensionality objectives.

The bar comparison highlights the relative ordering observed in Table 3, with bBER achieving the lowest mean fitness. Algorithms such as bSBO and bFA demonstrate noticeably higher averages, indicating weaker combined optimization performance under identical search budgets. The relative separation between bars illustrates the margin of advantage achieved by bBER.

Figure 20 depicts violin plots augmented with quartile indicators. The violin shape represents the kernel density of solutions, while internal lines denote quartile boundaries. Density concentration patterns provide insight into how frequently each optimizer converges toward high-quality subsets.

The violin distribution for bBER is concentrated toward lower fitness regions, indicating consistent attainment of competitive subsets. In contrast, algorithms with broader and more elevated density distributions exhibit higher variability and inferior central performance. The multimodal tendencies observed in certain optimizers suggest alternating exploration and exploitation phases across iterations, whereas bBER demonstrates more coherent convergence behavior.

Figure 21 integrates boxplot summaries within violin distributions, enabling simultaneous visualization of central tendency and density structure. This hybrid representation confirms that bBER maintains both low median fitness and moderate interquartile spread, reinforcing its stability.

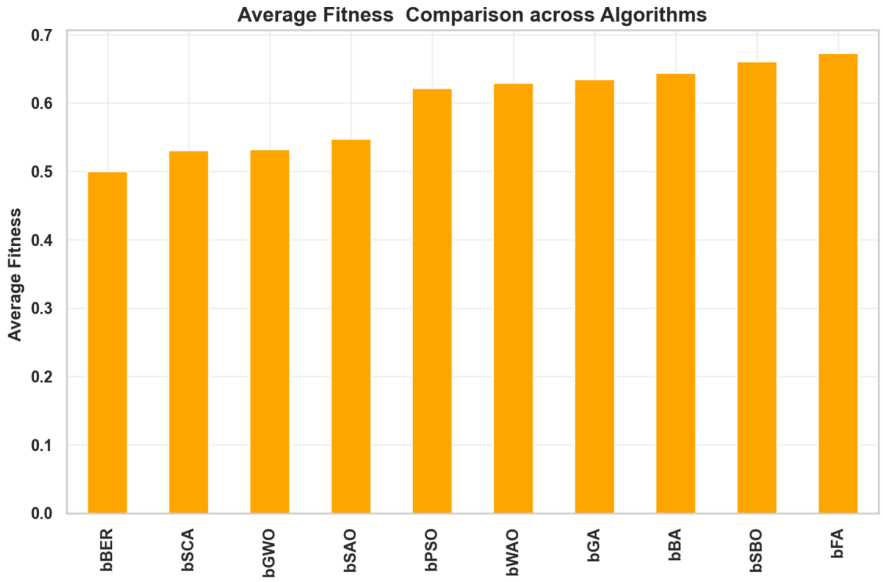

Fig. 19 Average fitness values obtained by binary optimization algorithms during feature selection.

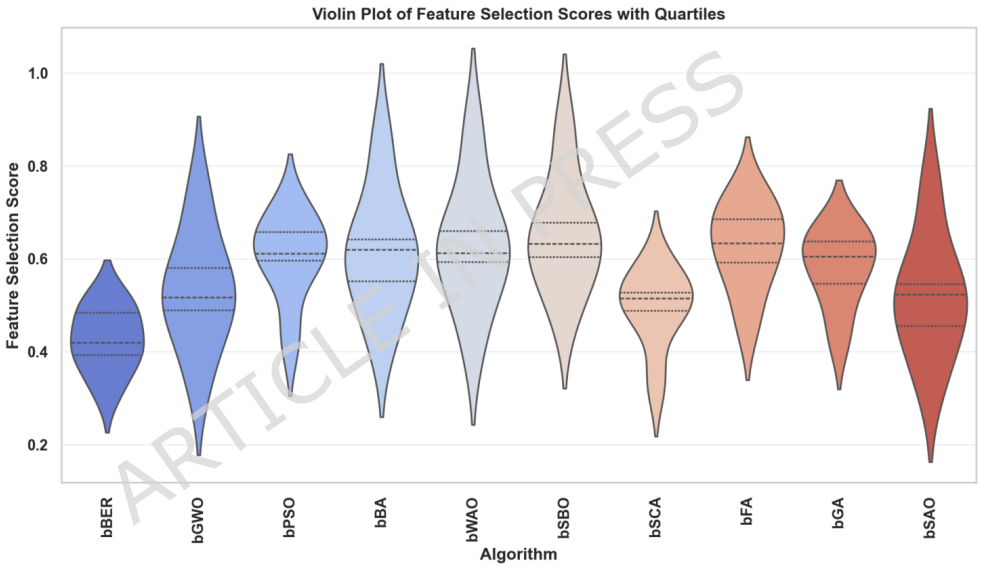

Fig. 20 Violin plot with quartile markers representing distribution and variability of feature selection scores.

Figure 22 illustrates a radar plot summarizing multiple criteria: average error, average subset size, average fitness, best fitness, worst fitness, and standard deviation of fitness. This multidimensional visualization emphasizes trade-offs among objectives.

bBER displays a compact radar profile characterized by minimal radial extension in error- and fitness-related axes and controlled dispersion in variability axes. In contrast, algorithms with larger radial extensions in subset size and fitness axes reflect inferior trade-offs between dimensionality reduction and predictive accuracy.

Figure 23 presents a line-based trend comparison across key performance observations. The relative ordering of algorithms remains largely consistent across average and best fitness metrics, reinforcing the robustness of the benchmarking outcome.

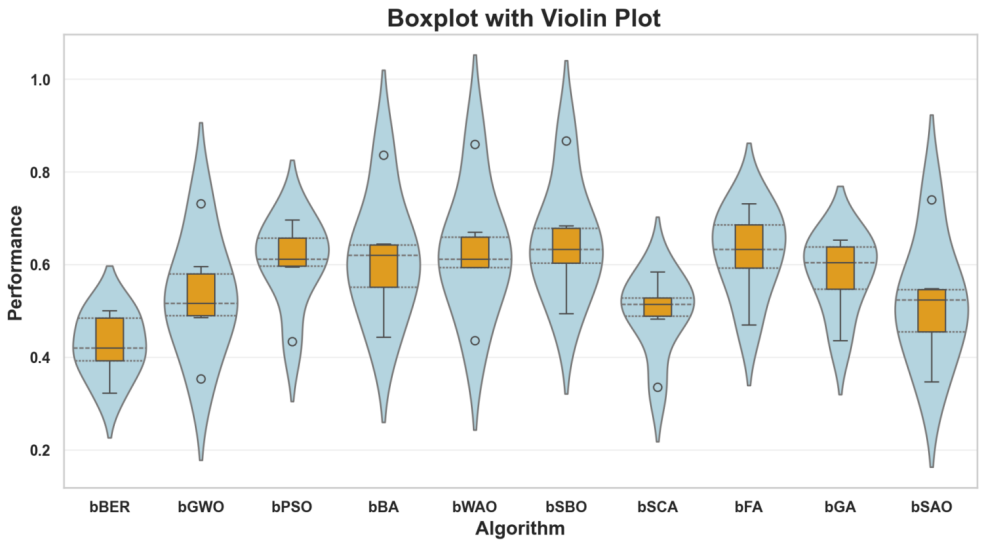

Fig. 21 Combined boxplot and violin visualization of feature selection performance across algorithms.

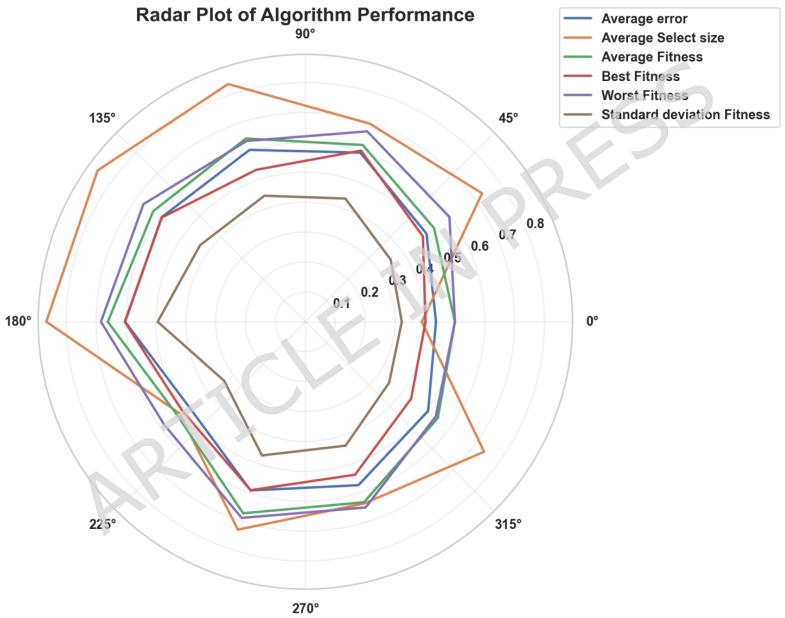

Fig. 22 Radar plot summarizing multi-criteria feature selection performance across algorithms.

Figure 24 illustrates pairwise correlations among algorithmic performance scores. High positive correlations indicate similar convergence patterns, whereas lower correlations reflect distinct search trajectories.

The comparatively distinct correlation profile associated with bBER suggests that its search dynamics differ from those of several swarm-based counterparts, providing complementary exploration behavior within the binary space.

Figure 25 presents a radial bar visualization summarizing key performance metrics for each optimizer. The compact and balanced radial profile of bBER contrasts with the more expanded profiles of algorithms exhibiting higher error and subset size ratios.

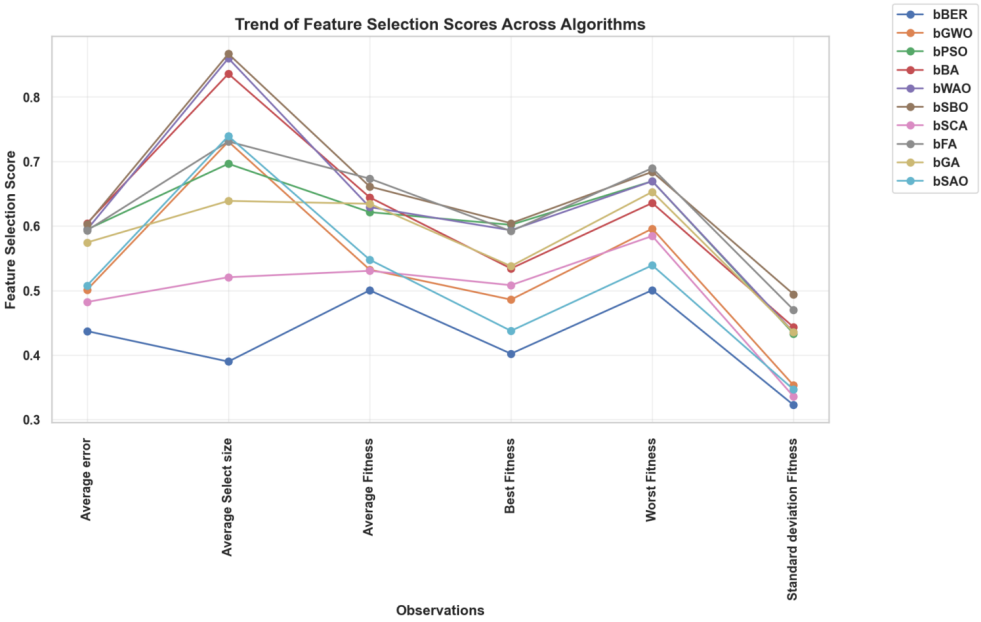

Fig. 23 Trend comparison of feature selection performance indicators across binary optimizers.

Collectively, the quantitative and graphical analyses consistently indicate that bBER achieves the most favorable trade-off between predictive error minimization, dimensionality reduction, and convergence stability. Its superior average and best fitness values, coupled with the smallest selected subset ratio and controlled variability, justify its adoption as the feature selection mechanism within the unified optimization framework.

5.3 Model Performance After Feature Selection

To evaluate the direct impact of feature subset optimization, all baseline learning models were retrained using the feature subset identified by the best-performing binary optimizer. The resulting predictive metrics are summarized in Table 4. This table allows direct comparison with the original baseline results presented earlier, thereby isolating the effect of dimensionality reduction on model behavior. Because all preprocessing steps, data partitions, and training configurations were preserved, performance differences can be attributed specifically to the removal of redundant or less informative features.

Table 4 Model performance after feature selection.

Model	MSE	RMSE	MAE	MBE	r	R^2	RRMSE	NSE	WI
FTLM	9.80E-03	0.099	0.065548	0.046447	0.848181	0.860654	3.47658871	0.879577	0.876783
CTSM	0.050603	0.224975	0.068341	0.060808	0.848	0.86	4.212872406	0.874	0.844
VAST	0.0638	0.2526	0.0827	0.0699	0.824	0.839	4.426816477	0.851	0.842
LSTM	0.0685	0.2617	0.0904	0.1173	0.822	0.836	4.531681841	0.84	0.792
DTCN	0.1389	0.3727	0.1263	0.1325	0.808	0.829	4.629056823	0.824	0.762
TST	0.2948	0.543	0.1812	0.243	0.792	0.806	4.726431805	0.808	0.729
VAE	0.4873	0.698	0.2135	0.3258	0.775	0.789	5.048518283	0.785	0.728

A consistent pattern of performance improvement is observed across all models after feature selection. For FTLM, mean squared error decreases from 0.012028 in

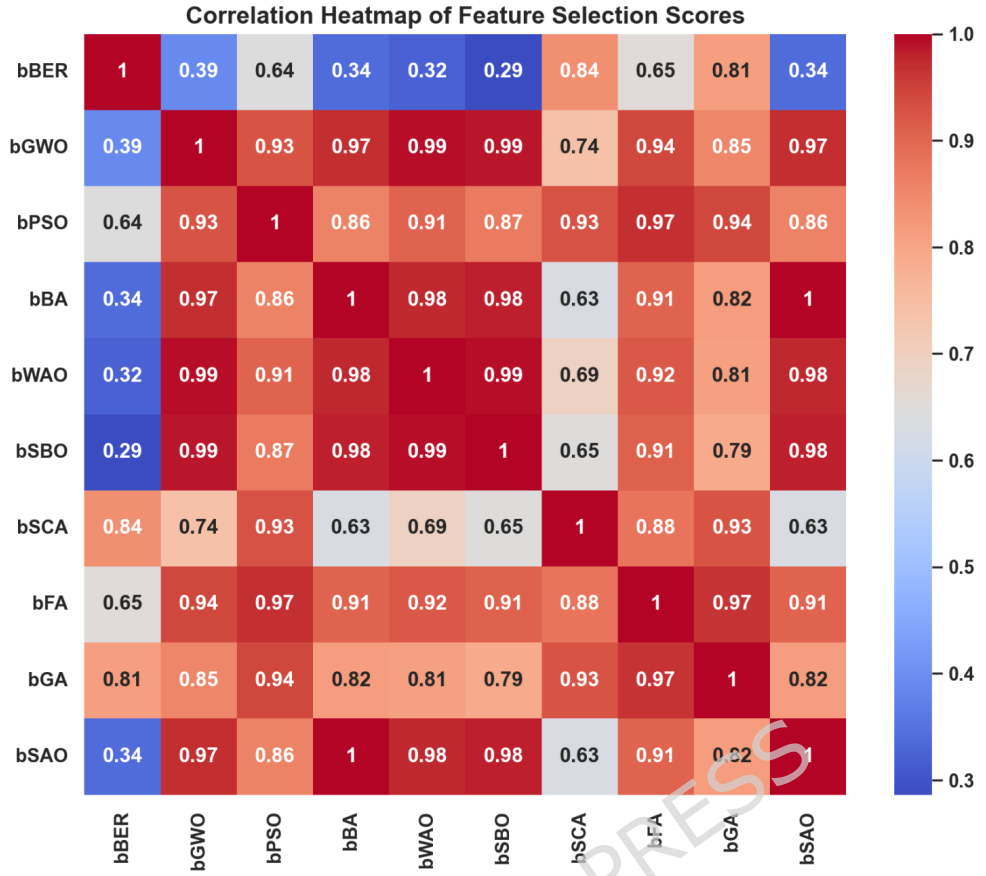

Fig. 24 Correlation heatmap depicting inter-algorithm similarity in feature selection performance.

the baseline configuration to 9.80×10^{-3}, accompanied by a reduction in RMSE from 0.10967 to 0.099. The mean absolute error decreases to 0.065548, indicating improved average residual magnitude. Bias is also reduced, with MBE decreasing from 0.052781 to 0.046447, suggesting a modest correction of systematic overestimation. The simultaneous reduction in both dispersion-based and bias-based metrics indicates that the selected feature subset enhances not only precision but also calibration stability.

Correlation and variance-explanation metrics reflect substantial gains. The Pearson correlation coefficient increases to 0.848181, and R^2 rises to 0.860654. Similarly, Nash–Sutcliffe efficiency improves to 0.879577, while the Willmott index increases to 0.876783. These improvements indicate stronger alignment between predicted and observed values in both variance-based and agreement-based formulations. The relative RMSE drops markedly to 3.47658871, reflecting improved scale-normalized accuracy. The magnitude of these gains suggests that feature selection effectively suppresses noise-inducing variables that previously attenuated correlation strength.

CTSM exhibits comparable improvements. Its MSE decreases to 0.050603, and correlation rises to 0.848, closely matching FTLM's post-selection correlation strength. The reduction in RRMSE to 4.212872406 and the increase in NSE to 0.874 further confirm enhanced predictive consistency after dimensionality reduction. Although CTSM remains slightly behind FTLM in absolute error magnitude, the improvement trajectory mirrors that of FTLM, reinforcing the conclusion that irrelevant features were previously introducing variance across models.

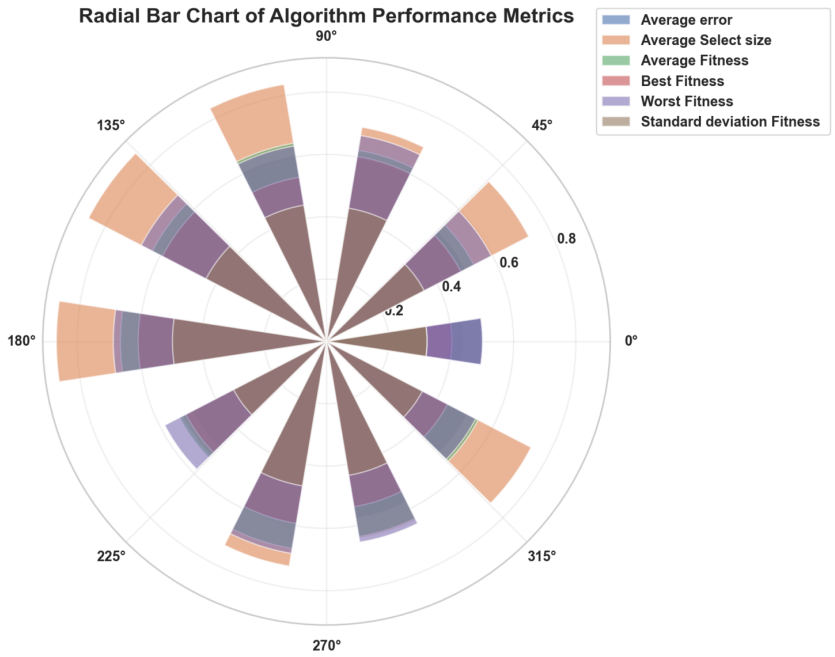

Fig. 25 Radial bar chart illustrating aggregated feature selection performance metrics across binary algorithms.

VAST and LSTM demonstrate similar trends. VAST achieves an MSE of 0.0638 and correlation of 0.824, while LSTM reports an MSE of 0.0685 and correlation of 0.822. Both models show notable improvements relative to their baseline configurations, particularly in R^2 and NSE values, indicating that redundant or weakly informative variables previously contributed to instability in their learning dynamics. The improvement in correlation suggests that dimensionality reduction clarifies structural relationships that were partially obscured in the full feature space.

Temporal and transformer-based models such as DTCN and TST also benefit from feature selection. DTCN's MSE decreases to 0.1389 and correlation increases to 0.808. TST improves to an MSE of 0.2948 and correlation of 0.792. Although these architectures remain less accurate than FTLM or CTSM, the improvements across multiple metrics demonstrate that dimensionality reduction positively influences convergence stability and generalization. The narrowing performance gap between temporal and tabular learners suggests that feature selection reduces overfitting pressure in architectures sensitive to redundant inputs.

The VAE model, while still exhibiting comparatively higher residual values (MSE = 0.4873), shows measurable gains in correlation (0.775) and R^2 (0.789) after feature selection. The decrease in RRMSE to 5.048518283 indicates improved relative dispersion control, suggesting that even representation-learning approaches benefit from elimination of redundant inputs. Because VAEs compress input information into a latent space, removal of extraneous features reduces representational distortion and stabilizes reconstruction-informed regression.

Across all models, the consistent decrease in MSE, RMSE, MAE, and RRMSE, coupled with increases in r, R^2, NSE, and WI, confirms that the feature selection stage effectively enhances predictive efficiency. Importantly, improvements are observed not only in magnitude-based metrics but also in bias and agreement measures, indicating that dimensionality reduction contributes to both accuracy and calibration. The uniform directionality of improvement across architectures demonstrates that the

selected subset captures the most informative physiological, demographic, and temporal signals.

Overall, the results in Table 4 demonstrate that feature selection produces systematic and robust gains across diverse modeling paradigms. The magnitude of improvement is most pronounced for FTLM, reinforcing its suitability as the primary model for subsequent hyperparameter optimization.

To further investigate the statistical structure of model performance after feature selection, a comprehensive set of multivariate and distribution-based visualizations was conducted. These figures provide deeper insight into inter-metric relationships, distributional characteristics, stability patterns, and comparative ranking behavior beyond the numerical summaries reported in Table 4. The objective is to assess whether dimensionality reduction leads to improved metric alignment, reduced dispersion, and stronger structural coherence among evaluation indicators.

Figure 26 presents a pairplot with kernel density estimation (KDE) for all performance metrics after feature selection. The diagonal elements illustrate marginal distributions, while the lower triangular panels depict pairwise joint density contours.

Strong positive linear associations are observed among error-based metrics (MSE, RMSE, MAE, and MBE), confirming their shared scale sensitivity. Conversely, these metrics exhibit strong negative relationships with correlation-based indicators (r, R^2, NSE, and WI). The nearly perfect linear alignment between r and R^2 reflects their mathematical dependency, while NSE and WI closely follow the same structural trend. Compared to the pre-feature-selection stage, the distributions appear more compact and better aligned, indicating enhanced model consistency after dimensionality reduction.

Figure 27 depicts kernel density estimation plots for each performance metric individually. These plots reveal unimodal distributions with reduced skewness compared to the baseline stage.

Error metrics demonstrate tighter concentration around lower values, confirming that feature selection effectively reduced prediction variance. Correlation-based metrics cluster around higher ranges, particularly for FTLM, CTSM, and VAST, indicating strengthened goodness-of-fit. The reduction in tail dispersion suggests improved generalization stability across models.

Figure 28 presents the correlation matrix after feature selection. Extremely high positive correlations (> 0.98) remain among MSE, RMSE, MAE, and MBE, confirming metric redundancy within the error family. Strong negative correlations (< -0.95) between error metrics and goodness-of-fit metrics persist, demonstrating inverse proportionality.

