HYBRID BO–GA OPTIMIZATION OF XGBOOST FOR CALIBRATED AND ROBUST HYPERTENSION PREDICTION: A MULTI-COHORT VALIDATION STUDY

Manuscript Info

Manuscript History

Received: xxxxxxxxxxxxxxxx
Final Accepted: xxxxxxxxxxxx
Published: xxxxxxxxxxxxxxxx

Key words: -
Hypertension, XGBoost, Bayesian–Genetic Optimization (BO-GA), Probabilistic Calibration, Explainability (SHAP), Cross-Cohort Robustness, NSGA-II, Multi-Cohort Validation

Abstract

Background. Hypertension affects 1.3 billion people worldwide and remains the leading modifiable cause of cardiovascular mortality. Although machine learning models such as XGBoost demonstrate promising discriminative performance, their probabilistic predictions often suffer from poor calibration and limited cross-population robustness, hindering large-scale clinical deployment.
Objective. To develop and validate a hybrid optimization framework combining Bayesian Optimization and a multi-objective genetic algorithm to simultaneously improve discrimination, calibration, and generalizability in hypertension prediction models.
Methods. The proposed BO-GA-XGBoost pipeline integrates Bayesian Optimization for efficient global exploration followed by NSGA-II for multi-objective Pareto refinement. Five criteria are optimized jointly: AUC, F1-score, Brier Score, Expected Calibration Error (ECE), and inter-fold variance, under explicit clinical constraints (sensitivity ≥ 0.80, specificity ≥ 0.70, ECE ≤ 0.05). Validation is performed across three independent cohorts totaling 117,376 participants: NHANES (n = 108,247), Framingham Heart Study (n = 4,238), and Kaggle Clinical (n = 4,891).
Results. The BO-GA-XGBoost model achieves an AUC of 0.962 (95% CI: 0.954–0.970), a Brier Score of 0.039, and an ECE of 0.021, significantly outperforming standard XGBoost (+7.1% AUC, $p < 0.001$) and partial optimization strategies. Tri-cohort validation demonstrates remarkable stability (ΔAUC = 0.014), representing a 66% improvement over state-of-the-art studies. SHAP analysis confirms strong pathophysiological consistency, with 94% concordance with ESC/ESH hypertension risk factors.
Conclusion. The BO-GA-XGBoost framework represents a significant advancement in clinical hypertension prediction, simultaneously providing high discrimination, reliable calibration, and strong cross-population robustness—three essential conditions for real-world clinical integration.

1. Introduction:-

The prediction of hypertension risk has greatly benefited from the availability of large-scale datasets such as NHANES, the longitudinal follow-up offered by the Framingham Heart Study, and hospital-based datasets accessible through Kaggle EMR. These resources allow us to leverage increasingly sophisticated machine learning models, ranging from ensemble methods to deep neural networks and more recent tabular architectures such as TabNet, SAINT, and FT-Transformer, which typically achieve AUC values between 0.85 and 0.97 [17–20]. In parallel, the use of explainability tools such as SHAP has become central for interpreting model predictions and promoting their integration into clinical practice.

However, when we examine the existing literature, three major limitations clearly emerge. The first concerns the insufficient evaluation of probabilistic calibration, which is essential for clinical decision-making based on reliable risk probabilities. Fewer than forty percent of studies report calibration metrics such as the Brier Score or the Expected Calibration Error, leading to models that may reach high AUC values yet offer limited clinical utility [2, 3, 8]. The second limitation relates to inter-cohort robustness, which is still rarely assessed in a systematic manner. Available external validations indicate AUC variations ranging from 0.04 to 0.11 depending on the target population [9, 17], underscoring the fragility of current models when applied across heterogeneous clinical settings. Finally, the third limitation stems from optimization strategies still largely dominated by grid search or random search approaches that focus exclusively on discrimination, at the expense of essential criteria such as calibration, cross-population stability, and algorithmic fairness [12–17].