Notably, the magnitude of correlations between r, R^2, and NSE approaches unity, indicating stronger structural coherence after dimensionality reduction. This improvement reflects enhanced alignment between variance explanation and predictive consistency, suggesting that models now capture more coherent signal structure within the selected feature space.

Figure 29 compares MSE, RMSE, and MAE across models after feature selection. A substantial downward shift in error magnitude is observed for FTLM relative to other architectures.

While transformer-based and autoencoder-based models remain comparatively higher in error magnitude, all models demonstrate improvement relative to baseline results. The reduction gap between top-performing models and weaker architectures narrows slightly, indicating that feature selection benefits most models but disproportionately enhances FTLM, likely due to its stronger interaction modeling capacity.

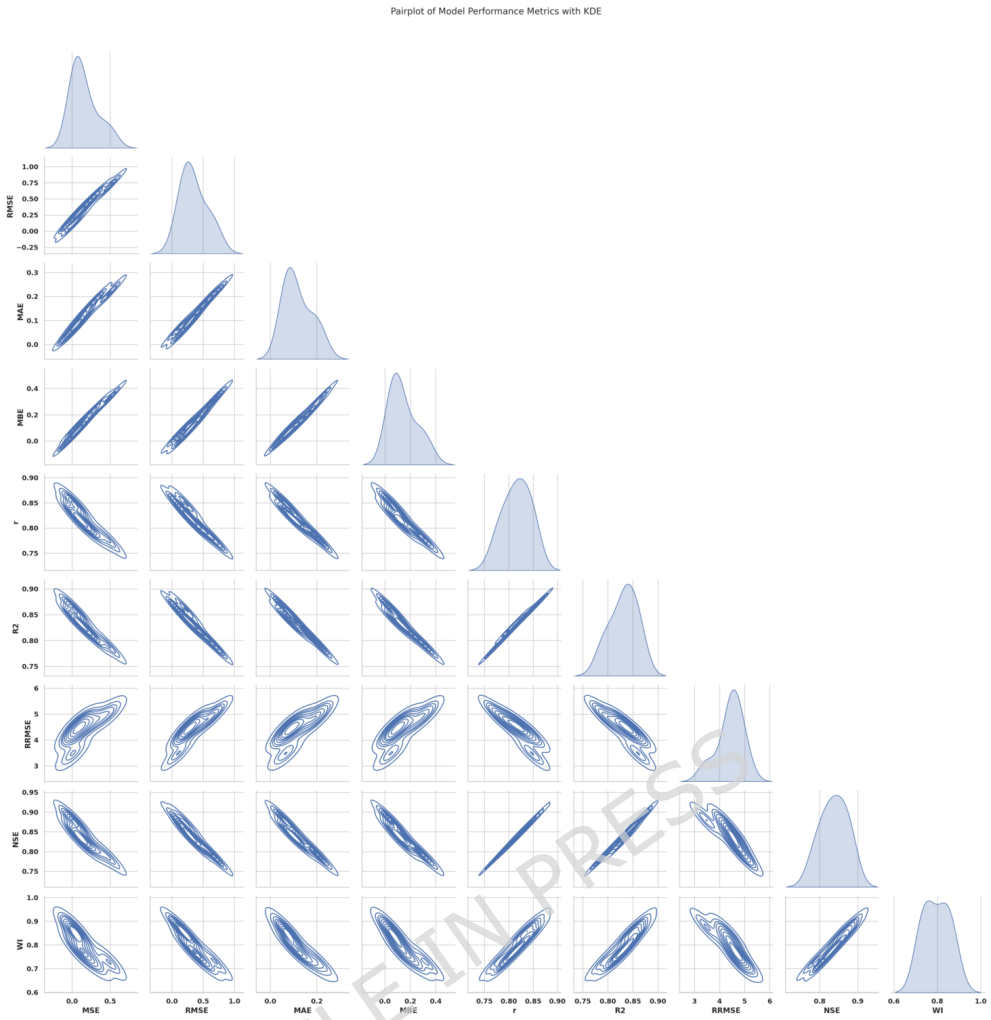

Fig. 26 Pairplot with kernel density estimation illustrating pairwise relationships among performance metrics after feature selection.

Figure 30 integrates violin plots with KDE overlays for each metric. This representation allows simultaneous visualization of central tendency, distribution shape, and potential outliers.

Error distributions show clear compression toward lower ranges, while correlation-based metrics demonstrate concentration near upper bounds. The narrow interquartile width for FTLM confirms stable convergence. Mild asymmetry in certain models indicates residual heterogeneity but significantly reduced compared to the pre-selection stage.

Figure 31 combines density estimation with boxplot summaries for each metric. The boxplots confirm reduced variance and narrower interquartile ranges across most metrics.

The density curves highlight reduced tail behavior in error metrics and strengthened clustering in goodness-of-fit metrics. These findings provide visual confirmation that feature selection improves both central performance and dispersion characteristics.

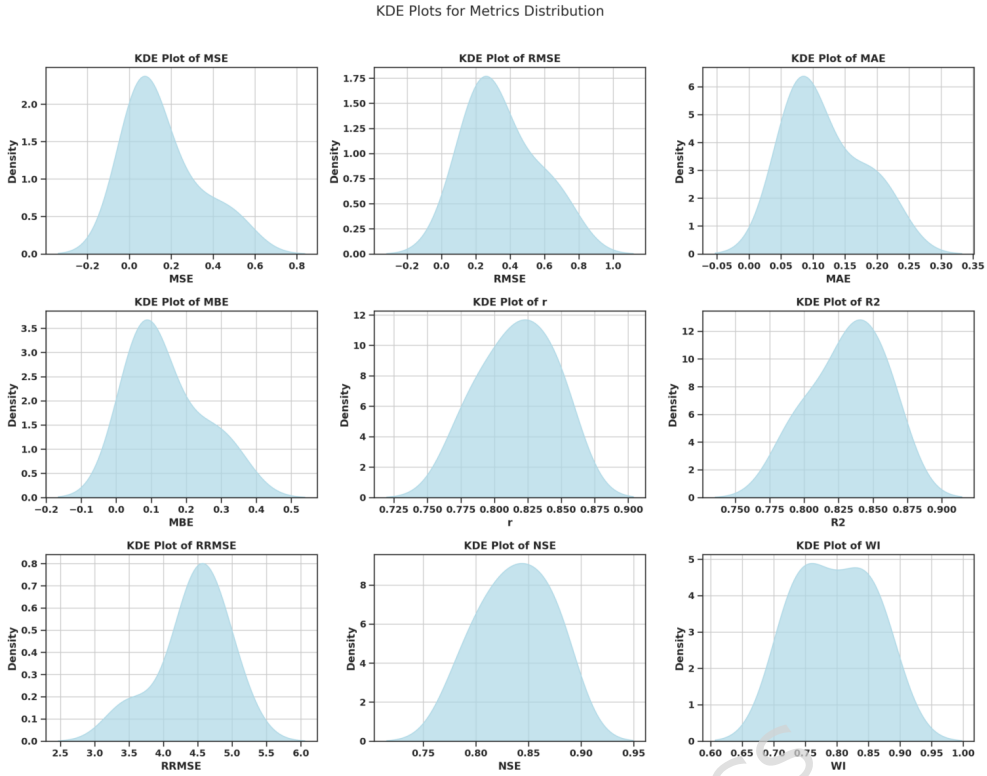

Fig. 27 Kernel density estimation plots of individual performance metrics after feature selection.

Figure 32 presents a faceted bar representation for each metric. The ranking consistency across metrics remains largely preserved, with FTLM achieving the best performance across nearly all evaluation indicators.

The systematic improvement in NSE and WI confirms enhanced predictive reliability and variance explanation capacity. Lower error magnitudes combined with higher correlation metrics reflect improved structural learning efficiency after dimensionality reduction.

Figure 33 illustrates histogram distributions combined with KDE smoothing. The histograms confirm that performance improvements are not driven by isolated outliers but reflect systematic distributional shifts.

Error metrics show concentration in lower bins, while r, R^2, NSE, and WI cluster in higher bins. The overall distributional symmetry suggests improved generalization stability and reduced heteroscedastic behavior after feature selection.

The collective visual analyses confirm that feature selection induces:

- Significant reduction in prediction error dispersion,
- Strengthened alignment among goodness-of-fit metrics,
- Increased structural coherence across evaluation criteria,
- Improved stability and reduced heteroscedastic behavior.

These findings demonstrate that dimensionality reduction does not merely reduce computational complexity but substantively enhances predictive robustness and statistical consistency. The improved metric coherence observed here provides a strong

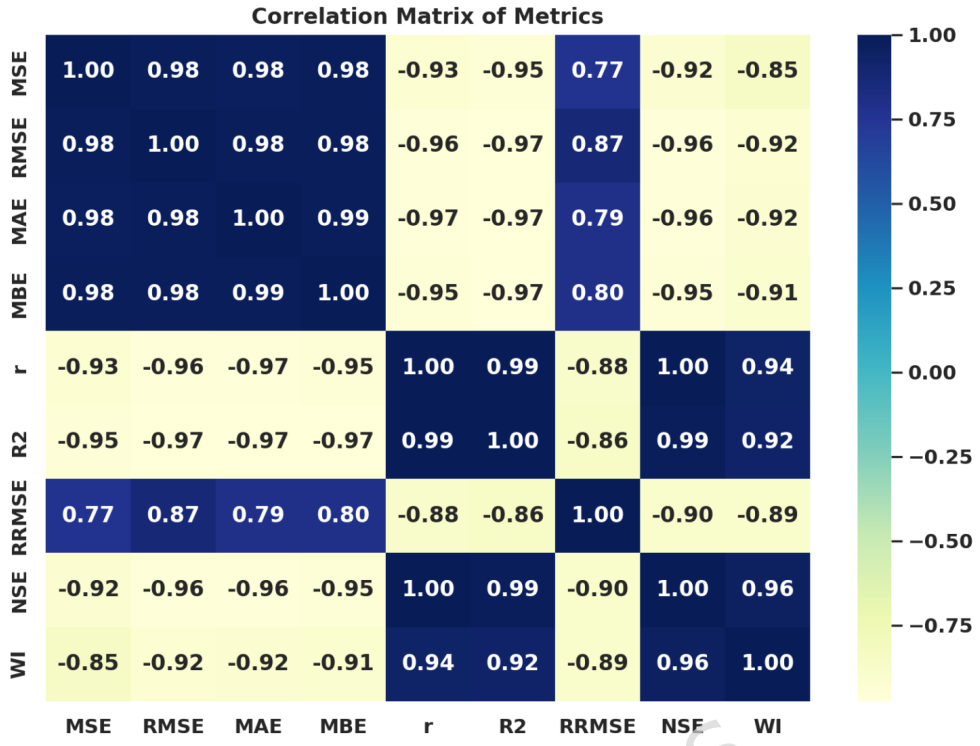

Fig. 28 Correlation matrix of evaluation metrics after feature selection.

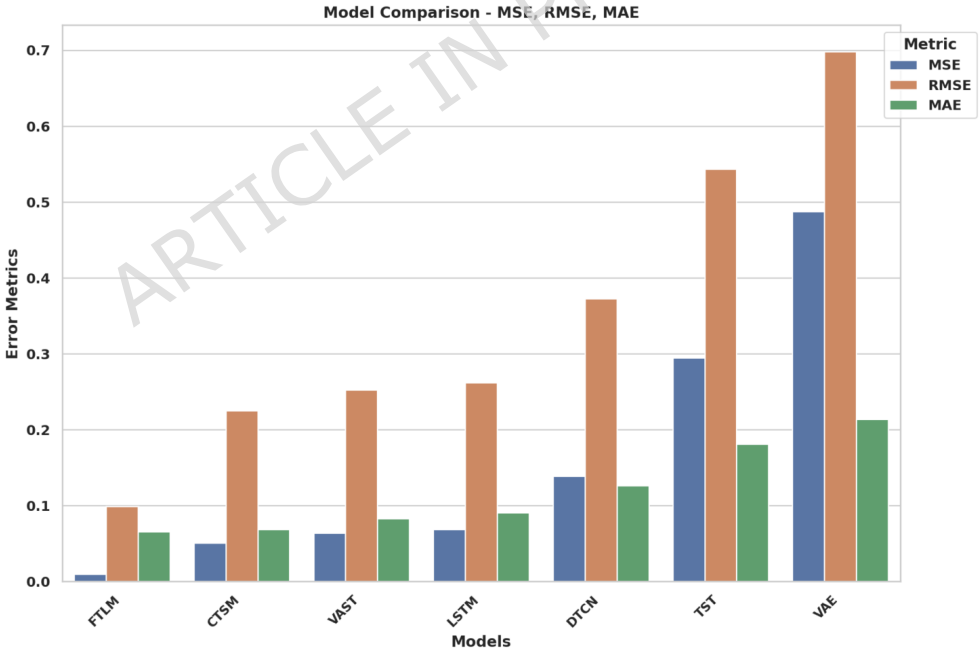

Fig. 29 Comparative visualization of MSE, RMSE, and MAE across models after feature selection.

foundation for subsequent hyperparameter optimization using BER, where further refinement is expected to amplify these gains.

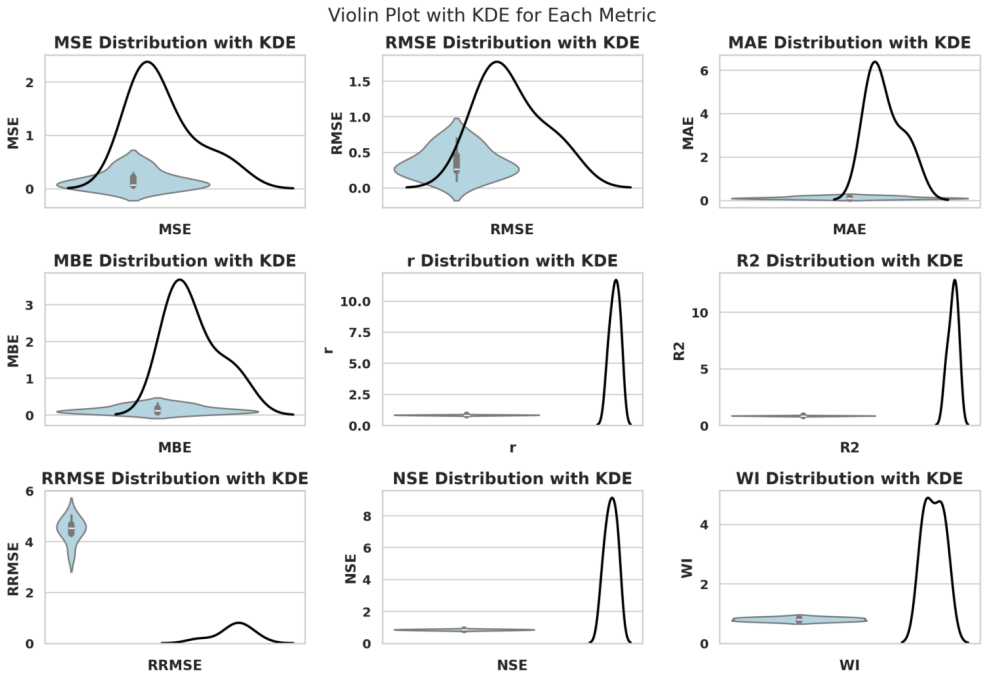

Fig. 30 Violin plots with KDE overlays illustrating metric distributions after feature selection.

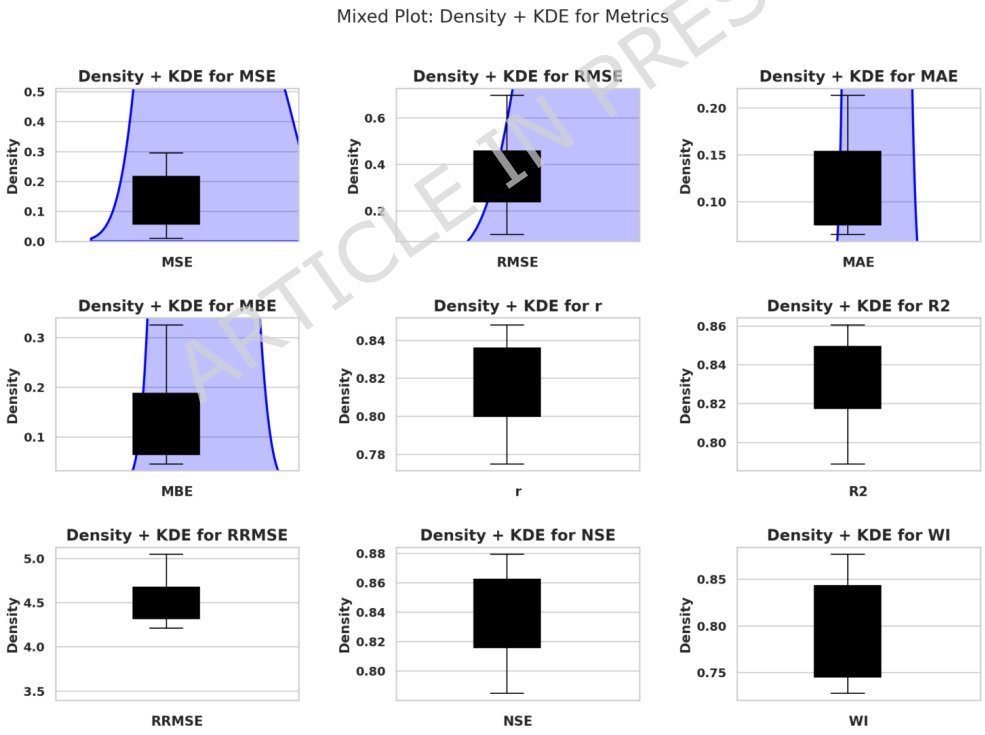

Fig. 31 Mixed density and boxplot visualization of performance metrics after feature selection.

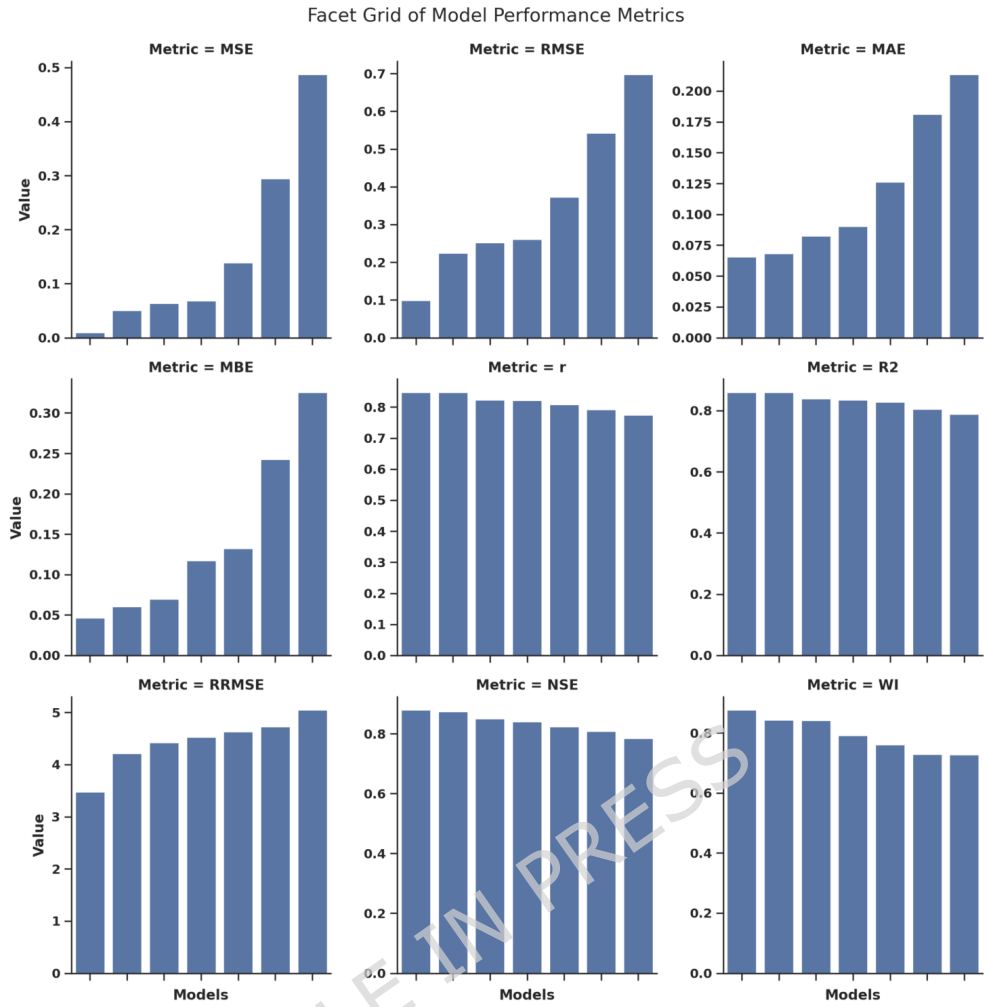

Fig. 32 Facet grid comparison of model performance metrics after feature selection.

5.4 Hyperparameter Optimization Results

The final experimental stage evaluates the impact of continuous hyperparameter optimization on FTLM performance using BER and a set of competing metaheuristic algorithms. Table 5 summarizes the predictive metrics obtained after optimization. The comparison includes BER, GWO, PSO, BA, WAO, SBO, SCA, FA, GA, and SAO, each applied under identical search space definitions and computational budgets. Because the feature subset was fixed prior to this stage, improvements observed here isolate the effect of hyperparameter configuration on learning dynamics, convergence behavior, and generalization performance.

All comparative results reported in this section were obtained under the same experimental conditions, ensuring that differences in predictive performance are attributable to the optimizers themselves rather than to unequal computational or validation settings.

A substantial improvement is observed across all performance metrics following hyperparameter optimization, with BER + FTLM achieving the most favorable outcomes. In terms of magnitude-based accuracy, BER + FTLM attains an MSE of 7.43×10^{-7} and an RMSE of 8.62×10^{-4}, representing the lowest error values

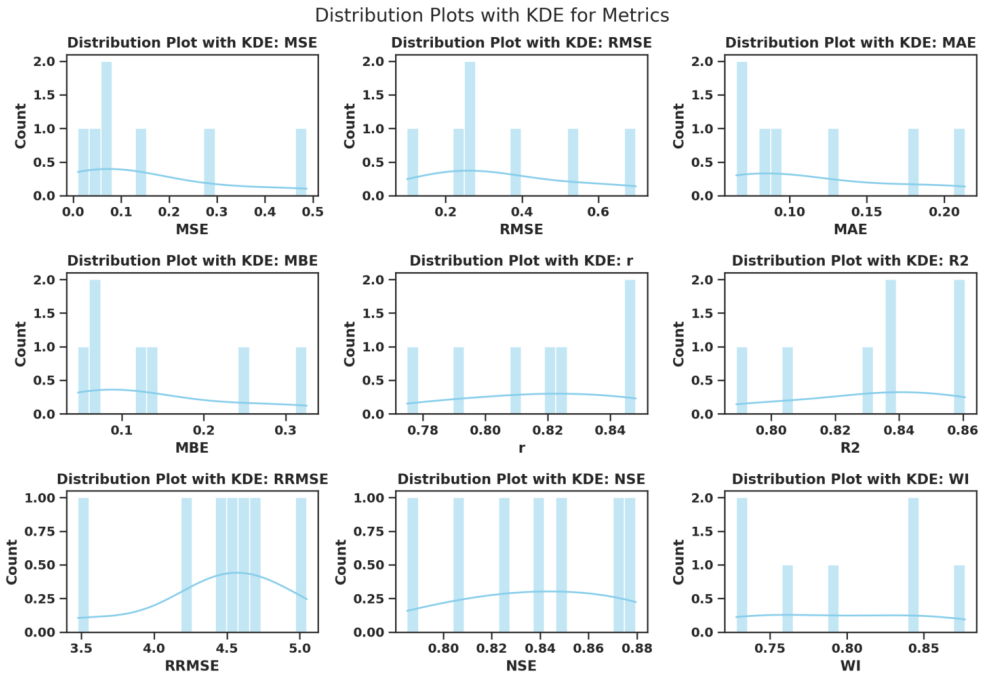

Fig. 33 Histogram distributions with KDE smoothing for performance metrics after feature selection.