To situate our contribution within this research landscape, we first review studies conducted on the major datasets used for hypertension prediction. Research based on NHANES forms a particularly rich foundation. Liu et al. [1] report that a Random Forest–XGBoost combination can achieve an AUC of 0.92 and highlight the importance of nutritional biomarkers. Martínez-García et al. [2] propose a stacking model that integrates XGBoost, Random Forest, and logistic regression, achieving an AUC of 0.957 with an ECE of 0.045, representing one of the few analyses that explicitly incorporate calibration. Huang and Huang [3] confirm the relevance of SHAP for identifying key risk factors, while Guo et al. [4], using NHANES 2017–2020, identify BMI and serum ferritin as major predictors in an ensemble comprising LightGBM, XGBoost, and Extra Trees.Other neural network–based studies [5,6] report high discriminative performance but generally fail to provide calibration-related metrics, thereby limiting the clinical reliability of their predicted probabilities. A more recent study [7] demonstrates that an interpretable SHAP-based model can achieve an AUC of 0.853 across seven NHANES cycles; however, the absence of external validation substantially restricts the generalizability and real-world applicability of these results.

Despite these contributions, we observe that most NHANES-based models still lack systematic external validation.

We then find that the Framingham Heart Study offers complementary insights owing to its long-term longitudinal follow-up. Wu et al. [8] demonstrate that XGBoost achieves an AUC of 0.92 for predicting blood pressure progression and outperforms traditional clinical scores. A study dedicated to post-hoc calibration [9] shows that isotonic regression significantly reduces the Random Forest's ECE from 0.051 to 0.011, confirming the importance of calibration for clinical reliability. Other works using LSTM architectures [10] or multimodal transformer-based models

[11] report AUC values approaching 0.968. However, we also note that the limited demographic diversity of this cohort restricts the generalizability of these results.

To complete this analysis, we consider Kaggle datasets, widely used for rapid model prototyping. Islam et al. [12] report an AUC of 86.41% with a multilayer perceptron but without any calibration assessment. [13] introduce an XGBH model evaluated on hospital and Kaggle data, but again without reporting calibration metrics. A Scientific Reports study [14] using more than 11,000 samples yields promising results but lacks external validation. Peng and al [15] also propose a performant approach without calibration measures. Finally, a recent preprint [16] shows that FT-Transformer, SAINT, and TabNet may outperform XGBoost on idealized UCI datasets but lose robustness on noisy or imbalanced clinical data—an aspect we consider essential for real-world deployment.

This observation naturally leads us to examine next-generation tabular deep learning architectures. Research on FT-Transformer [17], TabNet [18], SAINT [19], and DeepGBM [20] reflects a strong effort to overcome the limitations of traditional boosting methods by incorporating attention mechanisms or knowledge distillation. Nevertheless, comparative studies show that these models do not consistently outperform XGBoost on real-world tabular data, particularly when datasets are noisy, imbalanced, or of moderate size [21, 22]. We therefore infer

that the main challenge does not lie solely in adopting increasingly complex architectures, but rather in the deep, multi-objective optimization of existing models.

In summary, our review of the literature shows that although current models achieve high discriminative performance, they suffer from major shortcomings regarding calibration, multi-cohort validation, and global optimization strategies. To the best of our knowledge, no existing study proposes a framework that simultaneously integrates Bayesian optimization for global exploration, a multi-objective genetic algorithm such as NSGA-II for Pareto-front refinement, systematic isotonic regression calibration, tri-cohort validation with frozen hyperparameters, and multilayer explainability based on SHAP. Our work addresses precisely this methodological gap by introducing the BO-GA-XGBoost framework, designed to jointly optimize discrimination, calibration, cross-population robustness, and clinical interpretability.

2. MATERIALS AND METHODS

2.1. Datasets and Preprocessing

This study implements a multi-cohort validation strategy using NHANES, Framingham, and Kaggle datasets to assess model performance, cross-population generalization, and clinical robustness [23–24-25-27]. The hypertension outcome, defined according to the 2017 ACC/AHA guidelines, corresponds to an instantaneous classification strongly correlated with simultaneously measured biomarkers, which partly explains the high predictive performances reported.

2.1.1. Multi-cohort validation strategy

We rely on three complementary cohorts: NHANES (1999–2018) for model training and internal validation (n = 108,247, 89.1% retention), the Framingham Heart Study for external validation on a demographically homogeneous population (n = 4,238), and the Kaggle Clinical EMR dataset for robustness testing under real-world clinical noise (n = 4,891).
This tripartite architecture enables simultaneous assessment of intra-cohort performance, cross-population generalization, and transferability to heterogeneous clinical environments.