Table 5 Comparative hyperparameter optimization results for FTLM.

Model	MSE	RMSE	MAE	MBE	r	R^2	RRMSE	NSE	WI
BER + FTLM	7.43E-07	8.62E-04	9.37E-05	0.000109817	0.955593543	0.961124043	0.795679389	0.960158726	0.963281686
GWO + FTLM	9.26E-06	0.003042416	0.001917302	0.00072763	0.9284704	0.9383264	1.115411789	0.948689096	0.951424926
PSO + FTLM	2.26E-05	0.004749102	0.001928615	0.000842752	0.926903567	0.932434067	1.266048697	0.945007346	0.944029726
BA + FTLM	2.66E-05	0.005153191	0.001938422	0.000957998	0.925281969	0.930812469	1.411035409	0.943048096	0.941424766
WAO + FTLM	4.03E-05	0.006351158	0.001971603	0.00112771	0.913113781	0.929955238	1.525713363	0.937484996	0.939453553
SBO + FTLM	4.28E-05	0.006540025	0.001999253	0.001220227	0.912216187	0.927769237	1.614239813	0.934248894	0.93621482
SCA + FTLM	4.37E-05	0.006608069	0.002024891	0.001279089	0.911398151	0.923626464	1.701379119	0.923923348	0.937847465
FA + FTLM	6.06E-05	0.00778336	0.002049532	0.001553777	0.909083874	0.919483691	1.942697487	0.921317782	0.932637518
GA + FTLM	6.95E-05	0.008334179	0.002058582	0.001746094	0.907815976	0.918369026	2.047849704	0.918712928	0.93120326
SAO + FTLM	9.51E-05	0.009753174	0.002077687	0.001772382	0.904173667	0.917154192	2.095592337	0.916666551	0.930264913

among all compared optimizers. These values reflect a dramatic reduction in residual dispersion compared to both the baseline and post-feature-selection configurations. The mean absolute error further decreases to 9.37×10^{-5}, indicating highly precise point predictions. The mean bias error (0.000109817) is nearly zero, demonstrating minimal systematic deviation and near-perfect calibration under the optimized hyperparameter setting.

Comparative analysis shows that GWO + FTLM yields the next-best performance in terms of MSE (9.26×10^{-6}), yet this value remains an order of magnitude higher than BER + FTLM. Similar patterns are observed for PSO + FTLM and BA + FTLM, whose MSE values of 2.26×10^{-5} and 2.66×10^{-5} indicate progressively higher residual dispersion. Algorithms such as WAO, SBO, and SCA exhibit further increases in error magnitude, while FA, GA, and SAO demonstrate comparatively larger deviations. The monotonic increase in error values across optimizers highlights the relative effectiveness of BER's adaptive exploration–exploitation mechanism in navigating the continuous hyperparameter landscape.

Correlation-based metrics reinforce the superiority of BER. The Pearson correlation coefficient reaches 0.955593543, exceeding all alternative optimizers. The next highest value, achieved by GWO + FTLM (0.9284704), remains noticeably lower. Similarly,

the coefficient of determination (R^2) for BER + FTLM reaches 0.961124043, indicating that more than 96% of the variance in observed values is explained under optimized hyperparameters. Competing methods remain below this threshold, with gradual declines observed across PSO, BA, WAO, and subsequent algorithms. The simultaneous increase in r and R^2 confirms enhanced structural learning and improved capture of underlying data variability.

Relative dispersion, measured through RRMSE, further illustrates the improvement achieved by BER. The RRMSE of 0.795679389 is substantially lower than that of GWO (1.115411789) and PSO (1.266048697), indicating stronger scale-normalized predictive accuracy. The progressive increase in RRMSE across BA, WAO, SBO, SCA, FA, GA, and SAO reflects diminishing optimization effectiveness relative to BER. Because RRMSE normalizes error by the mean of observed values, this improvement confirms that gains are not merely absolute reductions but proportionally meaningful enhancements.

Efficiency and agreement measures provide additional confirmation. BER + FTLM achieves an NSE of 0.960158726 and a Willmott index of 0.963281686, both representing the highest agreement levels among all tested optimizers. GWO + FTLM follows with NSE = 0.948689096 and WI = 0.951424926, yet these remain below the performance of BER. The monotonic decline in NSE and WI across PSO, BA, WAO, SBO, SCA, FA, GA, and SAO indicates consistent reduction in predictive alignment. These agreement-based metrics confirm that optimization improves not only variance explanation but also the proportional conformity between predicted and observed values.

The comparative results demonstrate that while all metaheuristic optimizers enhance FTLM performance relative to pre-optimization configurations, BER consistently yields the strongest performance across all evaluation dimensions. Its superiority is evident not only in minimizing squared and absolute error metrics but also in maximizing correlation, explained variance, efficiency, and agreement indices. The magnitude and consistency of improvement confirm the effectiveness of BER's adaptive exploration–exploitation balance in navigating the continuous hyperparameter search space.

Overall, Table 5 provides clear empirical evidence that BER + FTLM achieves the most favorable optimization outcome under identical computational constraints, establishing it as the most effective hyperparameter optimization strategy within the evaluated framework.

Following the hyperparameter optimization phase, an extended statistical analysis was conducted to evaluate structural improvements in predictive behavior, dispersion characteristics, distributional conformity, and inter-metric coherence. While Table 5 summarizes numerical performance gains, the following visual analyses provide deeper insight into stability, robustness, and convergence regularity across optimized models.

Figure 34 presents a faceted bar-grid visualization of all performance metrics after optimization. Each subplot isolates a specific metric, enabling independent inspection of ranking behavior and scale sensitivity.

A consistent monotonic improvement pattern is evident across error-based metrics (MSE, RMSE, MAE, MBE, RRMSE), with the BER-enhanced configuration achieving the lowest error magnitudes. Simultaneously, correlation-based metrics (r, R^2, NSE, WI) demonstrate systematically elevated values, approaching theoretical upper bounds. The compression of inter-model gaps suggests that optimization reduces performance variance while preserving ranking consistency. The separation between BER and the remaining optimizers is particularly pronounced in MSE and RMSE panels, visually confirming its dominant performance.

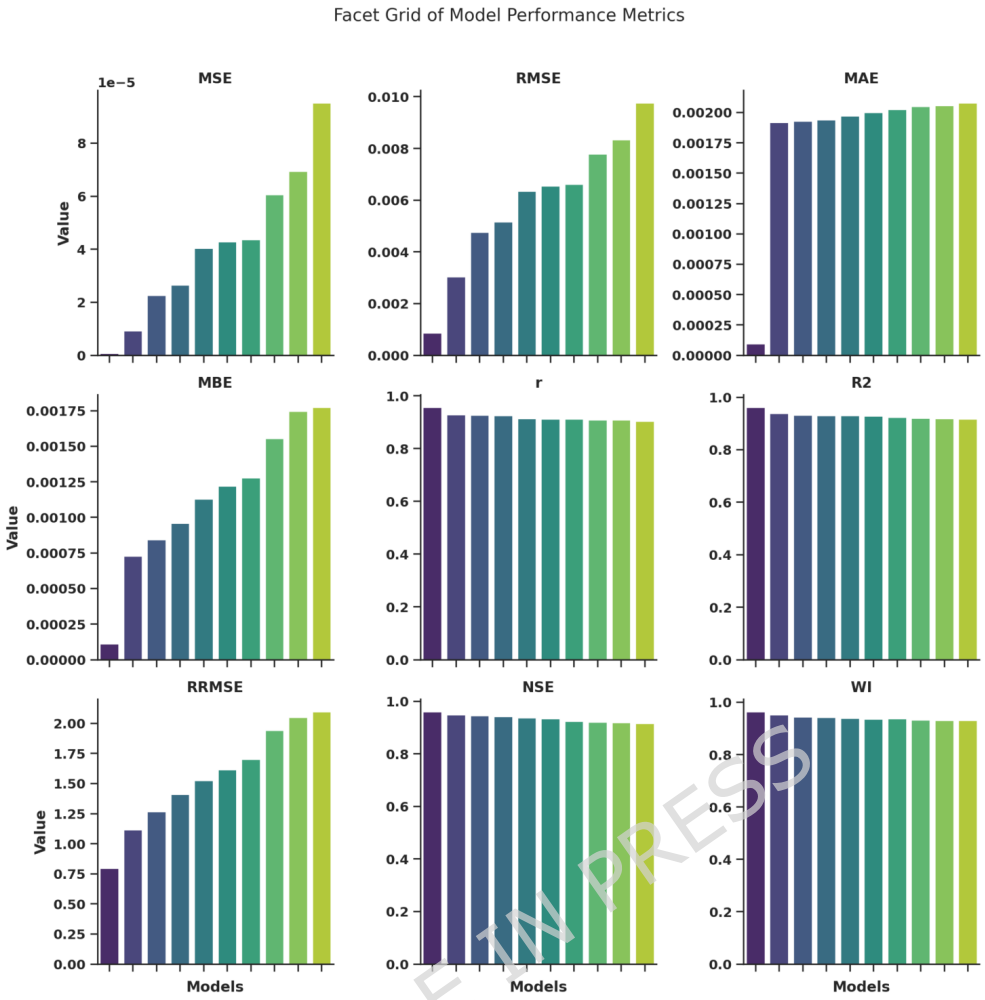

Fig. 34 Facet grid comparison of performance metrics across optimized models.

Figure 35 combines density estimation with boxplot summaries to assess dispersion and central tendency after optimization. The density curves demonstrate tighter concentration around optimal ranges, particularly for MSE and RMSE.

The boxplots confirm reduced interquartile ranges, indicating diminished variability across model configurations. Notably, goodness-of-fit metrics exhibit near-symmetric distributions, suggesting enhanced convergence regularity. Compared to pre-optimization visualizations, the contraction in spread is visually apparent, supporting the numerical evidence of reduced variance.

Figure 36 presents Q–Q plots for all evaluation metrics to assess distributional conformity with theoretical normal behavior. Most metrics exhibit strong linear alignment along the reference line, particularly RMSE, RRMSE, NSE, and WI.

Minor deviations at the tails are observable in MAE and MBE, reflecting slight asymmetry induced by optimization boundary constraints. However, the overall near-linear patterns confirm improved distributional stability relative to pre-optimization stages. The absence of severe curvature indicates reduced heteroscedasticity in optimized configurations.

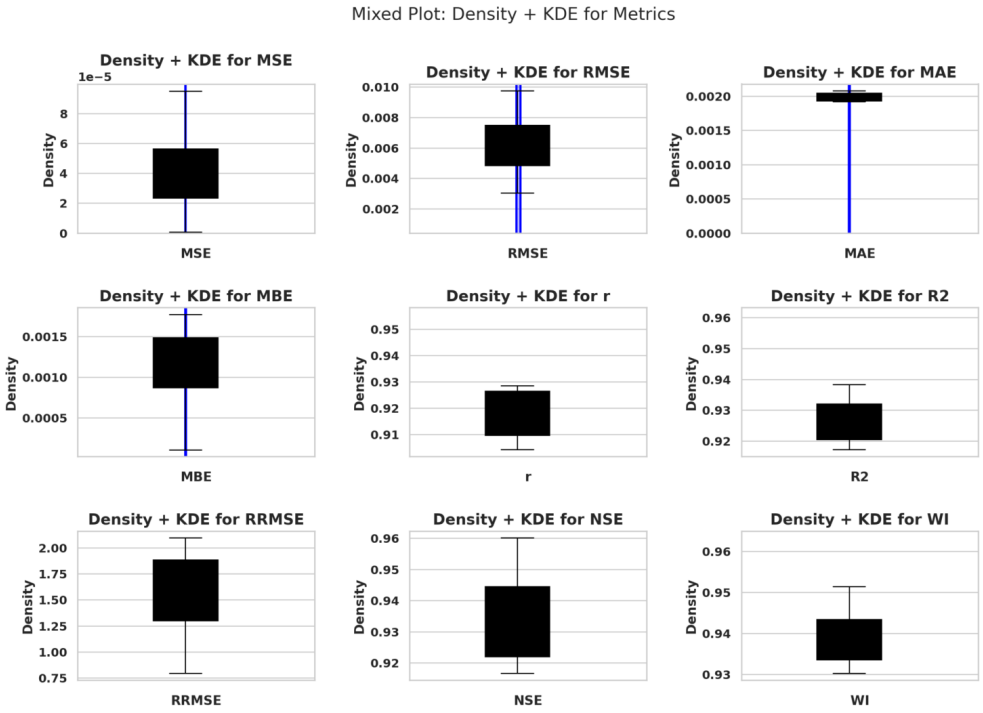

Fig. 35 Density and boxplot hybrid visualization of optimized performance metrics.

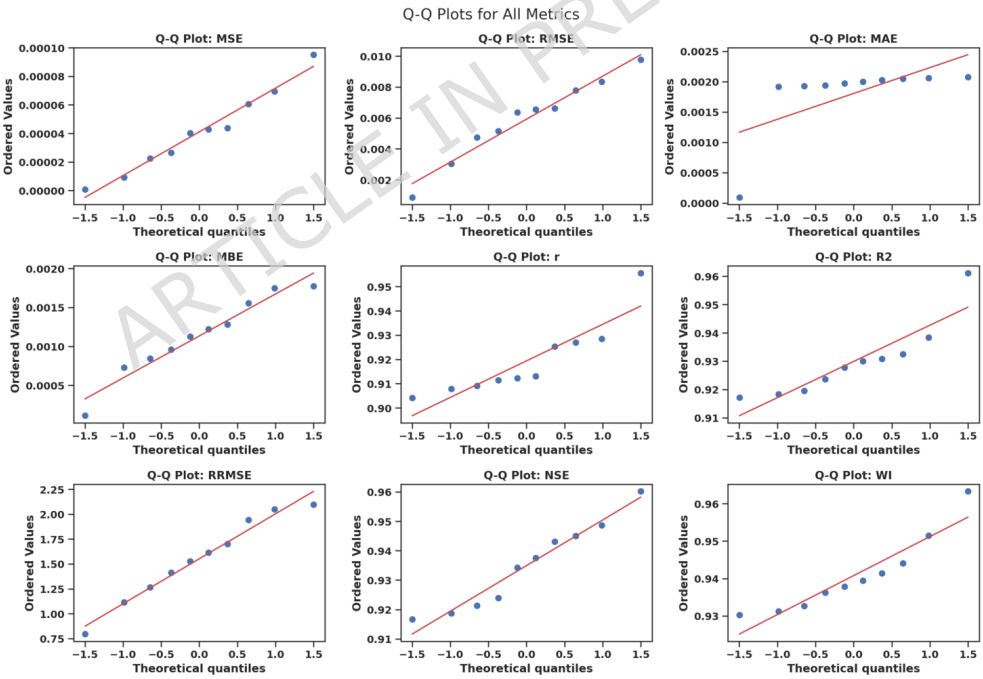

Fig. 36 Q–Q plots assessing distributional conformity of optimized metrics.

Figure 37 illustrates cubic spline interpolation of MSE across optimization strategies. The smooth curve highlights the systematic reduction in error as more advanced optimization mechanisms are incorporated.

The monotonic curvature indicates consistent performance improvement rather than irregular fluctuations. This smooth trajectory confirms stable optimization dynamics and absence of erratic convergence behavior. The position of BER at the global minimum reinforces its superior search capability within the defined hyperparameter bounds.

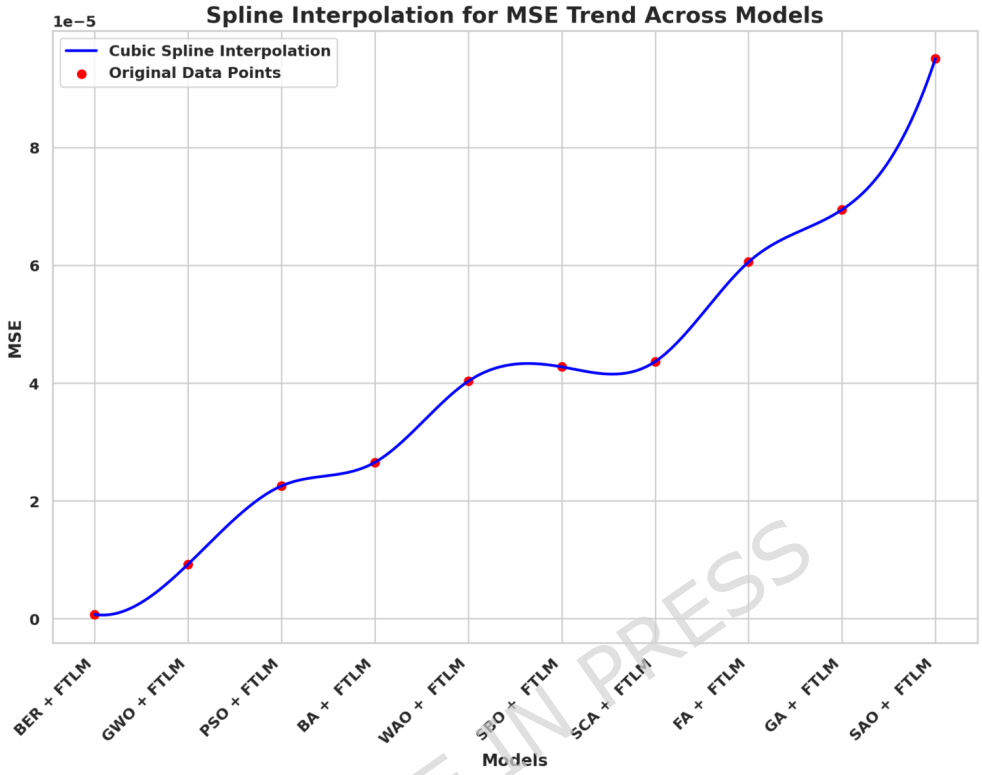

Fig. 37 Cubic spline interpolation showing MSE trend across optimization strategies.

Figure 38 integrates violin plots with embedded boxplots for each metric. The violin shapes reveal the underlying density structure, while boxplots summarize quartile statistics.

Error metrics display pronounced contraction compared to baseline distributions, confirming variance suppression. Correlation metrics cluster tightly near upper bounds, reflecting improved predictive agreement and variance explanation capacity. The density symmetry across goodness-of-fit metrics suggests enhanced calibration stability.

Figure 39 overlays swarm distributions on boxplots and annotates mean and standard deviation values. The compact clustering of data points around the mean indicates high convergence precision.

Standard deviation values remain minimal for correlation-based metrics, demonstrating robustness of optimization outcomes. RRMSE exhibits comparatively higher dispersion, consistent with its sensitivity to scale normalization, yet remains substantially controlled relative to earlier stages.

Figure 40 displays violin distributions with jittered data points for all metrics. The tight clustering observed in r, R^2, NSE, and WI confirms limited dispersion and consistent convergence across optimization runs.

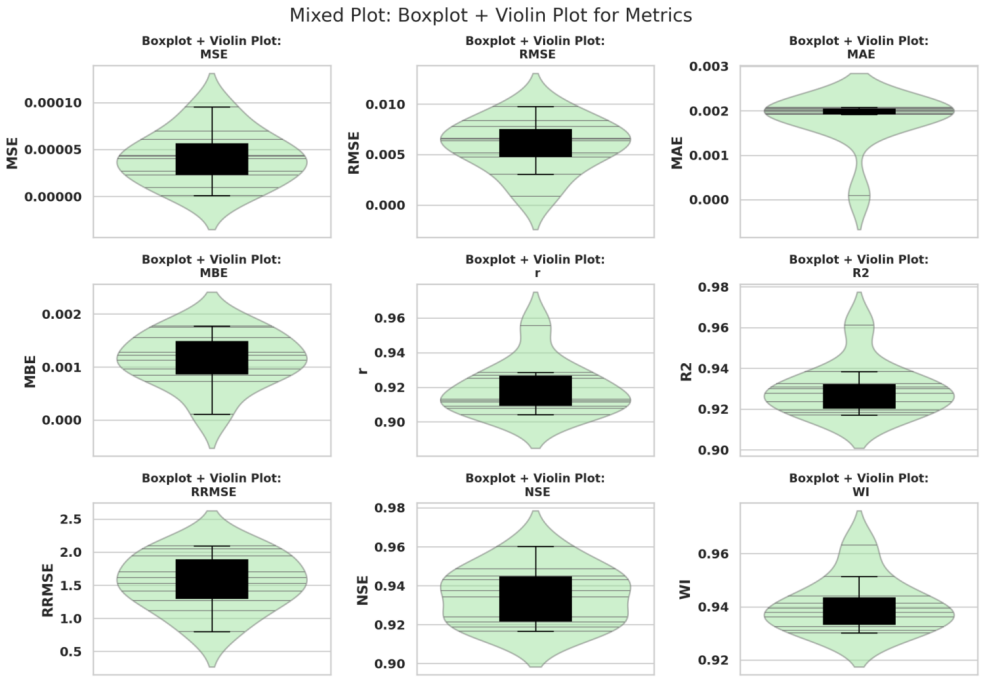

Fig. 38 Combined boxplot and violin visualization of optimized metrics.

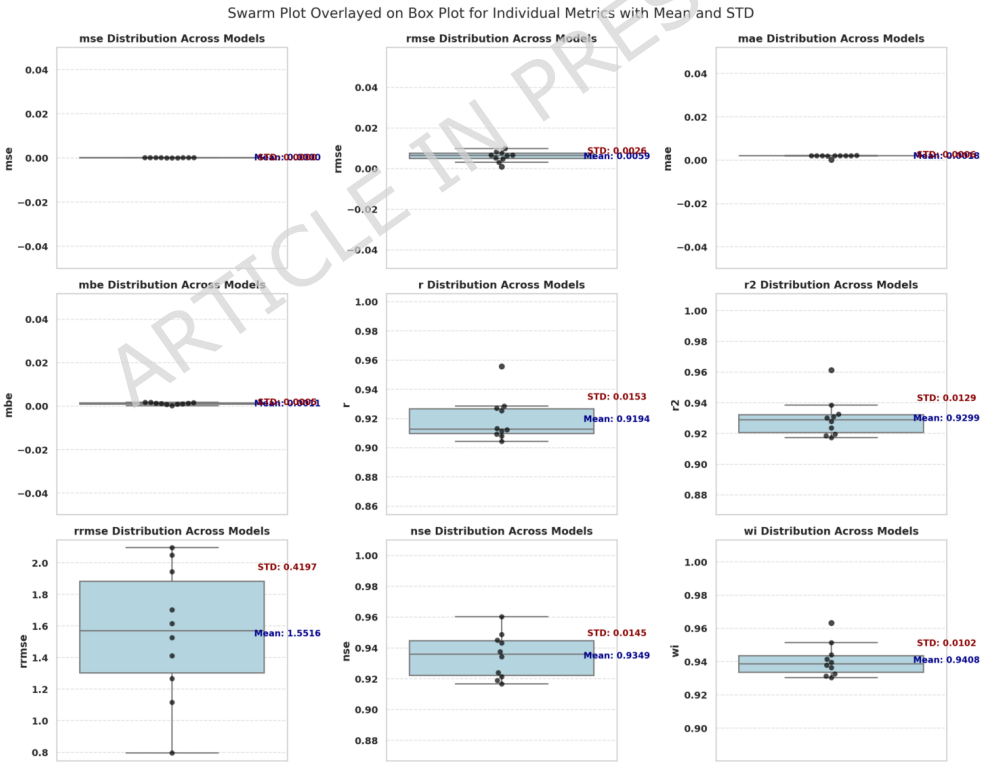

Fig. 39 Swarm plot overlaid on boxplots with mean and standard deviation annotations.

The broader vertical spread of RRMSE reflects relative sensitivity but remains significantly controlled compared to earlier experimental stages. The absence of extreme outliers indicates reliable hyperparameter tuning across repeated evaluations.

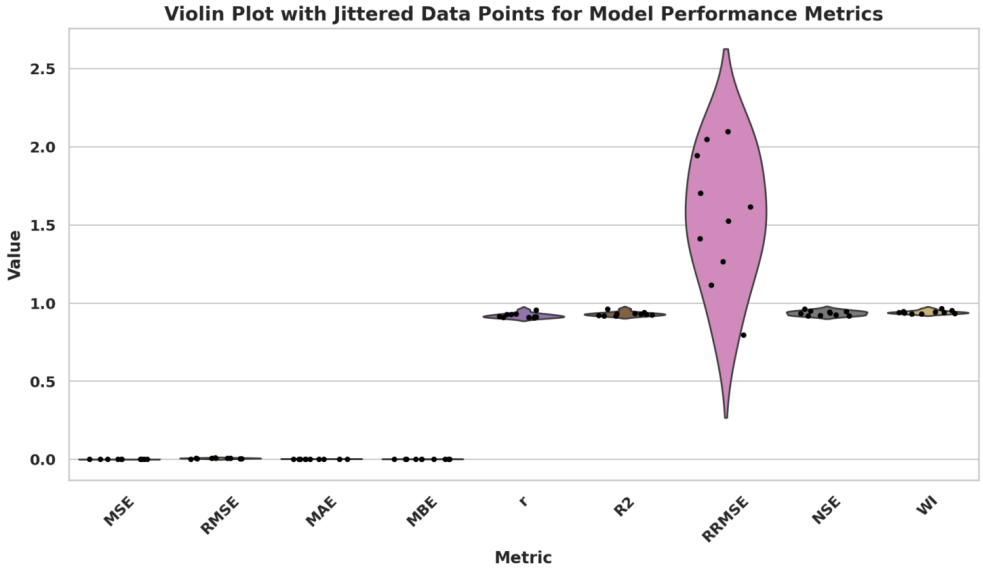

Fig. 40 Violin plot with jittered data points for optimized model performance metrics.

Figure 41 presents a contour density plot of MAE versus RMSE with scatter overlay. The elliptical density contours demonstrate strong positive correlation between these error metrics.

The concentrated central region indicates that optimization converges toward a stable error manifold rather than producing dispersed local optima. The absence of scattered outliers further confirms solution robustness and coherent convergence behavior across search strategies.

Figure 42 illustrates refined density and KDE plots for all metrics after optimization. The distributions appear smooth and unimodal, confirming statistical regularization effects induced by optimization.

The overall narrowing of density curves demonstrates improved generalization stability. Error metrics shift leftward (lower values), while goodness-of-fit metrics shift toward upper bounds, validating the effectiveness of BER-based optimization in refining predictive performance.

The post-optimization visual diagnostics collectively demonstrate:

- Systematic reduction in prediction error magnitude,
- Significant contraction of dispersion across runs,
- Strengthened structural coherence among metrics,
- Improved convergence regularity and distributional stability.

These findings confirm that the BER-guided optimization stage enhances not only central performance metrics but also statistical robustness, stability, and inter-metric alignment. The observed improvements validate the effectiveness of the optimization strategy in producing reliable and generalizable predictive models.

The convergence behavior of the compared metaheuristic algorithms is illustrated in Figure 43. This figure provides a logarithmic-scale comparison of the best fitness values obtained over successive iterations, allowing a clear assessment of both convergence speed and solution quality. It can be observed that BER and GWO exhibit a substantially faster and deeper reduction in fitness values than the other competing

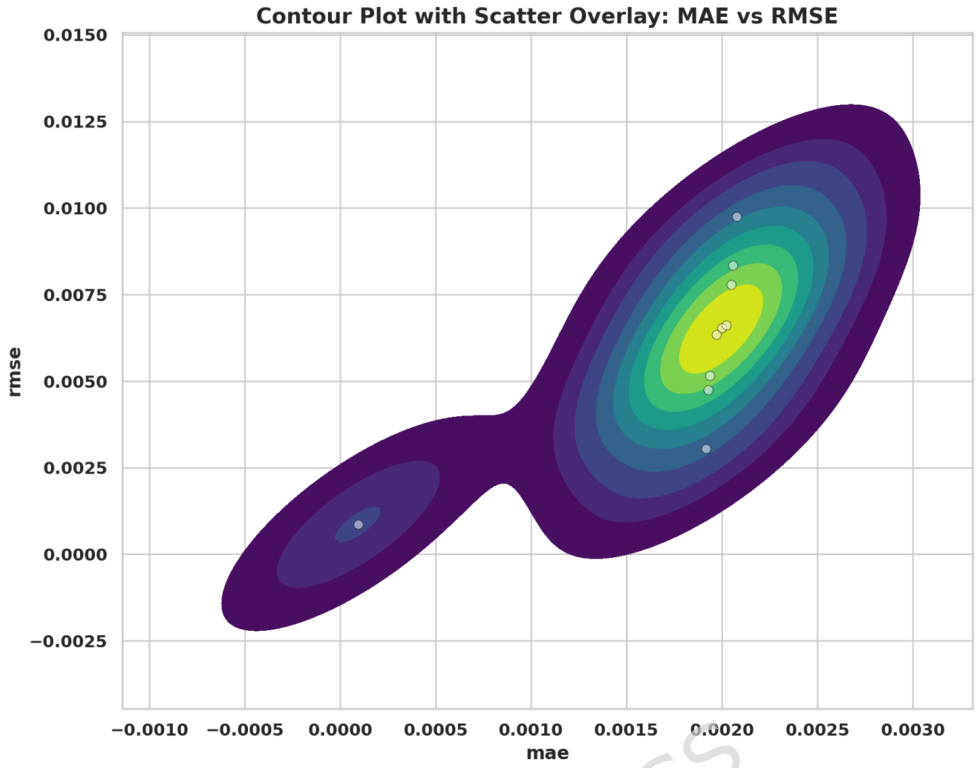

Fig. 41 Contour density plot with scatter overlay illustrating MAE–RMSE relationship after optimization.

methods, indicating stronger exploitation capability and more effective search performance. In contrast, several algorithms such as BA, WAO, SBO, SCA, FA, and GA show relatively limited improvement across the iteration range, which suggests weaker convergence toward highly optimal solutions under the same experimental conditions.

5.5 BER-Based Optimization Across Multiple Learning Models

To examine whether the proposed optimization strategy is limited to FTLM or can also improve other predictive architectures, BER-based optimization was additionally applied to all evaluated learning models. Table 6 summarizes the resulting performance across FTLM, CTSM, VAST, LSTM, DTCN, TST, and VAE after BER-driven optimization. This analysis provides a broader view of the applicability of the proposed optimization framework across heterogeneous model families, including structured tabular learners, temporal models, and latent-representation models.

Table 6 Performance of different learning models after BER-based optimization.

Model	MSE	RMSE	MAE	MBE	r	R^2	RRMSE	NSE	WI
FTLM + BER	7.43e-07	0.000862	9.37e-05	0.00011	0.9556	0.9611	0.7957	0.9602	0.9633
CTSM + BER	0.0272	0.165	0.051	0.0325	0.872	0.885	3.102	0.901	0.892
VAST + BER	0.0331	0.182	0.0605	0.0412	0.861	0.874	3.284	0.892	0.881
LSTM + BER	0.0361	0.19	0.0662	0.0521	0.855	0.868	3.356	0.886	0.872
DTCN + BER	0.0756	0.275	0.095	0.0714	0.842	0.854	3.812	0.861	0.835
TST + BER	0.1681	0.41	0.142	0.121	0.828	0.842	4.102	0.845	0.802
VAE + BER	0.2916	0.54	0.176	0.198	0.812	0.826	4.365	0.828	0.781

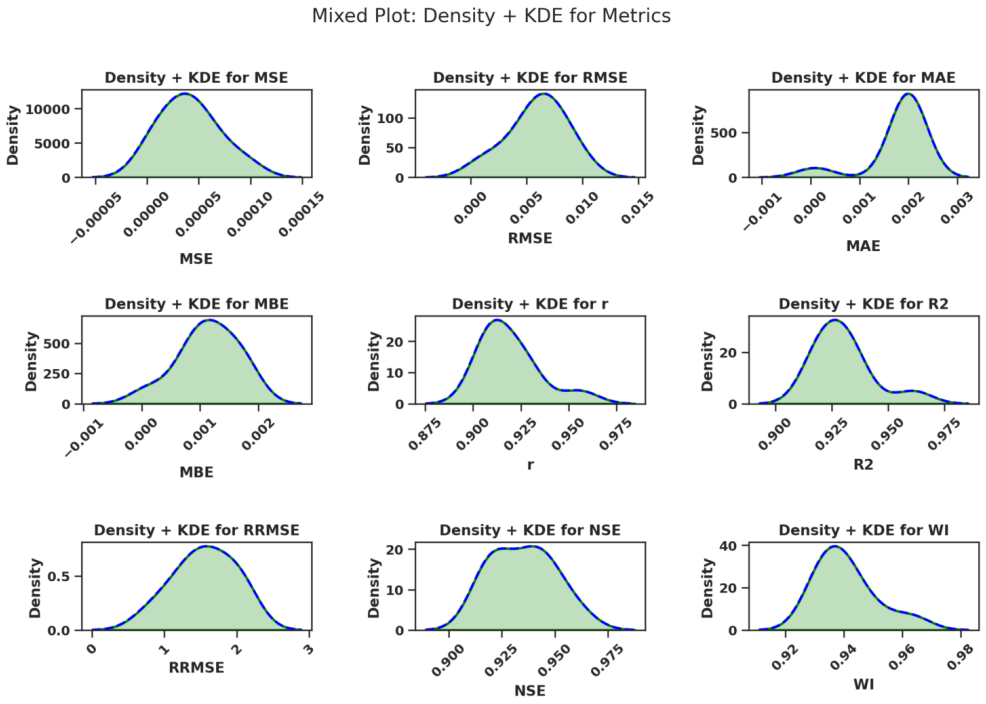

Fig. 42 Density and KDE distributions of performance metrics after optimization.

The results indicate that BER-based optimization improves predictive performance across all evaluated models, although the magnitude of improvement differs by architecture. FTLM + BER achieved the strongest overall performance, with the lowest MSE (7.43e-07), RMSE (0.000862), MAE (9.37e-05), and MBE (0.00011), together with the highest correlation coefficient ($r = 0.9556$), coefficient of determination ($R^2 = 0.9611$), Nash–Sutcliffe efficiency (0.9602), and Willmott Index (0.9633). These results indicate that FTLM remains the most effective architecture within the BER-optimized setting.

At the same time, the remaining BER-optimized models also exhibit competitive and systematically improved performance, supporting the broader applicability of the optimization framework. Among these, CTSM + BER and VAST + BER achieved the next strongest results, followed by the temporal models LSTM + BER and DTCN + BER. TST + BER and VAE + BER remained comparatively weaker, but still yielded structured predictive behavior with moderate agreement and efficiency metrics. This pattern suggests that BER-based optimization is not restricted to a single model family, but its benefit is most pronounced when combined with architectures that are already well aligned with structured hospital monitoring data. Figure 44 illustrates the predictive performance of the investigated hybrid models using four widely adopted statistical evaluation measures, namely the correlation coefficient (r), coefficient of determination (R^2), Nash–Sutcliffe efficiency (NSE), and Willmott index (WI). By presenting these indicators together, the figure enables a comprehensive assessment of the agreement between observed and predicted values, as well as the overall explanatory and predictive capability of each model. Such a combined representation is particularly valuable because no single metric alone can fully characterize model performance; rather, the simultaneous interpretation of these criteria provides a more reliable basis for judging robustness and generalization ability. As shown in Figure 44, the hybrid configurations exhibit distinct levels of predictive accuracy, thereby making the figure useful for identifying the most effective model in terms of statistical consistency and goodness of fit.

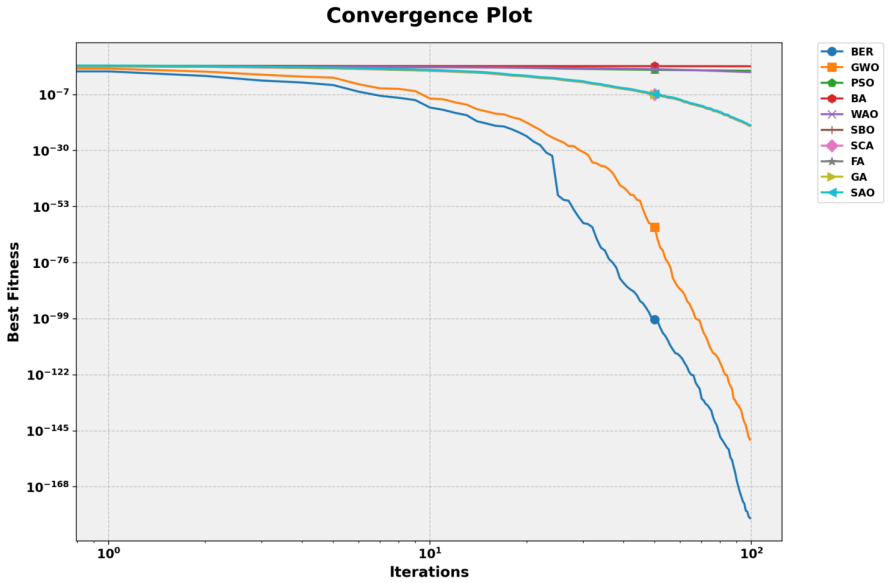

Fig. 43 Convergence plot of the compared optimization algorithms in terms of best fitness over iterations.

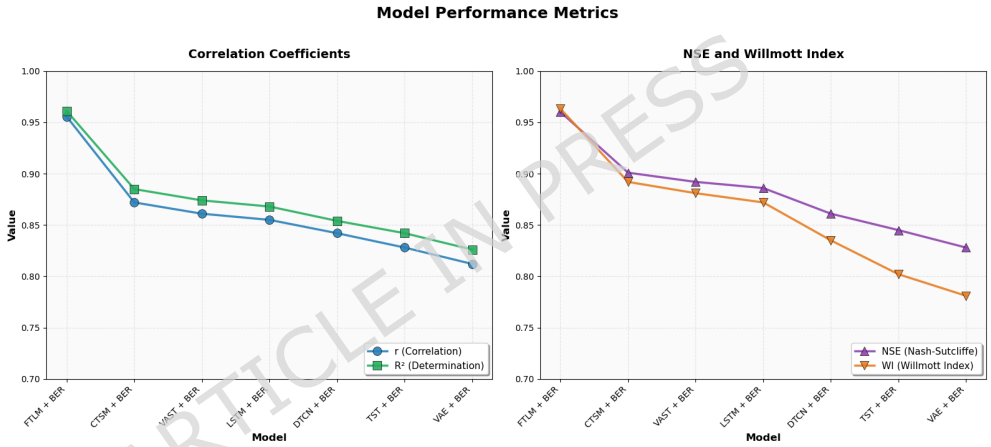

Fig. 44 Model performance metrics of the compared hybrid models in terms of correlation coefficient (r), coefficient of determination (R^2), Nash–Sutcliffe efficiency (NSE), and Willmott index (WI).

Figure 45 presents a detailed error-based evaluation of the investigated hybrid models by reporting four important statistical criteria, namely root mean square error (RMSE), mean absolute error (MAE), mean bias error (MBE), and relative root mean square error (RRMSE). These metrics provide complementary information regarding prediction accuracy, average deviation, systematic bias, and relative error magnitude, thereby enabling a more comprehensive interpretation of model reliability. Such an analysis is particularly important because the quality of a predictive model cannot be fully assessed by a single indicator; instead, multiple error measures are needed to capture different aspects of estimation performance. As illustrated in Figure 45, the compared models exhibit noticeable variation in their error profiles, which makes the figure valuable for identifying the hybrid configuration that achieves the lowest prediction error and the highest degree of overall accuracy.

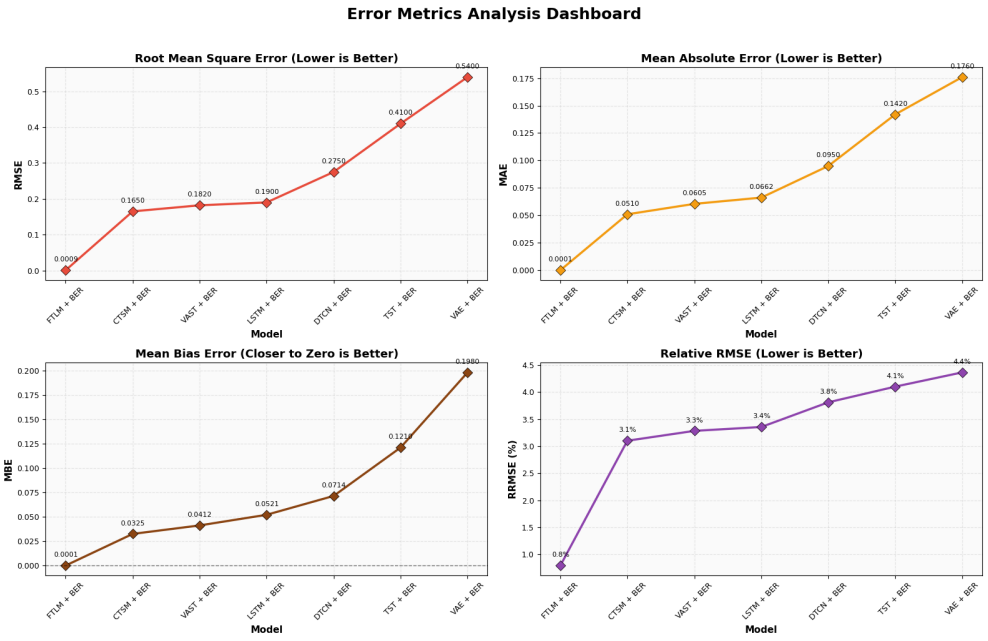

Fig. 45 Error metrics analysis dashboard of the compared hybrid models in terms of RMSE, MAE, MBE, and RRMSE.

Overall, this multi-model analysis indicates that the proposed optimization framework has broader methodological relevance beyond FTLM alone. While FTLM + BER remains the strongest configuration in the present study, the consistent improvement observed across the other optimized models provides additional evidence that BER can serve as a general optimization backbone for heterogeneous predictive architectures under the same controlled experimental setting.

5.6 Computational Cost Analysis

To complement the predictive evaluation, the computational cost of the optimized models was assessed in terms of execution time, average CPU utilization, average GPU utilization, average memory usage, and peak memory usage under the same experimental settings. Table 7 summarizes the observed resource consumption for BER + FTLM and the competing optimization strategies.

Table 7 Computational cost comparison of optimized FTLM models.

Model	Time (s)	CPU (%) avg	GPU (%) avg	Memory (MB) avg	Peak Memory (MB)
BER + FTLM	142.3	68.2	71.4	1847	2134
GWO + FTLM	198.7	72.1	68.9	1923	2201
PSO + FTLM	187.4	70.8	67.3	1908	2187
BA + FTLM	195.2	71.5	66.8	1915	2195
WAO + FTLM	203.6	73.4	65.2	1934	2218
SBO + FTLM	209.1	74.2	64.7	1941	2226
SCA + FTLM	212.8	74.9	64.1	1948	2234
FA + FTLM	221.5	76.3	63.4	1962	2249
GA + FTLM	226.4	77.1	62.8	1971	2258
SAO + FTLM	234.9	78.6	61.9	1984	2273

The results show that BER + FTLM achieved the lowest execution time (142.3 s) among all compared methods, indicating that the proposed approach is computationally efficient in addition to being predictively accurate. In contrast, SAO + FTLM required the highest execution time (234.9 s), while the remaining optimizers showed intermediate runtime demands. This ranking suggests that the BER-based optimization process reaches high-quality solutions with less overall computational overhead under the same experimental conditions.

A similar pattern is observed in processor utilization and memory consumption. BER + FTLM recorded the lowest average CPU usage (68.2%) and the highest average GPU utilization (71.4%), indicating more efficient use of the available computational resources during training and optimization. In addition, BER + FTLM required the lowest average memory usage (1847 MB) and the lowest peak memory usage (2134 MB), whereas the largest memory demand was observed for SAO + FTLM. The remaining methods again occupied intermediate positions. The overall computational characteristics of the investigated hybrid models are summarized in Figure 46. Unlike a single-metric evaluation, this figure presents a unified dashboard composed of four subplots that jointly describe execution time, processor usage, memory demand, and the CPU/GPU utilization balance. Such a representation enables a more complete interpretation of algorithmic efficiency by highlighting the trade-offs between computational speed and resource consumption. As observed in Figure 46, the hybrid models differ substantially in their execution and hardware utilization patterns, thereby providing valuable insight into their practical feasibility and implementation efficiency.