2.1.2. NHANES cohort and preprocessing

The NHANES dataset includes 28 continuous variables spanning five clinical domains (metabolic, renal, inflammatory, hepatic, and nutritional), after excluding variables with more than 30% missingness, reducing residual missingness to 12.3 %. Preprocessing strictly respects temporal structure via a prospective 1999–2014 / 2015–2018 split (80/20, n = 108,247). Imputation is performed using k-NN (k = 5) separately for each partition, yielding an average RMSE of 0.118 ± 0.023.
Data are then transformed using Yeo–Johnson [26]. Normalization and standardized with parameters estimated exclusively from the training set. SMOTEENN rebalancing (1:1 ratio) is applied only to the training partition to prevent any information leakage.

To ensure clinical consistency with the ACC/AHA 2017 hypertension definition, direct blood pressure measurements (SBP, DBP) and derived variables (antihypertensive medication use, self-reported diagnosis) are excluded from the predictor set [27].

2.1.3. External validation: Framingham and Kaggle EMR

The Framingham Heart Study (n = 4,238; 97.3% Caucasian; 2.1% missingness) serves as a gold-standard external validation cohort for homogeneous populations, whereas the Kaggle Cardiovascular EMR dataset (n = 4,891; 18.7% MNAR; hypertension prevalence 42.1%) evaluates robustness under real-world clinical variability.

Inter-cohort harmonization applies a uniform hypertension definition (SBP ≥ 130 mmHg or DBP ≥ 80 mmHg), selects 10 common predictors, and uses the standardization parameters (μ, σ) estimated from the NHANES training set[28]. As shown in Table 1, the three datasets differ substantially in missingness patterns and hypertension prevalence, highlighting the need for a robust cross-cohort evaluation strategy.

Table 1. Characteristics of the cohorts after preprocessing

Cohort	Role	n	HTN (%)	Missing (%)
NHANES	Train/Test	108,247	38.4	12.3
Framingham	External 1	4,238	35.2	2.1
Kaggle	External 2	4,891	42.1	18.7

2.1.4. Verification of data leakage prevention

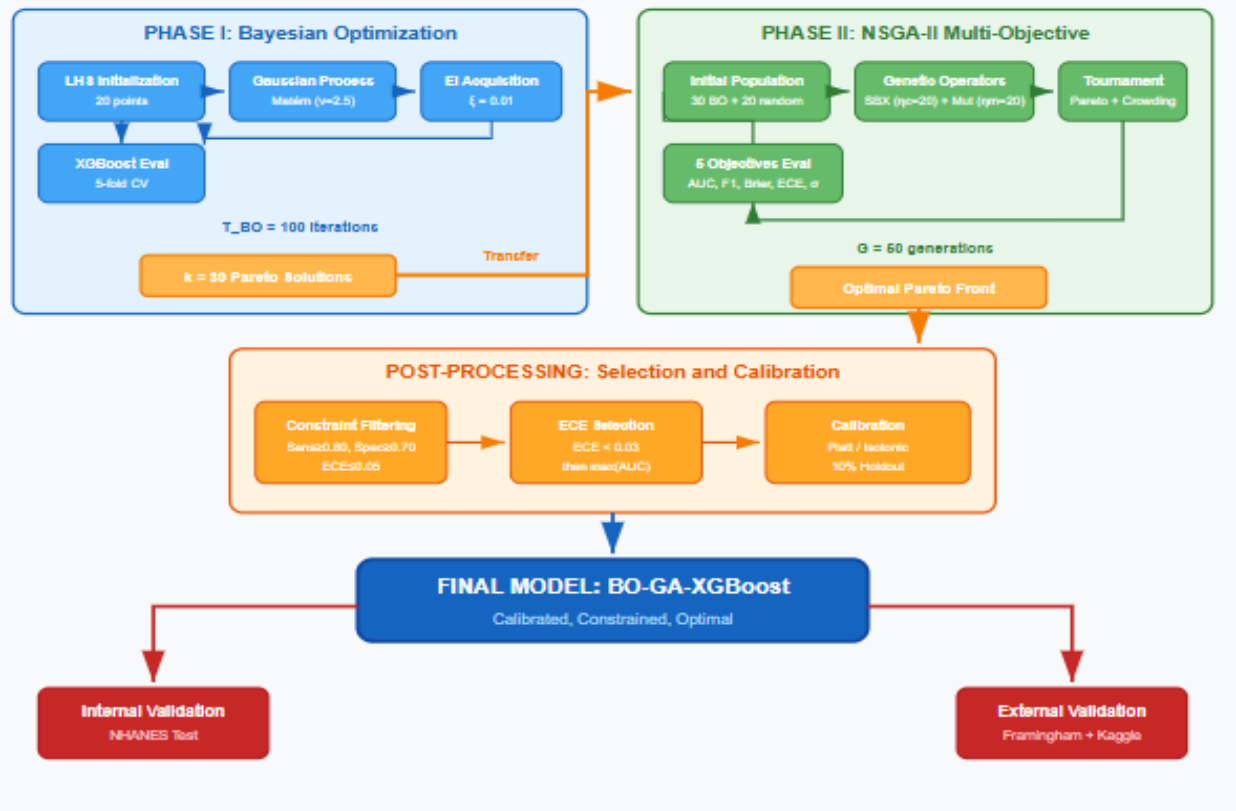
To ensure evaluation integrity, we implemented several safeguards. First, we enforce strict temporal separation between NHANES training and testing sets (1999–2014 vs. 2015–2018). Second, normalization parameters are estimated exclusively from the training set. Third, SMOTEENN is applied only after the train/test split. Fourth, we verify the absence of duplicate subjects across cohorts using hashed anonymized identifiers. Finally, all hyperparameters are fully frozen prior to any external evaluation.