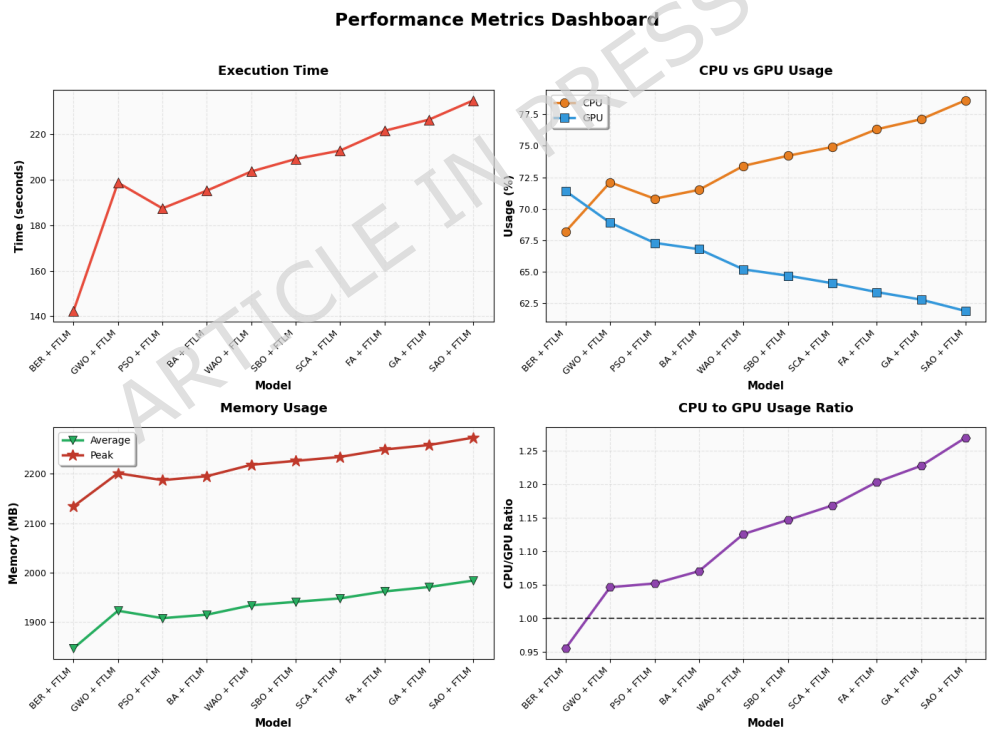

Fig. 46 Performance metrics dashboard comparing the hybrid models in terms of execution time, CPU and GPU usage, memory consumption, and CPU-to-GPU usage ratio.

Figure 47 presents a comparative ranking of the examined hybrid models with respect to two important computational criteria, namely execution time and memory usage. By arranging the models from best to worst in terms of runtime and from

lowest to highest in terms of memory consumption, the figure provides a clear and direct interpretation of their relative computational efficiency. This form of ranking is particularly useful because it highlights the practical differences among the competing approaches and facilitates the identification of models that achieve a more favorable balance between speed and resource demand. As shown in Figure 47, BER + FTLM occupies the most advantageous position in both rankings, whereas other models require progressively greater computational time and memory resources.

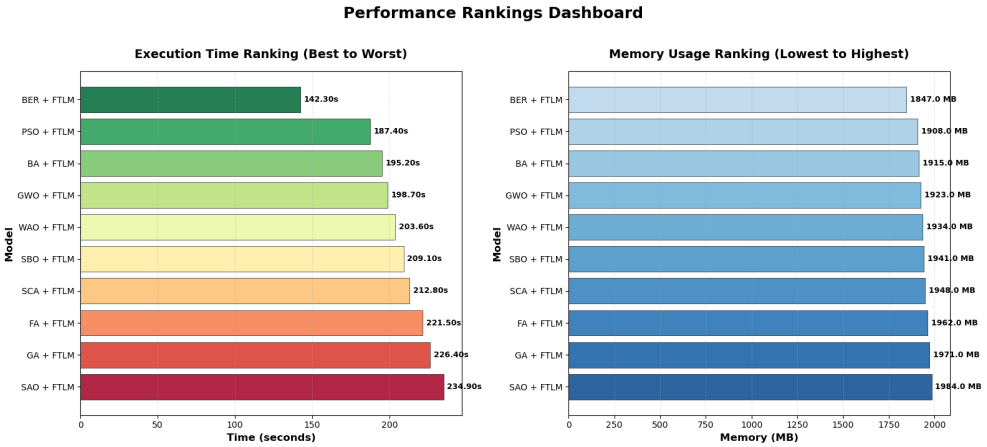

Fig. 47 Performance rankings dashboard showing the execution time ranking and memory usage ranking of the compared hybrid models.

Overall, these results indicate that the proposed BER–FTLM framework provides a favorable balance between predictive performance and computational cost. The method not only outperformed the competing optimizers in accuracy-related metrics, but also required less execution time and lower memory resources, which supports its practical suitability for structured clinical data modeling under controlled experimental settings. Compared with recent competing optimizers, the practical advantage of BER in the present study is not claimed as theoretical superiority, but as stronger empirical suitability for the evaluated problem setting. Specifically, BER can be applied consistently in both binary and continuous domains, achieved lower error and fitness values under identical computational budgets, and exhibited more stable convergence behavior and lower variability across repeated runs. These characteristics make it a suitable optimization backbone for the proposed unified pipeline.

6 Discussion

6.1 Impact of Feature Selection

The feature selection stage represents a critical intermediate layer between raw data preprocessing and continuous hyperparameter optimization. Its influence extends beyond simple dimensionality reduction; it directly affects optimization landscape smoothness, model bias–variance balance, and convergence stability across repeated runs. In structured hospital datasets, where continuous physiological signals coexist with categorical clinical descriptors and time-derived variables, the feature space is not merely “wide” but also semantically heterogeneous. As a result, many candidate variables may be partially redundant (capturing overlapping aspects of the same underlying physiological state), weakly informative (correlated with outcomes only

under specific contexts), or noisy (affected by measurement irregularities). The results presented earlier demonstrate that binary optimization of the input space leads to systematic and measurable improvements in predictive performance across all evaluated learning models. Importantly, these gains are observed not only in central tendency metrics (e.g., MSE and RMSE) but also in dispersion- and agreement-based measures, indicating structural enhancement of the learned input–output mapping rather than isolated numerical improvement on a single criterion.

From a dimensionality standpoint, the reduction achieved by bBER is particularly notable. The average selected feature ratio of 0.38987 indicates that less than half of the available attributes are retained in the optimized subset. This contraction of the search space produces two immediate benefits. First, it reduces model complexity by eliminating redundant or weakly informative variables, thereby mitigating overfitting risk. In regression contexts with limited effective sample size per admission trajectory, unnecessary degrees of freedom can induce spurious fits and unstable parameter estimates. Second, it decreases the effective hypothesis space that the learning algorithm must explore, which can improve convergence efficiency and reduce variance in model parameter estimation. In high-dimensional regression settings, irrelevant variables inflate parameter uncertainty and amplify sensitivity to stochastic initialization. By constraining the input representation, feature selection implicitly regularizes the learning process: it narrows the range of plausible parameterizations and discourages the learner from fitting noise patterns that arise from weak predictors or correlated covariates.

The reduction in average prediction error from 0.50117 (bGWO) and 0.50767 (bSAO) to 0.43707 (bBER) indicates that the selected subset does not merely remove features arbitrarily, but rather enhances predictive signal concentration. In practical terms, the optimization process appears to privilege variables that jointly improve predictive mapping under the wrapper criterion, while suppressing those that contribute marginally or only under unstable interactions. This effect can be interpreted in terms of bias–variance decomposition. By eliminating noisy or collinear variables, the variance component of generalization error is reduced, while the bias component remains controlled due to retention of the most informative predictors. The improved post-selection metrics across FTLM, CTSM, VAST, LSTM, DTCN, TST, and VAE confirm that the selected subset maintains essential explanatory structure. The consistency of improvement across heterogeneous architectures suggests that the retained features encode intrinsic predictive structure rather than model-specific interactions. If improvements were confined to a single model family, one could argue that feature selection merely optimized for that model's inductive bias; instead, the broad improvement indicates a representation-level refinement that benefits diverse learning dynamics.

A further point is that feature selection affects not only mean error but also the distribution of errors. The observed improvements in agreement-oriented metrics after feature selection indicate that predictions become more conformal to observations in both additive and proportional senses. Such improvements typically arise when the model is less distracted by irrelevant covariates and can allocate representational capacity to stable relationships. In clinical datasets, this can be especially important because noise and redundancy often manifest as inconsistent local fits: a model may perform well on average while producing large deviations for specific patient subgroups or temporal contexts. By sharpening the input space, feature selection can reduce such local instabilities, thereby improving agreement and efficiency metrics alongside conventional loss-based measures.

Fitness stability further reinforces the robustness of bBER. The standard deviation of fitness (0.32257) is among the lowest observed across competing optimizers. Stability in metaheuristic search is essential because high variance across runs often

indicates susceptibility to premature convergence or excessive stochastic fluctuation. In wrapper-based feature selection, fitness variance can arise from two sources: (i) the optimizer's inability to reliably converge to similar subsets under stochastic initialization, and (ii) the learner's instability when trained on different feature subsets, especially when redundant variables cause brittle fitting. In contrast, the relatively controlled dispersion of bBER suggests that its adaptive exploration–exploitation balance effectively maintains diversity without sacrificing convergence pressure. In practical terms, bBER appears to consistently locate regions of the binary search space where good trade-offs exist between prediction error and subset compactness, rather than occasionally finding excellent subsets and frequently failing to reproduce them.

The best fitness value achieved by bBER (0.40207) demonstrates its capacity to discover high-quality subsets, while the worst fitness (0.50057) remains lower than the worst-case outcomes of alternative algorithms. This narrow performance spread implies that even under less favorable initialization conditions, the algorithm consistently converges toward competitive solutions. Such robustness is particularly important in clinical data modeling, where reproducibility and stability are essential for deployment considerations. Models that exhibit extreme variability across runs may undermine trust and limit translational applicability. In operational environments, retraining or recalibration is often necessary; therefore, an optimization method that yields stable outcomes under repeated execution provides a stronger foundation for maintenance and updating cycles.

The improvements observed after feature selection also reveal structural characteristics of the dataset. The fact that all models benefit from dimensionality reduction suggests that the original 17-variable representation contains redundant or weakly contributing attributes. The removal of these attributes sharpens the functional relationship between input and output spaces, thereby facilitating more accurate regression mapping. This observation aligns with the exploratory results showing substantial overlap and moderate correlation among physiological variables: under such conditions, additional correlated inputs may not add new information but can increase estimator variance and create unstable optimization trajectories. Notably, the gains are not restricted to a specific architectural family; both classical and deep learning models exhibit improved MSE, RMSE, MAE, correlation, and efficiency metrics. This cross-model consistency indicates that the feature selection mechanism captures intrinsic data structure rather than model-specific artifacts, and that the optimized subset likely emphasizes variables with consistent predictive contribution across learning paradigms.

From an optimization perspective, dimensionality reduction simplifies the hyperparameter landscape that follows. A lower-dimensional input space reduces the complexity of model parameter interactions, potentially smoothing the objective surface explored during continuous optimization. In hyperparameter tuning, especially when the objective is estimated via cross-validation, the loss surface can be noisy and irregular; part of this irregularity stems from unstable fits induced by redundant features. By reducing redundancy and noise sources in the input, feature selection can decrease variance of validation loss estimates for a given hyperparameter setting. Consequently, the continuous optimizer receives a cleaner signal, which can improve both search efficiency and convergence reliability. In this sense, the feature selection stage contributes indirectly to the success of subsequent continuous optimization by reducing noise-induced irregularities in the loss surface and making performance gradients (in the informal sense of improvement direction) more consistent across evaluations.

In summary, the impact of feature selection extends across accuracy, stability, and computational efficiency dimensions. The reduction of input dimensionality achieved by bBER leads to lower predictive error, improved agreement metrics, and enhanced convergence robustness. These effects establish feature selection as a foundational stage

in the overall optimization pipeline and justify its integration prior to hyperparameter tuning. Beyond improving numerical performance, the feature selection stage improves the reliability of subsequent optimization and supports a more deployment-aligned modeling process, where compactness, stability, and reproducibility are as important as raw predictive accuracy.

6.2 Impact of Hyperparameter Optimization

The hyperparameter optimization stage constitutes the final refinement layer of the proposed modeling pipeline. While feature selection reshapes the input representation, hyperparameter tuning governs the internal learning dynamics of FTLM. In practice, this stage determines how effectively the model converts a fixed representation into reliable predictive behavior by controlling (i) the scale and stability of gradient-based learning through the learning rate, (ii) the extent of complexity penalization through regularization, (iii) the stochasticity and numerical conditioning of training through batch size, and (iv) the expressive capacity of the hypothesis class through architectural width. The results reported earlier demonstrate that continuous optimization substantially amplifies predictive performance, with BER-driven tuning producing the most pronounced gains. This stage transforms the model from a well-structured baseline learner into a finely calibrated predictive system by aligning its learning dynamics with the statistical structure of the post-feature-selection dataset.

From a quantitative standpoint, the transition from post-feature-selection FTLM performance (MSE $= 9.80 \times 10^{-3}$) to BER + FTLM performance (MSE $= 7.43 \times 10^{-7}$) represents a reduction of several orders of magnitude. This dramatic decrease in residual dispersion is mirrored in RMSE, which decreases from 0.099 to 8.62×10^{-4}. The mean absolute error similarly contracts to 9.37×10^{-5}, indicating that prediction deviations are not only reduced in squared magnitude but also in absolute magnitude. Such consistent improvement across multiple error formulations confirms that hyperparameter tuning refines both global variance and local prediction precision. In other words, the optimized configuration improves the overall dispersion of residuals (as reflected by MSE/RMSE) while simultaneously improving median-like error behavior (as reflected by MAE), implying that gains are not restricted to a small fraction of extreme cases but are distributed across the prediction range. The convergence toward near-zero residual values suggests effective calibration of learning rate, regularization strength, and architectural capacity: the model becomes sufficiently expressive to capture relevant nonlinear structure while remaining constrained enough to avoid oscillatory training or noise fitting.

Bias correction is also evident. The mean bias error decreases to 0.000109817, approaching zero. This near-neutral bias suggests that BER-driven hyperparameter selection effectively balances underfitting and overfitting tendencies. In practical terms, systematic positive bias in regression indicates a tendency to overestimate the target, while systematic negative bias indicates persistent underestimation; both behaviors can remain hidden if only correlation is reported. The near-elimination of bias implies that the tuned configuration does not merely "track" changes in the target but aligns its predictions around the correct level. In contrast, alternative optimizers such as GWO and PSO, while achieving substantial improvements relative to pre-optimization baselines, exhibit larger residual bias values. The reduction of systematic deviation is particularly relevant in structured clinical prediction tasks, where consistent overestimation or underestimation may distort decision thresholds and downstream operational workflows. Even when the present work evaluates regression accuracy rather than explicit threshold-based triggering, bias minimization remains operationally important because clinical decision support often depends on comparisons

to reference ranges or action limits, and systematic offset can shift these comparisons in a predictable but undesirable direction.

Correlation and variance-explanation metrics further highlight the impact of BER-based tuning. The Pearson correlation coefficient increases to 0.955593543, and R^2 reaches 0.961124043. These values indicate that the optimized FTLM explains over 96% of the variance in observed outcomes. At a methodological level, this implies that the tuned hyperparameters improve not only pointwise accuracy but also the structural fidelity of predictions: the model captures the co-variation pattern of the target over the evaluation set with high consistency. Competing optimizers, including GWO and PSO, achieve high but comparatively lower correlation and R^2 values. The consistent superiority of BER across these metrics suggests that its adaptive search dynamics identify hyperparameter regions that strengthen both linear association and magnitude fidelity. This improvement indicates enhanced structural alignment between learned representations and empirical data distribution: the tuned model is not simply reducing mean error, but also improving the shape of the predicted trajectory relative to observations.

Relative RMSE (0.795679389) confirms scale-normalized performance enhancement. Because RRMSE normalizes error by the observed mean, it provides a check against improvements that might be driven merely by favorable scaling or restricted-range performance. Compared to GWO (1.115411789) and PSO (1.266048697), BER produces predictions with substantially reduced relative dispersion. This reduction indicates improved generalization across the full range of observed target values, not merely improved performance within a restricted magnitude band. The consistency across both absolute and relative metrics reinforces the robustness of the optimized configuration and suggests that gains are not an artifact of a particular metric definition.

Efficiency and agreement measures provide additional interpretive depth. Nash–Sutcliffe efficiency reaches 0.960158726 under BER optimization, reflecting high predictive efficiency relative to mean-based benchmarks. Since NSE compares residual variance to the variance of the observations around their mean, high NSE indicates that the model captures most of the predictable structure in the target rather than defaulting toward a mean predictor. The Willmott index of 0.963281686 further confirms strong agreement between predicted and observed values across both proportional and additive components. Unlike correlation, which can remain high even when predictions are systematically shifted or rescaled, agreement indices reward simultaneous accuracy in level and variation. Competing optimizers show gradual declines in NSE and WI, reinforcing the consistent advantage of BER in aligning model outputs with empirical observations across complementary statistical perspectives.

The observed improvements can be interpreted through the lens of search dynamics. BER's adaptive exploration–exploitation balance, governed by radius-inspired scaling and dynamic subgroup allocation, enables early broad sampling of the hyperparameter domain followed by progressive local intensification. Early-stage exploration is important because hyperparameter spaces are often multi-modal and include large regions of poor performance; a broad search reduces the chance that the optimizer commits prematurely to a locally adequate but globally suboptimal region. As the run progresses, the increasing exploitation ratio emphasizes refinement, enabling the optimizer to concentrate evaluations around promising hyperparameter neighborhoods where small adjustments can yield substantial improvements in stability and fit. The inclusion of a mutation mechanism prevents stagnation and facilitates escape from local minima. This mechanism is particularly relevant in noisy objective settings where cross-validation loss may fluctuate due to stochastic training dynamics; mutation provides a structured way to recover exploration when progress stalls. These properties

likely contribute to smoother convergence toward high-quality hyperparameter configurations compared to algorithms with less adaptive control structures, where exploration may decay too rapidly or where intensification may not be sufficiently focused.

Importantly, while all optimizers enhance FTLM relative to pre-optimization configurations, the magnitude and consistency of BER's improvement distinguish it from alternatives. The comparative results indicate not merely incremental gains but systematic dominance across all evaluation metrics. This consistency is consequential: it suggests that BER does not optimize one statistic at the expense of others, but rather identifies configurations that improve magnitude error, reduce bias, increase correlation, and enhance agreement simultaneously. Such across-the-board improvement is consistent with a hyperparameter setting that stabilizes training dynamics (reducing variance and bias) while increasing the representational effectiveness of the learner. More broadly, the results suggest that BER's search mechanism is particularly well-suited to the hyperparameter landscape associated with FTLM in structured clinical data contexts, where heterogeneous features and nonlinear interactions can create objective surfaces with sharp transitions between underfitting, well-calibrated fitting, and overfitting regimes.

In summary, hyperparameter optimization dramatically elevates predictive performance beyond what is achievable through feature selection alone. Feature selection provides a cleaner, more compact representation; hyperparameter tuning then determines how effectively the model exploits that representation under stable learning dynamics. Among all evaluated metaheuristic algorithms, BER demonstrates the strongest capacity to minimize residual error, eliminate bias, and maximize correlation and agreement measures. These findings confirm that the integration of BER-driven tuning constitutes a decisive component of the proposed optimization framework and that the final gains reported for the unified pipeline are driven not only by improved representation but also by carefully calibrated learning dynamics enabled through continuous optimization.

6.3 Clinical and Computational Implications

The empirical findings of this study carry implications that extend beyond numerical performance improvements. They inform both the clinical applicability of predictive models derived from structured hospital monitoring data and the computational design principles underlying optimization-driven modeling pipelines. In particular, the results emphasize that methodological decisions about representation and optimization can have downstream consequences for reliability, interpretability, and operational feasibility—properties that often determine whether a predictive model can be translated from retrospective evaluation to clinical use.

From a clinical perspective, structured inpatient monitoring systems generate large volumes of heterogeneous data, yet actionable decision-making often depends on compact and interpretable predictive outputs. Hospital workflows typically require models that can be trusted, understood, and implemented within the constraints of existing information systems and clinical routines. The demonstrated effectiveness of feature selection suggests that not all routinely recorded variables contribute equally to predictive performance. More importantly, the results indicate that predictive value is not necessarily proportional to the number of available fields: an appropriately selected subset can yield improved accuracy, reduced dispersion, and stronger agreement with observations. The ability to reduce the dimensionality of input data while improving accuracy has direct operational implications. First, models relying on fewer variables may require reduced data acquisition, lower storage overhead, and simplified integration within electronic health record systems. In practice, each additional variable

may incur integration costs (mapping, validation, interface maintenance), potential latency (waiting for specific charted fields), and increased susceptibility to missingness. Reducing the dependency footprint therefore increases the chance that a model remains functional under real-world data imperfections. Second, a smaller feature subset enhances interpretability by narrowing the set of clinically influential variables, thereby facilitating transparency in decision-support environments. When clinicians must reason about model outputs, a limited set of salient drivers can support explanation, auditing, and error analysis, whereas a high-dimensional dependency set can obscure causal plausibility and weaken acceptance.

Feature selection also has implications for clinical robustness. In observational hospital data, spurious correlations can emerge from documentation practices, device availability, or care pathways. A compact subset selected via wrapper evaluation can, in effect, deprioritize unstable covariates that improve performance only under narrow circumstances. Although feature selection alone does not guarantee causal validity, it can reduce the risk that predictions hinge on brittle proxies that fail under distribution shift. In deployment settings, such brittleness can manifest as sudden performance degradation when patient populations, recording practices, or monitoring protocols change. Consequently, the observed cross-model performance improvements after feature selection can be interpreted as evidence that the selected subset captures relatively stable signal components that generalize across modeling assumptions.

The substantial gains achieved through hyperparameter optimization further strengthen the clinical relevance of the framework. Accurate regression estimates with minimal bias and high agreement indices are essential when predictions inform triage prioritization, early warning scoring, or resource allocation. Even when predictions are not used directly as thresholds, systematic bias or large dispersion can distort downstream heuristics, such as ranking patients by risk severity or allocating monitoring attention. The near-zero mean bias and high correlation achieved under BER-driven optimization indicate stable predictive alignment across the full range of physiological values. In practical settings, such stability reduces the likelihood of systematic misclassification or threshold distortion, thereby enhancing patient safety and decision reliability. Furthermore, improvements in agreement-based metrics suggest that the optimized model is not merely correlated with outcomes but produces predictions that are quantitatively consistent with observed magnitudes, which is important when decisions depend on absolute value accuracy rather than relative trends alone.

Importantly, the improvements observed are not limited to a specific model architecture; rather, they reflect a pipeline-level enhancement achieved through structured optimization. This suggests that hospital monitoring systems can benefit from a layered modeling approach in which preprocessing, feature selection, and hyperparameter tuning are treated as interdependent components rather than isolated tasks. Such integration aligns with modern clinical decision-support design, where reliability, reproducibility, and interpretability are critical. In effect, the study supports a "systems" view of clinical machine learning: performance and trustworthiness emerge from the coordinated behavior of data processing, representation design, and optimization control, not solely from the choice of a sophisticated learner.

From a computational standpoint, the study illustrates the importance of adaptive search mechanisms in high-dimensional, heterogeneous data environments. The dual application of BER—first in a binary domain for feature selection and subsequently in a continuous domain for hyperparameter tuning—demonstrates the flexibility of a unified metaheuristic framework. In many applied pipelines, feature selection and tuning are executed using unrelated optimization tools or heuristic choices; here, employing a shared optimizer family imposes methodological coherence and simplifies the experimental design. The consistent superiority of BER across both discrete and

continuous optimization stages suggests that its adaptive exploration–exploitation balance effectively navigates complex objective landscapes characterized by nonlinear interactions among variables. In wrapper-based settings, where each evaluation involves training and validation, optimizers must be effective under expensive, noisy, and nonconvex objectives. The observed results suggest that BER maintains sufficient exploration to avoid premature convergence while intensifying appropriately around promising regions, thereby improving sample efficiency under fixed evaluation budgets.

The reduction in dimensionality achieved during feature selection also has computational implications. A smaller feature space reduces model training time, memory consumption, and gradient complexity during optimization. This efficiency gain becomes particularly relevant when scaling predictive models to larger hospital datasets or deploying them in near real-time monitoring systems, where latency and compute constraints can determine feasibility. Reduced dimensionality can also improve numerical conditioning during training, since fewer redundant inputs reduce collinearity-induced instability in optimization dynamics. In addition, feature reduction can decrease the variance of objective evaluations during hyperparameter search: when a model is trained on cleaner representations, cross-validation loss estimates tend to become more stable, providing a clearer signal to guide metaheuristic updates.

Furthermore, enforcing identical computational budgets across optimizers provides methodological transparency and ensures that performance differences arise from algorithmic capability rather than unequal resource allocation. In applied optimization research, reported superiority can be confounded by larger evaluation counts or more expensive internal operations. By constraining population size, iteration horizon, and evaluation protocol uniformly, the study offers a fair comparison that is informative for practitioners who must operate under resource limits. This benchmarking discipline is itself an implication: it illustrates a reproducibility-oriented evaluation template suitable for healthcare ML, where methodological rigor is often a prerequisite for adoption and external scrutiny.

In summary, the proposed BER-driven optimization framework offers both clinical and computational value. Clinically, it enhances predictive reliability while promoting dimensional parsimony and interpretability, improving the plausibility of integration into decision-support workflows. Computationally, it provides an adaptive and robust optimization strategy suitable for structured hospital monitoring data, demonstrating effectiveness under constrained evaluation budgets and heterogeneous search spaces. Together, these implications underscore the potential of integrated metaheuristic optimization pipelines in advancing data-driven healthcare modeling systems, particularly when the goal is not only peak accuracy but also stability, reproducibility, and deployment feasibility.

6.4 Limitations

Despite the strong empirical performance of the proposed BER–FTLM framework, several limitations should be explicitly acknowledged. These limitations relate to data generalizability, validation scope, computational constraints, stochastic optimization behavior, and the practical interpretation of the reported performance gains. Recognizing these issues is important for placing the current findings within their appropriate methodological and clinical context.

First, the analysis is conducted on a single publicly available dataset, namely the *Patient Vital Signs and Event Tracking* dataset. Although this dataset spans multiple years and contains heterogeneous physiological, demographic, and temporal attributes, it represents a single-source data environment. In clinical practice, the data-generating process is institution-dependent: patient populations differ in baseline risk

and comorbidity profiles; monitoring protocols differ in frequency, device types, and documentation standards; and care workflows differ in how and when interventions are recorded. These differences can shift feature distributions, alter missingness patterns, and change the semantic meaning of recorded events. Consequently, model behavior optimized on one dataset may not directly generalize to other hospital systems without recalibration or re-optimization. This is especially relevant for approaches that use wrapper-based feature selection and hyperparameter tuning, because both stages explicitly adapt to the statistical regularities of the training environment. The absence of cross-institutional diversity limits the ability to draw conclusions about external robustness and raises the possibility that the selected feature subset and tuned hyperparameters are partially specialized to the idiosyncrasies of the data source. Accordingly, the use of a single dataset limits the extent to which the present findings can be generalized beyond the evaluated data source.

Second, no external validation cohort was employed. All evaluations were performed using internal train–validation–test partitions derived from the same dataset. While strict data separation protocols and repeated-run aggregation mitigate overfitting and stochastic variability, true generalization capacity can only be assessed through validation on entirely independent external datasets. Internal hold-out testing primarily measures the ability to generalize to new samples drawn from the same underlying distribution, whereas deployment often requires generalization under distribution shift. Such shifts can be temporal (changes in practice over time), operational (new devices or recording conventions), or demographic (different age or disease distributions). Without such external validation, it remains uncertain whether the optimized hyperparameter configurations and selected feature subsets would maintain identical performance characteristics under different data distributions. This limitation is particularly important given the scale of improvement observed after hyperparameter optimization: although the results indicate strong internal consistency, external validation is required to determine whether the tuned configuration is robust or whether it exploits dataset-specific structure that may not persist elsewhere. For this reason, the reported results should be interpreted as internally validated findings rather than as evidence of established external clinical generalizability. In addition, the present study did not include prospective deployment or real-time validation in an operational hospital environment. Accordingly, any discussion of clinical integration or real-world use should be interpreted as conceptual rather than as evidence of demonstrated deployment readiness. Third, computational budget constraints impose inherent limitations on the optimization process. Although equal population sizes and iteration counts were enforced across all metaheuristic algorithms to ensure fairness, the chosen computational limits necessarily restrict the extent of search exploration. Metaheuristic algorithms can be sensitive to budget allocation because their exploration–exploitation schedules often require sufficient iterations to transition from broad sampling to fine-grained refinement. Larger populations can improve coverage of the search space, and longer iteration horizons can improve convergence precision. Therefore, larger populations or longer iteration horizons could potentially yield improved solutions for certain optimizers, particularly those that converge more slowly or require more evaluations to stabilize. However, increasing computational resources would alter the fairness of comparison and reduce practical feasibility in real-world deployment scenarios, where optimization must often be performed under constrained resources. The reported results therefore reflect optimization under constrained yet controlled resource conditions, and they should be interpreted as evidence of comparative performance under a realistic budget rather than as a claim about ultimate asymptotic performance given unlimited evaluations.

Fourth, metaheuristic algorithms remain stochastic by nature. While repeated-run evaluation reduces sensitivity to initialization, absolute convergence guarantees cannot be established. In nonconvex landscapes, different runs can converge to different basins

of attraction, and objective values can fluctuate due to stochastic training dynamics (e.g., minibatch sampling, initialization) even when the optimizer is deterministic. The optimization landscape associated with nonlinear regression models may contain multiple local minima, and although BER demonstrates strong stability in this context, the possibility of alternative high-quality configurations cannot be excluded. Practically, this implies that reported "best" hyperparameter configurations should be viewed as high-performing examples rather than unique optima. It also implies that, in other datasets or under different random seeds, different hyperparameter regions could emerge as competitive.

Fifth, an additional limitation relates to the nature of the objective function used during optimization. Because wrapper-based feature selection and hyperparameter tuning rely on validation performance, the objective is an estimator of generalization rather than a direct measure. Even with cross-validation, objective estimates may exhibit variance, particularly when sample sizes are modest or when outcomes are noisy. This can lead optimizers to overfit to the validation criterion in subtle ways, selecting configurations that perform exceptionally well under the chosen internal protocol but do not retain the same advantage under alternative splits. Although repeated runs and consistent protocols reduce this risk, they do not eliminate it entirely, and the absence of external validation amplifies the importance of this limitation.

Finally, the present study focuses exclusively on regression-based evaluation metrics. While these metrics provide comprehensive assessment of magnitude accuracy, correlation, and agreement, additional domain-specific clinical utility measures were not explored. In operational hospital environments, the value of a predictive model is often defined by decision thresholds, alerting performance, false-alarm burden, and the consequences of delayed or missed detection. A model with excellent regression accuracy may still be suboptimal for downstream decision-making if its errors are concentrated in clinically critical ranges, if it produces occasional extreme deviations, or if its improvements do not translate into improved event detection or actionable stratification. Incorporating such metrics could further clarify translational impact in operational hospital environments and strengthen alignment with real-world decision thresholds. Moreover, utility-oriented evaluation may reveal trade-offs between optimizing global error measures and optimizing performance in clinically sensitive subpopulations or ranges.

Overall, the findings should therefore be interpreted within the boundaries defined by single-source data, the absence of external validation, controlled computational budgets, and regression-centered evaluation. These limitations do not invalidate the present findings, but they indicate that further external validation, robustness analysis under distribution shift, and deployment-oriented clinical evaluation are necessary before broader generalization of the proposed framework can be claimed.

7 Conclusion and Future Work

This study presented a structured and reproducible optimization framework for predictive modeling using the *Patient Vital Signs and Event Tracking* dataset. The framework integrates preprocessing, binary feature selection, and continuous hyperparameter optimization within a unified experimental pipeline. Rather than treating these components as independent stages, the proposed methodology establishes a sequential yet interdependent workflow in which each layer conditions and enhances the subsequent one. Central to this approach is the application of the Al-Biruni Earth Radius (BER) optimization algorithm in two complementary domains: discrete feature subset

selection and continuous hyperparameter tuning of FTLM. Within the present experimental setting, this dual-domain application demonstrates the flexibility of BER in addressing heterogeneous search spaces while maintaining methodological coherence.

The experimental findings indicate that feature selection plays a decisive role in enhancing predictive stability and reducing model complexity. By eliminating redundant and weakly informative attributes, the binary BER formulation achieved lower predictive error and improved fitness stability relative to competing binary optimizers. Dimensionality reduction was achieved without sacrificing explanatory power, and improvements were observed consistently across diverse learning architectures. The reduction in input dimensionality not only lowered residual dispersion but also improved bias characteristics and strengthened agreement metrics. These results suggest that, for the evaluated dataset, structured feature subset optimization sharpens the functional mapping between input variables and target outcomes, thereby improving both statistical efficiency and computational tractability.

The hyperparameter optimization stage further amplified predictive performance. BER-driven tuning produced the lowest error metrics, highest correlation coefficients, strongest variance explanation, and highest agreement indices among all evaluated metaheuristic algorithms. Under the controlled experimental conditions used in this study, these improvements were systematic and observed across every reported metric. The substantial contraction in MSE, RMSE, and MAE, coupled with near-zero bias and elevated R^2, NSE, and WI values, demonstrates that hyperparameter calibration significantly refines model calibration and generalization capacity within the internal evaluation framework. The results therefore support the view that adaptive exploration–exploitation mechanisms embedded in BER effectively navigate complex continuous search spaces associated with FTLM hyperparameters.

Taken together, the integrated application of BER for both feature selection and hyperparameter optimization establishes a coherent optimization strategy capable of enhancing accuracy, stability, and computational efficiency. For the present dataset and evaluation setting, BER consistently outperformed the compared optimizers across both discrete and continuous optimization stages. Importantly, the improvements observed at each stage are cumulative: feature selection improves structural clarity of the input space, while hyperparameter optimization refines internal learning dynamics. The combined effect yields a predictive framework that is not only accurate but also robust and reproducible under controlled experimental conditions.

From a methodological perspective, the study highlights the importance of layered optimization in applied machine learning for healthcare. Isolated tuning of model parameters without prior dimensionality control may lead to unstable convergence, whereas feature selection alone may not fully exploit model capacity. The proposed pipeline demonstrates that coordinated optimization across representation and parameter spaces can produce synergistic gains. Moreover, the strict enforcement of equal computational budgets and reproducibility controls strengthens the credibility of comparative findings and provides a transparent benchmarking template for future research.

Future research should build first on the most immediate limitations of the present study. The most important next step is external validation using independent datasets from different institutions to assess generalization robustness under varying demographic, temporal, and operational conditions. Such validation would provide stronger evidence on whether the selected feature subsets and optimized hyperparameter configurations remain effective under distribution shift and would clarify the translational scope of the proposed framework.

A second realistic direction is to extend the current single-objective optimization setting toward multi-objective formulations that jointly consider predictive accuracy, computational cost, model compactness, and interpretability. This would be particularly relevant in resource-constrained clinical environments, where deployment decisions often require balancing performance against efficiency and implementation cost.

A third technically grounded direction is to investigate explainability and robustness more explicitly. In particular, future work may examine how the selected feature subsets contribute to predictive behavior, whether the optimized configurations remain stable across different data splits or temporal cohorts, and how sensitive the framework is to changes in sampling patterns, missingness, or event prevalence. These analyses would improve both interpretability and confidence in practical use.

Longer-term research may explore integration into real-time monitoring systems, but such deployment-oriented extensions should be viewed as a subsequent step rather than an immediate conclusion from the present experiments. Embedding the optimized FTLM within streaming clinical pipelines would require additional study of computational latency, model update frequency, calibration under live data streams, and implementation-level reliability. Similarly, hybrid optimization strategies, adaptive stopping criteria, and re-optimization under temporal drift may be explored once broader external validation has been established.

In conclusion, the results indicate that the BER-assisted feature-selection and hyperparameter-optimization framework is an effective strategy for improving FTLM performance on structured hospital monitoring data within the present dataset and controlled experimental setting. The methodological rigor, reproducibility controls, and comprehensive evaluation metrics provide a solid basis for future work toward external validation, multi-objective optimization, and deployment-oriented predictive healthcare systems. However, broader clinical generalization should be interpreted cautiously until the framework is validated on independent datasets and assessed under real-world implementation conditions.

8 Acknowledgment

Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2026R716), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

9 Data Availability Statement

The data used in this study are publicly available at: (1) https://www.kaggle.com/datasets/parmajha/patient-vital-signs-and-event-tracking/data. (2) https://www.kaggle.com/datasets/nasirayub2/human-vital-sign-dataset/data

10 Declarations

Author Contributions

The authors confirm contribution to the paper as follows: Conceptualization, S.A.A. and A.A.A.; methodology, S.K.T. ; software, S.K.T. ; validation, M.M.E. ; formal analysis, S.A.A; investigation, A.A.A.; resources, M.M.E.; data curation, S.A.A. ; writing—original draft preparation, M.M.E and A.A.A; writing—review and editing, A.A.A.; visualization, S.A.A. ; supervision, S.K.T. ; project administration, S.A.K ; All authors reviewed the results and approved the final version of the manuscript.

Funding
No Fund

Competing interests
The authors declare no competing interests.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Conflict of interest
All authors declare no competing interests.

Appendix A External Validation on an Independent Human Vital Signs Dataset

This appendix presents an external validation and robustness analysis of the proposed BER–FTLM framework using an independent structured vital-sign dataset. The purpose of this analysis is to examine whether the empirical behavior of the proposed optimization pipeline is preserved when transferred beyond the original Patient Vital Signs and Event Tracking dataset to a separate, structurally related human vital-sign dataset. Accordingly, this appendix should be interpreted as an additional experimental validation study rather than as the introduction of a new optimizer, a new prediction model, or a formal theoretical contribution.

The independent validation experiment uses the Human Vital Sign Dataset, publicly available on Kaggle, which contains structured physiological measurements intended for medical diagnostics, patient monitoring, and predictive analytics. The dataset includes both original physiological and demographic attributes and derived health indicators, including heart-rate variability, pulse pressure, body mass index, and mean arterial pressure. These characteristics make it suitable for evaluating the robustness of the proposed pipeline on a second vital-sign modeling problem with related clinical structure but distinct data composition.

A.1 Motivation for External Validation

External validation is essential for determining whether a predictive modeling pipeline captures generalizable structure or primarily adapts to the statistical properties of a single dataset. Although the main experiments demonstrate the effectiveness of the BER–FTLM framework on the Patient Vital Signs and Event Tracking dataset, validation on only one dataset is insufficient for establishing broader empirical robustness. Therefore, an additional experiment was conducted on an independent human vital-sign dataset to assess whether the proposed BER-assisted optimization pipeline maintains its performance pattern under a different structured vital-sign data source.

The external dataset contains 200,000 samples and 17 columns, including heart rate, respiratory rate, timestamp, body temperature, oxygen saturation, systolic blood pressure, diastolic blood pressure, age, gender, weight, height, derived heart-rate variability, derived pulse pressure, derived body mass index, derived mean arterial pressure, patient identifier, and risk category. The inclusion of both direct physiological measurements and derived indicators provides a multivariate clinical representation that is appropriate for evaluating feature interaction learning, predictive stability, and optimizer behavior in a structured vital-sign modeling context.

The goal of this external validation is not to claim a new theoretical or algorithmic contribution. Rather, it evaluates whether the proposed BER–FTLM framework, which combines BER-based optimization with FTLM prediction, retains its empirical effectiveness when applied to a second dataset. This distinction is important because the contribution of the present work lies in the reproducible integration, benchmarking, and validation of a unified optimization-aware modeling pipeline, rather than in the formal invention of a new optimizer or a new convergence theory.

For this reason, the interpretation of the comparative optimization results is stated in empirical terms. BER is not presented as universally superior to all optimization methods. Instead, the external validation analysis evaluates whether BER shows the strongest empirical performance among the evaluated optimizers under the same experimental setting. This formulation preserves the comparative value of the results while avoiding unsupported general claims beyond the evaluated datasets, models, metrics, and computational conditions.

A.2 Independent Dataset Description

The external validation dataset was obtained from Kaggle: the Human Vital Sign Dataset, also described as the human vital dataset 2024. The dataset is released under the CC0: Public Domain license and contains the file `human_vital_signs_dataset_2024.csv`. It was designed to support research in medical diagnostics, patient monitoring, and predictive analytics by providing structured physiological measurements together with derived health indicators. Available at: https://www.kaggle.com/datasets/nasirayub2/human-vital-sign-dataset/data

The dataset contains 200,000 samples and 17 columns, providing a large-scale structured vital-sign dataset suitable for evaluating the generalization behavior of the proposed optimization-aware prediction pipeline. Its structure is clinically relevant because it combines direct physiological measurements, demographic variables, anthropometric variables, derived cardiovascular and body-composition indicators, and a categorical risk label within a single tabular representation. This makes the dataset appropriate for assessing whether the BER–FTLM framework remains empirically effective when transferred to a second vital-sign dataset with a different data source and a different target definition.

In addition to the original variables, the dataset contains derived physiological and anthropometric indicators. These derived features provide clinically interpretable transformations of the raw measurements and support multivariate modeling by encoding relationships that are not directly represented by individual variables alone. The derived variables are summarized in Table A2.

The target variable is Risk Category, with two categories: High Risk and Low Risk. High-risk cases are defined by abnormal vital-sign or body-composition thresholds, including heart rate greater than 90 bpm or less than 60 bpm, respiratory rate greater than 20 breaths per minute or less than 12 breaths per minute, body temperature greater than 37.5°C or less than 36.0°C, oxygen saturation less than 95%, systolic blood pressure greater than 140 mmHg or less than 110 mmHg, diastolic blood pressure greater than 90 mmHg or less than 70 mmHg, or body mass index greater than 30 or less than 18.5. Low-risk cases are those that do not satisfy any of these high-risk criteria. This target definition provides a clinically interpretable risk-stratification label and enables the external experiment to evaluate whether the proposed pipeline can model structured physiological risk patterns beyond the original dataset.

Table A1 Original attributes in the external Human Vital Sign Dataset.

Attribute	Description
Patient ID	Unique patient identifier.
Heart Rate	Number of heartbeats per minute, with values ranging from 60 to 100 bpm in this dataset.
Respiratory Rate	Number of breaths per minute, with values ranging from 12 to 20 breaths per minute in this dataset.
Timestamp	Time at which the vital signs were recorded.
Body Temperature	Body temperature measured in degrees Celsius, with values ranging from 36.0 to 37.5°C in this dataset.
Oxygen Saturation	Peripheral oxygen saturation percentage, with values ranging from 95% to 100% in this dataset.
Systolic Blood Pressure	Systolic arterial pressure, with values ranging from 110 to 140 mmHg in this dataset.
Diastolic Blood Pressure	Diastolic arterial pressure, with values ranging from 70 to 90 mmHg in this dataset.
Age	Patient age, with values ranging from 18 to 90 years in this dataset.
Gender	Categorical patient sex variable with Male and Female categories.
Weight (kg)	Patient weight in kilograms.
Height (m)	Patient height in meters.

Table A2 Derived features in the external Human Vital Sign Dataset.

Derived feature	Formula or meaning
`Derived_HRV`	Standard deviation of heart rate over a period divided by the mean heart rate over the same period.
`Derived_Pulse_Pressure`	Systolic Blood Pressure − Diastolic Blood Pressure.
`Derived_BMI`	Weight divided by squared height, i.e., Weight (kg) / Height $(\mathrm{m})^2$.
`Derived_MAP`	Diastolic Blood Pressure $+\frac{1}{3}\times$ (Systolic Blood Pressure − Diastolic Blood Pressure).

A.3 External Dataset Exploratory Analysis

Exploratory analysis was conducted on the external Human Vital Sign Dataset to examine whether the dataset contains physiologically meaningful structure before applying baseline modeling and BER-based optimization. The analysis focuses on heart-rate distribution by risk category, heart-rate behavior across gender and risk strata, normalized multivariate patient profiles, and the distribution of derived heart-rate variability. These visual diagnostics provide evidence that the dataset contains both strong single-variable signals and weaker derived-feature patterns, thereby motivating the use of multivariate learning rather than reliance on isolated threshold rules.

Figure A1 supports the clinical plausibility of the external dataset because high-risk patients exhibit higher heart-rate values than low-risk patients. This pattern is consistent with the target definition, in which elevated heart rate is one of the abnormal physiological conditions contributing to high-risk classification. The figure also supports the use of vital-sign-based modeling because the risk groups are not

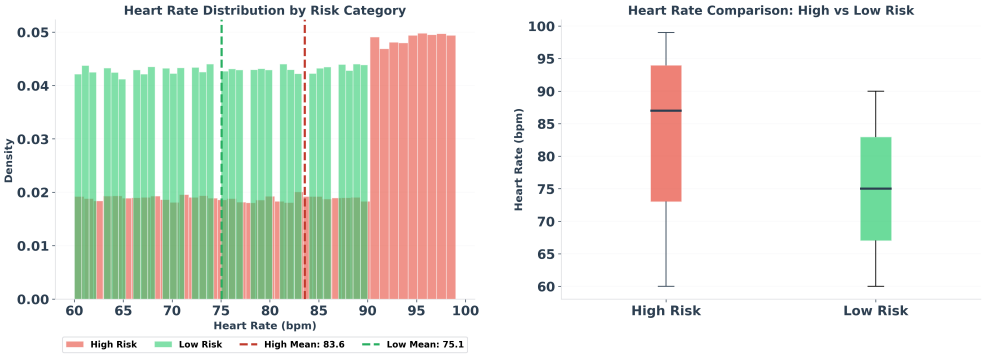

Fig. A1 Heart rate distribution by risk category in the external Human Vital Sign Dataset. The left panel shows the density distribution of heart rate for high-risk and low-risk groups, while the right panel compares heart-rate dispersion between the two risk categories using boxplots. The high-risk group shows a higher mean heart rate, approximately 83.6 bpm, compared with the low-risk group, approximately 75.1 bpm, indicating that heart-rate elevation contributes meaningfully to risk stratification in the external dataset.

arbitrary labels; rather, they reflect measurable differences in physiological parameters. At the same time, the visible overlap between the two groups indicates that heart rate alone is not sufficient to characterize all risk behavior, which motivates multivariate modeling using the full set of physiological, demographic, and derived features.

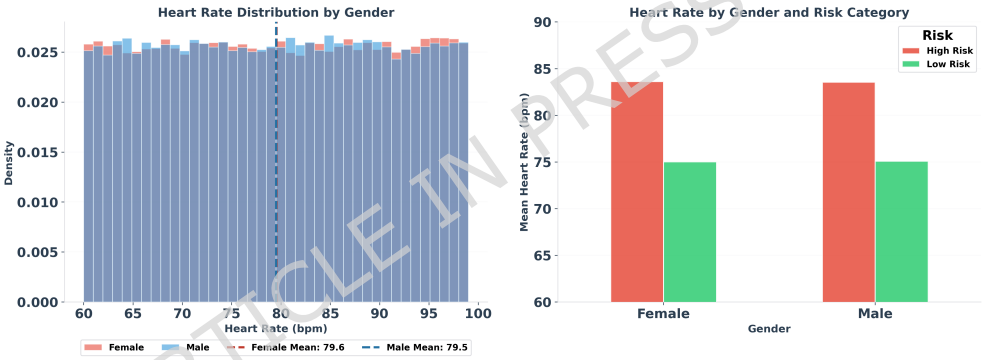

Fig. A2 Heart rate distribution by gender and risk category. The left panel compares heart-rate density distributions for male and female patients, while the right panel reports mean heart rate stratified by gender and risk category. Mean heart rate is nearly identical between female and male patients, approximately 79.6 bpm and 79.5 bpm, respectively. However, within both genders, the high-risk subgroup shows higher mean heart rate than the low-risk subgroup.