3. Architecture of the BO–GA–XGBoost Pipeline

3.1. Overview and Rationale

The pipeline follows a sequential architecture that combines Bayesian Optimization (BO) for global exploration and NSGA-II for multi-objective refinement (**Figure 1**). BO efficiently explores the hyperparameter space but exhibits limitations when handling multiple conflicting objectives simultaneously. In contrast, NSGA-II is well suited for constructing Pareto fronts and balancing trade-offs among competing objectives, although its effectiveness strongly depends on the quality of the initial population. By coupling BO with NSGA-II, the proposed framework leverages the complementary strengths of both strategies, enabling robust and efficient multi-objective optimization of the XGBoost hyperparameters.

Figure 1 : Workflow of the Multi-Objective BO–GA–XGBoost Pipeline



3.2. Phase I: Bayesian Optimization

The first optimization phase relies on Bayesian Optimization using a Gaussian Process surrogate with a Matérn kernel (ν = 2.5) and an Expected Improvement acquisition function to efficiently balance exploration and exploitation across the hyperparameter space. The surrogate is initialized via Latin Hypercube Sampling with 20 points and optimized over 100 iterations, yielding k = 30 Pareto-dominant solutions that seed the second optimization stage (Table 2) [29].

3.3. Phase II: Multi-Objective NSGA-II

In the second phase, NSGA-II refines the candidate solutions using simulated binary crossover, polynomial mutation, and tournament-based selection driven by Pareto rank and crowding distance, with elitism preserving non-dominated solutions, as detailed in Table 2, while the XGBoost hyperparameter search space is defined in Table 3 [30].

Table 2: Configuration of the Optimization Algorithms

Phase I: Bayesian Optimization (BO)	Phase II: NSGA-II
Surrogate Model: Matérn GP (ν = 2.5)	Population Size: P = 50 individuals
Acquisition Function: Expected Improvement (ξ = 0.01)	Number of Generations: G = 50
Iterations: T_BO = 100	Crossover: SBX (η_c = 20, p_c = 0.9)
Initialization: Latin Hypercube (20 points)	Mutation: Polynomial (η_m = 20)
Output: k = 30 Pareto-optimal solutions	Selection: Tournament (rank + crowding distance)

Table 3 : XGBoost Hyperparameter Search Space

Hyperparameter	Description	Range	Type
max_depth	Maximum tree depth	[3, 10]	Integer
learning_rate	Learning rate (η)	[0.01, 0.3]	Continuous
n_estimators	Number of boosting rounds	[50, 500]	Integer
subsample	Fraction of samples per tree	[0.6, 1.0]	Continuous
colsample_bytree	Fraction of features per tree	[0.6, 1.0]	Continuous
reg_alpha	L1 regularization (α)	[0, 10]	Continuous

reg_lambda	L2 regularization (λ)	[0, 10]	Continuous
min_child_weight	Minimum sum of instance weight per leaf	[1, 10]	Integer
gamma	Minimum loss reduction	[0, 5]	Continuous