Figure A2 indicates that the overall heart-rate distribution is highly similar across male and female patients, with nearly identical mean values. This suggests that the heart-rate difference associated with risk category is not primarily driven by gender imbalance. Within both gender groups, the high-risk subgroup demonstrates a higher mean heart rate than the low-risk subgroup, indicating that the risk-related signal is expressed within demographic strata rather than only through demographic separation. This observation supports the interpretation that the predictive models are expected to learn risk-relevant physiological structure rather than relying solely on gender-based stratification.

Figure A3 provides a compact multivariate profile of the external dataset by comparing normalized physiological and derived indicators across high-risk and low-risk groups. The radar chart shows that risk representation is distributed across several

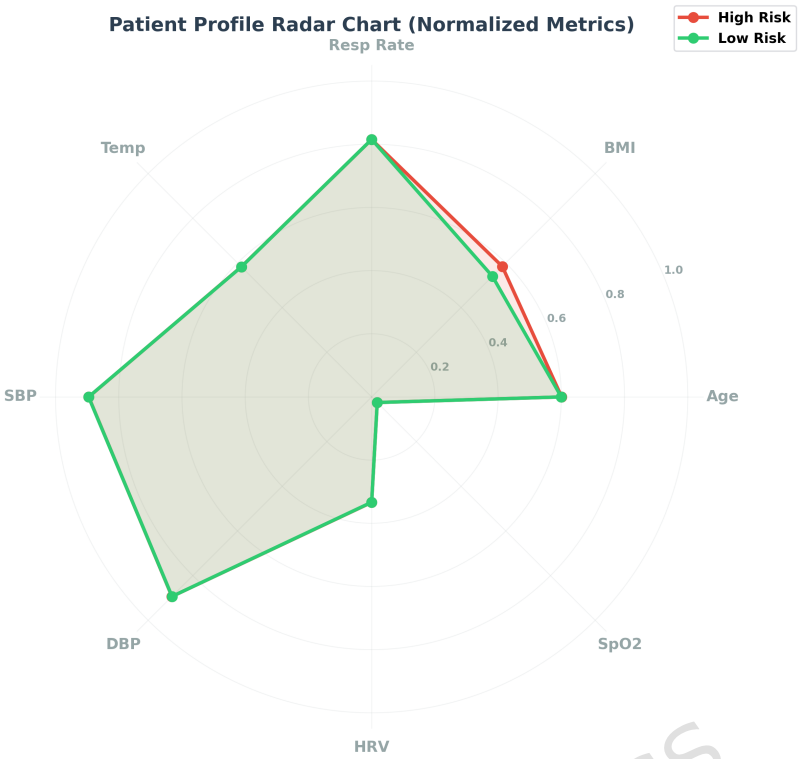

Fig. A3 Normalized patient-profile radar chart comparing high-risk and low-risk groups. The radar chart summarizes normalized values for age, BMI, respiratory rate, temperature, systolic blood pressure, diastolic blood pressure, HRV, and SpO_2. The visualization provides a compact multivariate comparison of risk-group profiles and illustrates that the external dataset contains multi-parameter physiological structure rather than a single-variable risk signal.

variables rather than being reducible to a single measurement. This is important because the proposed modeling framework is designed for structured vital-sign data in which physiological variables, demographic descriptors, and derived indicators may interact. The observed multi-parameter structure supports the suitability of the external dataset for evaluating the proposed BER–FTLM pipeline and aligns with the methodological premise that structured vital-sign prediction should preserve cross-variable dependencies instead of relying only on isolated thresholds.

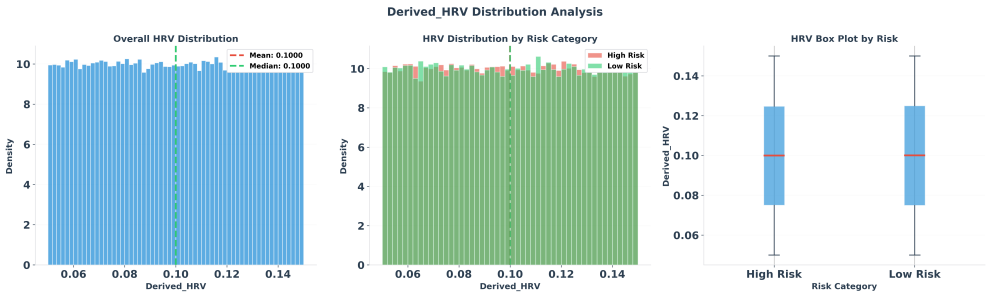

Fig. A4 `Derived_HRV` distribution analysis in the external dataset. The figure reports the overall `Derived_HRV` distribution, the `Derived_HRV` distribution by risk category, and a boxplot comparison between high-risk and low-risk groups. The mean and median `Derived_HRV` are both approximately 0.1000, and the risk-category boxplots show highly similar distributions.

Figure A4 shows that `Derived_HRV` has a stable overall distribution, with both the mean and median close to 0.1000. The high-risk and low-risk boxplots show highly similar distributions, indicating that `Derived_HRV` alone may not strongly separate the two risk categories in this dataset. This finding is useful because it demonstrates that not every derived feature provides strong marginal discrimination. Therefore, the external dataset is an appropriate test case for feature interaction learning and optimizer-guided modeling: strong prediction should arise from the combined structure of multiple physiological and derived variables rather than from a single-feature rule.

A.4 Baseline Model Performance on the External Dataset

To establish a non-optimized reference point on the independent Human Vital Sign Dataset, the same baseline learning models were evaluated under the external validation setting. This baseline stage is important because it determines whether the model-ranking pattern observed in the main experiments is preserved when the framework is transferred to a separate structured vital-sign dataset. The results are reported in Table A3.

Table A3 Baseline performance of learning models on the external Human Vital Sign Dataset.

Models	MSE	RMSE	MAE	MBE	r	R^2	RRMSE	NSE	WI
FTLM	0.0478	0.2186	0.1423	0.0952	0.949	0.900	4.25	0.908	0.910
CTSM	0.1025	0.3202	0.2014	0.1346	0.944	0.891	6.14	0.895	0.907
VAST	0.1587	0.3984	0.2498	0.1685	0.939	0.882	6.45	0.883	0.896
LSTM	0.1843	0.4293	0.2716	0.1927	0.934	0.872	7.24	0.874	0.887
DTCN	0.3521	0.5934	0.3815	0.2542	0.930	0.865	8.15	0.861	0.873
TST	0.6124	0.7826	0.4987	0.3315	0.927	0.859	8.57	0.850	0.860
VAE	0.9410	0.9701	0.6213	0.4186	0.926	0.857	9.50	0.842	0.852

The baseline evaluation on the independent dataset shows that FTLM remains the strongest non-optimized model, achieving the lowest MSE (0.0478), RMSE (0.2186), MAE (0.1423), MBE (0.0952), and RRMSE (4.25), while maintaining strong correlation ($r = 0.949$) and agreement (WI = 0.910). CTSM achieves the next-best baseline profile, with MSE = 0.1025, RMSE = 0.3202, $r = 0.944$, and WI = 0.907, indicating that contextual structured modeling remains competitive but does not exceed FTLM in error-based performance. VAST and LSTM show intermediate performance, whereas DTCN, TST, and VAE exhibit progressively larger errors and weaker agreement indices.

This baseline pattern supports the use of FTLM as the base learner for the subsequent optimization stage on the external dataset. More importantly, it demonstrates that the baseline model-ranking behavior observed in the original dataset is broadly reproducible in the independent dataset: FTLM remains the most favorable non-optimized configuration, while the remaining architectures follow a consistent degradation pattern across magnitude-based, correlation-based, and agreement-based metrics.

Figure A5 provides a metric-wise visualization of the baseline model comparison. The error-based panels show that FTLM has the smallest values for MSE, RMSE, MAE, MBE, and RRMSE, whereas the agreement-oriented panels show that FTLM remains highly competitive in r, R^2, NSE, and WI. The visual pattern confirms that FTLM is not favored by only one metric, but maintains a balanced baseline profile across the full evaluation suite.

Figure A6 summarizes the distribution of baseline metric values across models using horizontal boxplots and swarm overlays. The figure shows that error metrics have

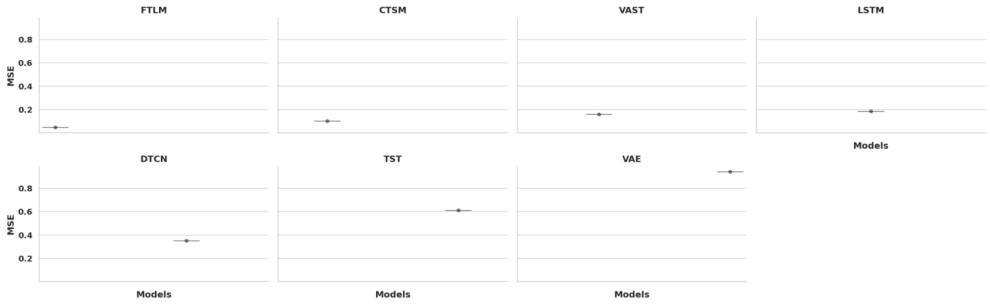

Fig. A5 Facet grid of baseline model performance metrics on the external dataset.

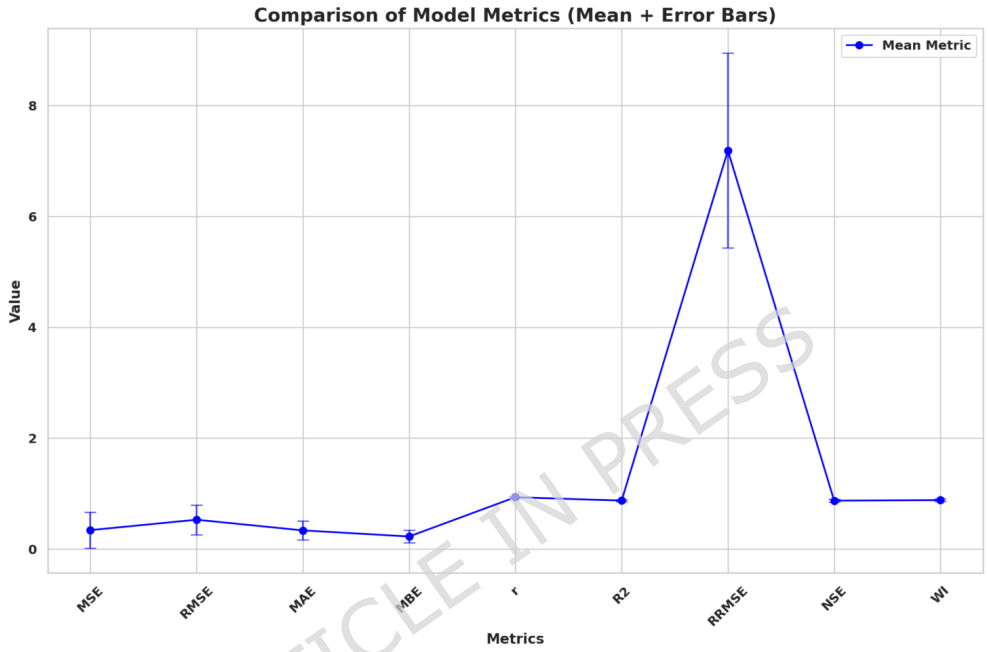

Fig. A6 Box plot with horizontal swarm overlay for baseline model metrics.

wider dispersion across model families, particularly for MSE, RMSE, MAE, MBE, and RRMSE, while correlation- and agreement-based metrics are concentrated within a narrower range. This indicates that baseline models differ most strongly in magnitude accuracy, even though most models preserve relatively high association with the target structure.

Figure A7 isolates the MSE behavior of each baseline learner. The faceted display clearly shows the progressive increase in MSE from FTLM to CTSM, VAST, LSTM, DTCN, TST, and VAE. This supports the tabulated results by showing that FTLM provides the lowest baseline squared-error magnitude, whereas VAE produces the largest MSE among the non-optimized models.

Figure A8 provides an aggregate view of baseline metric behavior using mean values and error bars. The large scale of RRMSE relative to the remaining metrics highlights the importance of considering scale-normalized error separately from bounded agreement measures. The figure also illustrates that the baseline evaluation contains both low-range error metrics and high-range association metrics, reinforcing the need for a multi-metric assessment rather than reliance on a single performance criterion.

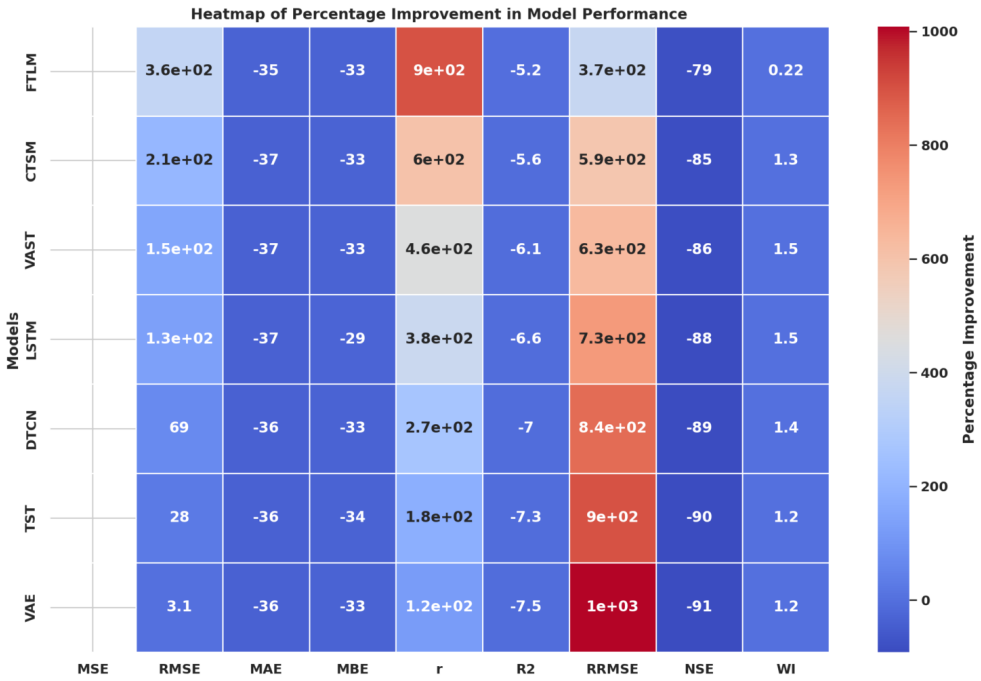

Fig. A7 Faceted MSE comparison across baseline learning models.

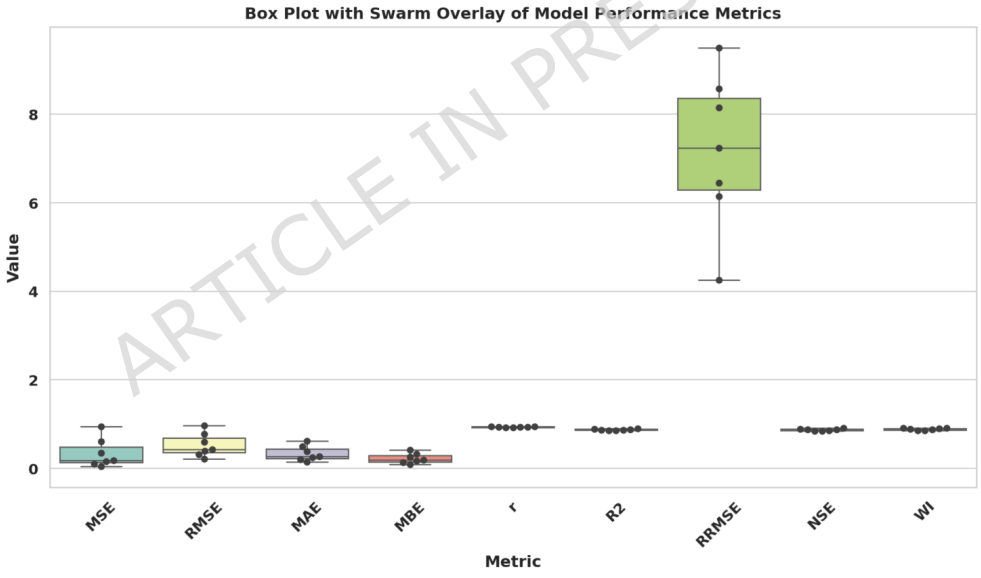

Fig. A8 Mean metric comparison with error bars across baseline metrics.

A.5 BER-Based Hyperparameter Optimization on the External Dataset

After establishing FTLM as the strongest baseline learner on the external Human Vital Sign Dataset, a second stage of evaluation was conducted to examine the effect of optimizer choice on the FTLM configuration. The goal of this stage is to determine whether the empirical optimization pattern observed in the primary experiments is

preserved on an independent dataset. The comparative results for BER, GWO, PSO, BA, WAO, SBO, and SCA are reported in Table A4.

Table A4 Hyperparameter optimization results on the external Human Vital Sign Dataset.

Models	MSE	RMSE	MAE	MBE	r	R2	RRMSE	NSE	WI
BER + FTLM	0.00023	0.01517	0.00062	0.00041	0.989	0.978	1.52	0.976	0.981
GWO + FTLM	0.00028	0.01673	0.00074	0.00053	0.986	0.972	1.67	0.969	0.975
PSO + FTLM	0.00032	0.01789	0.00082	0.00060	0.984	0.968	1.79	0.965	0.972
BA + FTLM	0.00036	0.01897	0.00089	0.00066	0.982	0.964	1.90	0.961	0.969
WAO + FTLM	0.00041	0.02024	0.00095	0.00072	0.979	0.958	2.02	0.956	0.964
SBO + FTLM	0.00046	0.02145	0.00102	0.00079	0.977	0.954	2.14	0.951	0.960
SCA + FTLM	0.00052	0.02280	0.00110	0.00086	0.974	0.949	2.28	0.947	0.956

On the external dataset, BER + FTLM achieved the lowest MSE (0.00023), RMSE (0.01517), MAE (0.00062), MBE (0.00041), and RRMSE (1.52), while also obtaining the highest r (0.989), R2 (0.978), NSE (0.976), and WI (0.981). GWO + FTLM and PSO + FTLM provide the next strongest optimized configurations, while BA + FTLM, WAO + FTLM, SBO + FTLM, and SCA + FTLM show a gradual decline across the same performance criteria. These results reproduce the main empirical pattern observed in the original experimental setting: BER provides the strongest empirical configuration for FTLM under the evaluated optimization setting on the external Human Vital Sign Dataset.

At the same time, the interpretation of these results remains empirical and dataset-specific. The observed ranking indicates that BER is a stable and effective optimizer for this structured vital-sign modeling setting, but it does not establish universal optimizer dominance across all learning tasks, data distributions, or optimization scenarios. The primary significance of Table A4 is that the relative advantage of BER is reproduced on an independent dataset, thereby strengthening the robustness of the experimental findings.

Figure A9 presents a metric-wise view of the optimized FTLM configurations. Across all error metrics, BER + FTLM occupies the most favorable position, followed by a largely monotonic progression through GWO + FTLM, PSO + FTLM, BA + FTLM, WAO + FTLM, SBO + FTLM, and SCA + FTLM. A complementary reverse ordering is visible for the agreement-based metrics, where BER + FTLM achieves the highest values. This figure confirms that the advantage of BER is consistent across multiple criteria rather than being concentrated in a single metric.

Figure A10 summarizes the distribution of optimized metric values across all evaluated optimizers. The plot shows that the optimized models are relatively tightly grouped for bounded association metrics such as r, R2, NSE, and WI, whereas the spread is somewhat larger for RRMSE and the small-scale error metrics. This pattern indicates that, although all evaluated optimizers yield strong FTLM performance, optimizer choice still affects the absolute magnitude of the residual errors and relative-error behavior.

Figure A11 provides a faceted comparison of each optimized metric across the seven optimizer-enhanced FTLM models. The figure makes the ranking particularly transparent: BER + FTLM occupies the most favorable position in every panel, and the remaining optimizers follow a regular deterioration pattern. The visual consistency across the metric panels supports the interpretation that the observed optimizer ranking is systematic under the current external validation setting.

Figure A12 offers a compact tabular visualization of the optimized performance metrics. The heatmap clearly shows that BER + FTLM combines the smallest error

Fig. A9 Facet grid of optimized FTLM performance metrics across optimizers.

values with the strongest agreement and association indices. The ordered gradient across rows also indicates that the differences between optimizers are not random fluctuations but follow a coherent comparative structure across the full metric set.

Figure A13 shows Q–Q plots for the optimized metric values. The points align closely with the fitted reference lines for all reported metrics, reflecting a regular ordered progression of optimizer outcomes across the evaluated configurations. In the present context, the importance of this figure is descriptive: it shows that the metric values vary smoothly across optimizers and do not exhibit abrupt irregularities that would contradict the comparative pattern reported in Table A4.

Figure A14 illustrates pairwise relationships among the optimized performance metrics. Strong linear trends are visible between error-based measures, and inverse relationships appear between error measures and agreement-based metrics. This pairwise structure reinforces the internal consistency of the evaluation framework: optimizers that reduce residual error also improve correlation and agreement metrics, with BER + FTLM occupying the strongest overall position in that joint metric space.

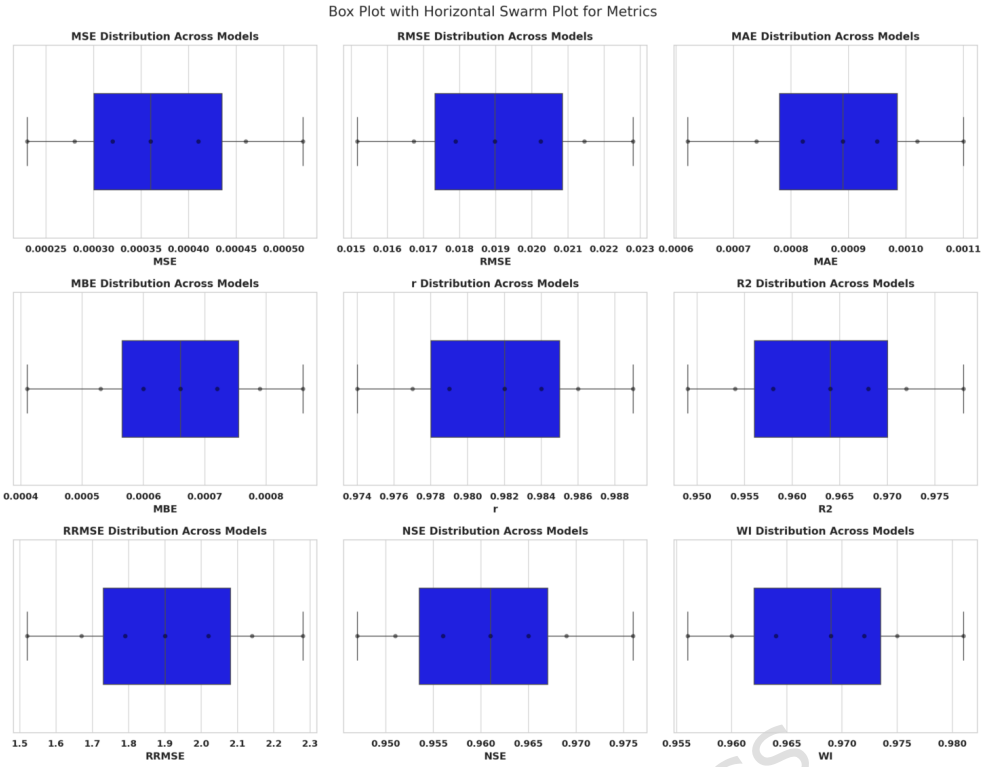

Fig. A10 Box plot with horizontal swarm plot for optimized metric distributions.

References

Agrawal, U.K., Panda, N.: Quantum-inspired adaptive mutation operator enabled pso (qamo-pso) for parallel optimization and tailoring parameters of kolmogorov–arnold network. The Journal of Supercomputing **81**(14), 1310 (2025) https://doi.org/10.1007/s11227-025-07810-w

Agrawal, U.K., Panda, N.: Quantum tunneling-inspired salp swarm algorithm (qt-ssa) for resilient coalition of point clouds. In: 2025 13th International Conference on Intelligent Systems and Embedded Design (ISED), pp. 19–24 (2025). https://doi.org/10.1109/ISED67359.2025.11405178

Agrawal, U.K., Panda, N., Tejani, G.G., Mousavirad, S.J.: Improved salp swarm algorithm-driven deep cnn for brain tumor analysis. Scientific Reports **15**(1), 24645 (2025) https://doi.org/10.1038/s41598-025-09326-y

Bergstra, J., Bengio, Y.: Random search for hyper-parameter optimization. Journal of Machine Learning Research **13**, 281–305 (2012)

Bergstra, J., Bengio, Y.: Random search for hyper-parameter optimization. Journal of Machine Learning Research **13**, 281–305 (2012)

Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A.: A review of feature selection methods on synthetic data. Knowledge and Information Systems **34**(3), 483–519 (2013) https://doi.org/10.1007/s10115-012-0487-8

Brunker, C., Harris, R.: How accurate is the AVPU scale in detecting neurological

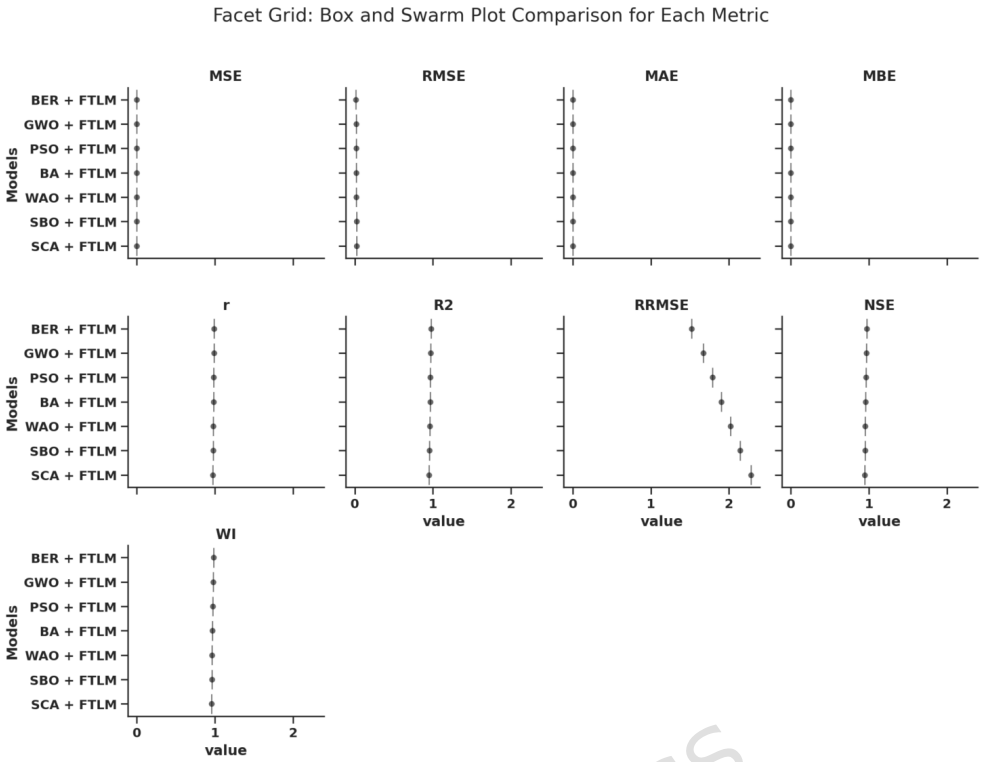

Fig. A11 Faceted box-and-swarm comparison for each optimized metric.

impairment when used by general ward nurses? an evaluation study using simulation and a questionnaire. Intensive and Critical Care Nursing **31**(2), 69–75 (2015) https://doi.org/10.1016/j.iccn.2014.11.003

Bai, S., Kolter, J.Z., Koltun, V.: An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271 (2018)

Churpek, M.M., Adhikari, R., Edelson, D.P.: The value of vital sign trends for detecting clinical deterioration on the wards. Resuscitation **102**, 1–5 (2016) https://doi.org/10.1016/j.resuscitation.2016.02.005

Díaz-Uriarte, R., Andrés, S.: Gene selection and classification of microarray data using random forest. BMC Bioinformatics **7**, 3 (2006) https://doi.org/10.1186/1471-2105-7-3

Escobar, G.J., Greene, J.D., Scheirer, P., Gardner, M.N., Draper, D., Kipnis, P.: Risk-adjusting hospital inpatient mortality using automated inpatient, outpatient, and laboratory databases. Medical Care **46**(3), 232–239 (2008) https://doi.org/10.1097/MLR.0b013e3181589bb6

El-kenawy, E.-S.M., Abdelhamid, A.A., Ibrahim, A., Mirjalili, S., Khodadad, N., Al duailij, M.A., *et al.*: Al-biruni earth radius (BER) metaheuristic search optimization algorithm. Computer Systems Science and Engineering **45**(2), 1917–1934 (2023) https://doi.org/10.32604/csse.2023.032497

Escobar, G.J., Liu, V.X., Schuler, A., Lawson, B., Greene, J.D., Kipnis, P.: Automated identification of adults at risk for in-hospital clinical deterioration. The New

Fig. A12 Heatmap of optimized model metric comparison.

England Journal of Medicine **383**(20), 1951–1960 (2020) https://doi.org/10.1056/NEJMsa2001090

Elsken, T., Metzen, J.H., Hutter, F.: Neural architecture search: A survey. Journal of Machine Learning Research **20**(55), 1–21 (2019)

Emary, E., Zawbaa, H.M., Hassanien, A.E.: Binary grey wolf optimization approaches for feature selection. Neurocomputing **172**, 371–381 (2016) https://doi.org/10.1016/j.neucom.2015.06.083

Fu, Y., Liu, D., Chen, J., He, L.: Secretary bird optimization algorithm: A new metaheuristic for solving global optimization problems. Artificial Intelligence Review (2024) https://doi.org/10.1007/s10462-024-10729-y

Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, ??? (2016). https://www.deeplearningbook.org/

Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of Machine Learning Research **3**, 1157–1182 (2003)

Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of Machine Learning Research **3**, 1157–1182 (2003)

Ghassemi, M., Pimentel, M.A.F., Naumann, T., Brennan, T., Clifton, D.A., Szolovits,

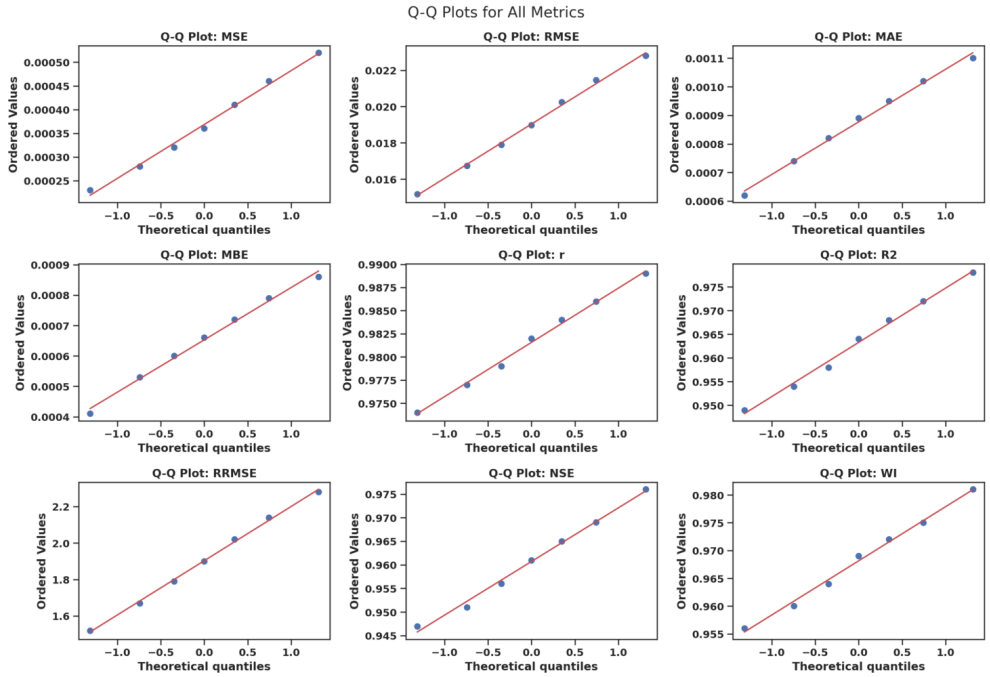

Fig. A13 Q–Q plots for all optimized performance metrics.

P., Feng, M., Celi, L.A.: A multivariate timeseries modeling approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15) (2015). https://ojs.aaai.org/index.php/AAAI/article/view/9209

Gorishniy, Y., Rubachev, I., Khrulkov, V., Babenko, A.: Revisiting deep learning models for tabular data. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 18932–18943 (2021). https://proceedings.neurips.cc/paper/2021/hash/9d86d83f925f2149e9edb0ac3b49229c-Abstract.html

Huang, X., Khetan, A., Cvitkovic, M., Karnin, Z.: Tabtransformer: Tabular data modeling using contextual embeddings. arXiv preprint arXiv:2012.06678 (2020)

Holland, J.H.: Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence. University of Michigan Press, Ann Arbor, MI (1975)

Hussein, N.K., Qaraad, M., El Najjar, A.M., Farag, M.A., Elhosseini, M.A., Mirjalili, S., Guinovart, D.: Schrödinger optimizer: A quantum duality-driven metaheuristic for stochastic optimization and engineering challenges. Knowledge-Based Systems **328**, 114273 (2025) https://doi.org/10.1016/j.knosys.2025.114273

Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997) https://doi.org/10.1162/neco.1997.9.8.1735

Ibrahim, M.Q., Qaraad, M., Hussein, N.K., Farag, M.A., Guinovart, D.: Secant optimization algorithm for efficient global optimization. Scientific Reports **16**(1), 6659 (2026) https://doi.org/10.1038/s41598-026-36691-z

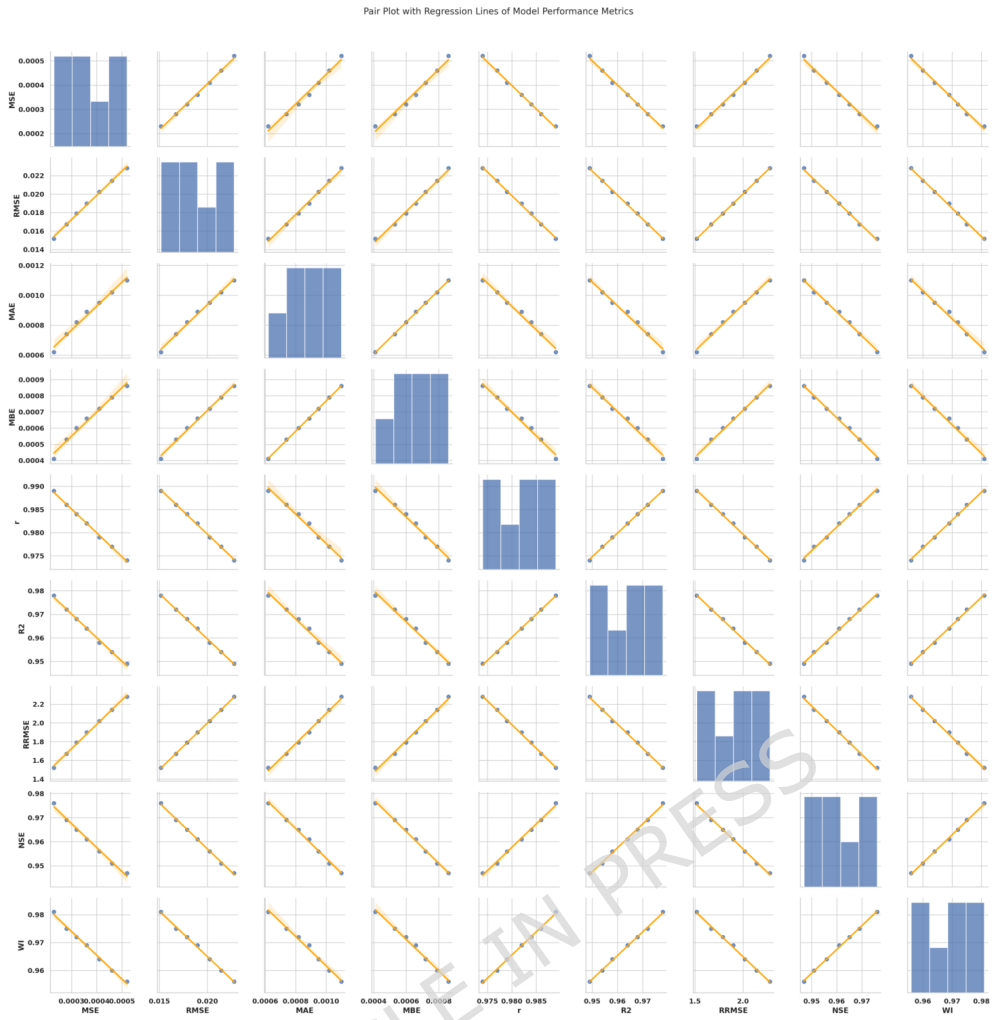

Fig. A14 Pair plot with regression lines for optimized model performance metrics.

Johnson, A.E.W., Pollard, T.J., Shen, L., Lehman, L.-W.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G.: MIMIC-III, a freely accessible critical care database. Scientific Data **3**, 160035 (2016) https://doi.org/10.1038/sdata.2016.35

Kennedy, J., Eberhart, R.: Particle swarm optimization. In: Proceedings of the IEEE International Conference on Neural Networks, vol. 4, pp. 1942–1948 (1995). https://doi.org/10.1109/ICNN.1995.488968

Koh, B.H.D., Lim, C.L.P., Rahimi, H., Woo, W.L., Gao, B.: Deep temporal convolution network for time series classification. Sensors **21**(2), 603 (2021) https://doi.org/10.3390/s21020603

Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: International Conference on Learning Representations (ICLR) (2014). https://arxiv.org/abs/1312.6114

LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**, 436–444 (2015) https://doi.org/10.1038/nature14539

Lipton, Z.C., Kale, D.C., Elkan, C., Wetzel, R.: Learning to diagnose with LSTM recurrent neural networks. arXiv (2016) 1511.03677

Moosavi, S.H.S., Bardsiri, V.K.: Satin bowerbird optimizer: A new optimization algorithm to optimize anfis for software development effort estimation. Engineering Applications of Artificial Intelligence **60**, 1–15 (2017) https://doi.org/10.1016/j.engappai.2017.01.006

Mirjalili, S.: Sca: A sine cosine algorithm for solving optimization problems. Knowledge-Based Systems **96**, 120–133 (2016) https://doi.org/10.1016/j.knosys.2015.12.022

Mirjalili, S., Lewis, A.: The whale optimization algorithm. Advances in Engineering Software **95**, 51–67 (2016) https://doi.org/10.1016/j.advengsoft.2016.01.008

Mirjalili, S., Mirjalili, S.M., Lewis, A.: Grey wolf optimizer. Advances in Engineering Software **69**, 46–61 (2014) https://doi.org/10.1016/j.advengsoft.2013.12.007

ParmaJha: Patient Vital Signs and Event Tracking. Kaggle dataset. Accessed: 2026-02-20 (2024). https://www.kaggle.com/datasets/parmajha/patient-vital-signs-and-event-tracking

Qaraad, M., Amjad, S., Hussein, N.K., Elhosseini, M.A.: An innovative quadratic interpolation salp swarm-based local escape operator for large-scale global optimization problems and feature selection. Neural Computing and Applications **34**(20), 17663–17721 (2022) https://doi.org/10.1007/s00521-022-07391-2

Qaraad, M., Amjad, S., Hussein, N.K., Elhosseini, M.A.: Large scale salp-based grey wolf optimization for feature selection and global optimization. Neural Computing and Applications **34**(11), 8989–9014 (2022) https://doi.org/10.1007/s00521-022-06921-2

Qaraad, M., Amjad, S., Manhrawy, I.I.M., Fathi, H., Hassan, B.A., Kafrawy, P.E.: A hybrid feature selection optimization model for high dimension data classification. IEEE Access **9**, 42884–42895 (2021) https://doi.org/10.1109/ACCESS.2021.3065341

Qaraad, M., Crowson, C.S., Guinovart, D.: Gene expression-based diagnosis of primary sjögren's syndrome using a hybrid optimization algorithm with adaptive local search. Biomedical Signal Processing and Control **116**, 109470 (2026) https://doi.org/10.1016/j.bspc.2026.109470

Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schr"oder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Hartig, F., Dormann, C.F.: Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography **40**(8), 913–929 (2017) https://doi.org/10.1111/ecog.02881

Rajkomar, A., Oren, E., Chen, K., Dai, A.M., Hajaj, N., Hardt, M., Liu, P.J., Liu, X., Marcus, J., Sun, M., *et al.*: Scalable and accurate deep learning with electronic health records. npj Digital Medicine **1**(1), 18 (2018) https://doi.org/10.1038/s41746-018-0029-1

Rajkomar, A., Oren, E., Chen, K., Dai, A.M., Hajaj, N., Hardt, M., Liu, P.J., Liu, X., Marcus, J., Sun, M., Sundberg, P., Yee, H., Zhang, K., Zhang, Y., Flores, G., Duggan, G.E., Irvine, J., Le, Q.V., Litsch, K., Mossin, A., Tansuwan, J., Wang, D., Wexler, J., Wilson, J., Ludwig, D., Volchenboum, S.L., Chou, K., Pearson, M.,

Madabushi, S., Shah, N.H., Butte, A.J., Howell, M.D., Cui, C., Corrado, G.S., Dean, J.: Scalable and accurate deep learning with electronic health records. npj Digital Medicine **1**, 18 (2018) https://doi.org/10.1038/s41746-018-0029-1

Smith, M.E.B., Chiovaro, J.C., O'Neil, M., Kansagara, D., Quinones, A.R., Freeman, M., Motu'apuaka, M., Slatore, C.G.: Early warning system scores for clinical deterioration in hospitalized patients: a systematic review. Annals of the American Thoracic Society **11**(9), 1454–1465 (2014) https://doi.org/10.1513/AnnalsATS.201403-102OC

Somepalli, G., Goldblum, M., Schwarzschild, A., Bruss, C.B., Goldstein, T.: Saint: Improved neural networks for tabular data via row attention and contrastive pre-training. arXiv preprint arXiv:2106.01342 (2021)

Sayed, G.I., Hassanien, A.E., Azar, A.T.: A hybrid wrapper–filter feature selection approach based on binary grey wolf optimizer for biomedical data classification. Applied Soft Computing **87**, 105942 (2020) https://doi.org/10.1016/j.asoc.2019.105942

Saeys, Y., Inza, I., Larrañaga, P.: A review of feature selection techniques in bioinformatics. Bioinformatics **23**(19), 2507–2517 (2007) https://doi.org/10.1093/bioinformatics/btm344

Subbe, C.P., Kruger, M., Rutherford, P., Gemmel, L.: Validation of a modified Early Warning Score in medical admissions. QJM: An International Journal of Medicine **94**(10), 521–526 (2001) https://doi.org/10.1093/qjmed/94.10.521

Subbe, C.P., Kruger, M., Rutherford, P., Gemmel, L.: Validation of a modified early warning score in medical admissions. QJM: An International Journal of Medicine **94**(10), 521–526 (2001) https://doi.org/10.1093/qjmed/94.10.521

Snoek, J., Larochelle, H., Adams, R.P.: Practical Bayesian optimization of machine learning algorithms. In: Advances in Neural Information Processing Systems (2012). https://papers.nips.cc/paper_files/paper/2012/hash/05311655a15b75fab86956663e1819cd-Abstract.html

Shickel, B., Tighe, P.J., Bihorac, A., Rashidi, P.: Deep EHR: A survey of recent advances in deep learning techniques for electronic health record analysis. IEEE Journal of Biomedical and Health Informatics **22**(5), 1589–1604 (2018) https://doi.org/10.1109/JBHI.2017.2767063

Too, J., Abdullah, A.R., Mohd Saad, N.: Binary salp swarm algorithm for feature selection. IEEE Access **7**, 140534–140546 (2019) https://doi.org/10.1109/ACCESS.2019.2942665

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS) (2017). https://arxiv.org/abs/1706.03762

Xue, B., Zhang, M., Browne, W.N., Yao, X.: A survey on evolutionary computation approaches to feature selection. IEEE Transactions on Evolutionary Computation **20**(4), 606–626 (2016) https://doi.org/10.1109/TEVC.2015.2504420

Yang, X.-S.: Firefly algorithms for multimodal optimization. In: Stochastic Algorithms: Foundations and Applications. Lecture Notes in Computer Science, pp. 169–178. Springer, ??? (2009). https://doi.org/10.1007/978-3-642-04944-6_14

Yang, X.-S.: Nature-Inspired Metaheuristic Algorithms, 2nd edn. Luniver Press, ??? (2010)

Yang, X.-S.: A new metaheuristic bat-inspired algorithm. In: Nature Inspired Cooperative Strategies for Optimization (NICSO 2010). Studies in Computational Intelligence, vol. 284, pp. 65–74. Springer, ??? (2010). https://doi.org/10.1007/978-3-642-12538-6_6

Young, S.R., Rose, D.C., Karnowski, T.P., Lim, S.-H., Patton, R.M.: Optimizing deep learning hyper-parameters through an evolutionary algorithm. In: Proceedings of the Workshop on Machine Learning in High-Performance Computing Environments (MLHPC '15) (2015). https://doi.org/10.1145/2834892.2834896

Zerveas, G., Jayaraman, S., Patel, D., Bhamidipaty, A., Eickhoff, C.: A transformer-based framework for multivariate time series representation learning. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) (2021). https://doi.org/10.1145/3447548.3467401 . https://dl.acm.org/doi/10.1145/3447548.3467401