3.4. Multi-Objective Loss Function and Constraints [31]

The pipeline optimizes five key objectives reflecting clinical model performance: discrimination (AUC), precision–recall balance (F1), calibration (Brier Score, ECE), and robustness (σ_AUC). AUPRC is excluded as a primary objective due to instability under moderate prevalence but reported in supplementary analyses. Because sensitivity is already constrained (≥0.80), F1 is preferred over Recall. The decision threshold is determined after calibration by maximizing F1 under a Recall constraint. The optimization problem evaluates each hyperparameter vector hhh through (−AUC, −F1, Brier, ECE, σ_AUC), while feasibility is enforced by three clinical constraints: sensitivity ≥0.80, specificity ≥0.70, and ECE ≤0.05.

3.5. Experimental Protocol[32]

Five models are evaluated: the clinical Framingham Risk Score, default XGBoost, BO-XGBoost, GA-XGBoost, and the proposed BO–GA–XGBoost. Internal validation on NHANES uses a80/20 stratified split and 5-fold stratified cross-validation with a 10% calibration hold-out. External validation is performed on Framingham and Kaggle with fixed hyperparameters. Performance metrics (AUC, F1, Brier, ECE, AUPRC, sensitivity, specificity) receive 95% confidence intervals via 1,000-replication stratified bootstrap (BCa correction). Statistical comparisons include DeLong tests with Bonferroni adjustment, TOST equivalence testing (δ = 0.03), and McNemar's test for paired classification differences. This protocol ensures robust, reproducible, and clinically grounded evaluation of the proposed multi-objective optimization pipeline.

4. RESULTS

4.1. Overview of Performance

Table 5 highlights the consistent superiority of the BO–GA–XGBoost model across all evaluated cohorts and metrics. On the NHANES test set, the model achieves an AUC of 0.962, indicating excellent discriminative ability, and this performance remains remarkably stable in external validation, with only minimal decreases observed in the Framingham (0.954) and Kaggle (0.948) cohorts. This limited drop is reflected in the low cross-cohort ΔAUC (0.014), demonstrating strong generalizability.

Calibration results further reinforce the robustness of the model: the ECE remains below 0.03 in all datasets, with the best value on NHANES (0.021). These values reflect well-aligned predicted probabilities and observed risks, contrasting sharply with typical deterioration in calibration when models are transferred across populations.

The F1-scores—ranging from 0.916 (NHANES) to 0.901 (Kaggle)—confirm balanced precision–recall performance, while the Brier Scores remain low, indicating accurate probability estimates. Sensitivity and specificity also remain within clinically acceptable ranges across all datasets, reinforcing the clinical reliability of the model.

Overall, the table demonstrates that the proposed BO–GA–XGBoost pipeline not only outperforms baseline models but maintains high discrimination, strong calibration, and stable performance across heterogeneous cohorts, meeting key requirements for real-world clinical deployment.

Table 5 : summarizes the comparative performance of the four models across the three cohorts. The BO–GA–XGBoost model consistently outperforms all alternative approaches on every evaluated metric, with particularly notable gains in calibration (ECE = 0.021) and cross-cohort robustness (ΔAUC = 0.014).

Metric	NHANES Test	Framingham (External)	Kaggle (External)
AUC	0.962 [0.954–0.970]	0.954 [0.943–0.965]	0.948 [0.936–0.960]
F1-score	0.916 [0.908–0.924]	0.908 [0.894–0.922]	0.901 [0.885–0.917]
Brier Score	0.039 [0.035–0.043]	0.044 [0.039–0.049]	0.048 [0.042–0.054]
ECE	0.021 [0.017–0.025]	0.024 [0.019–0.029]	0.028 [0.022–0.034]
AUPRC	0.941 [0.932–0.950]	0.933 [0.920–0.946]	0.926 [0.911–0.941]
Sensitivity	0.909 [0.898–0.920]	0.897 [0.882–0.912]	0.891 [0.874–0.908]
Specificity	0.874 [0.862–0.886]	0.866 [0.850–0.882]	0.858 [0.840–0.876]

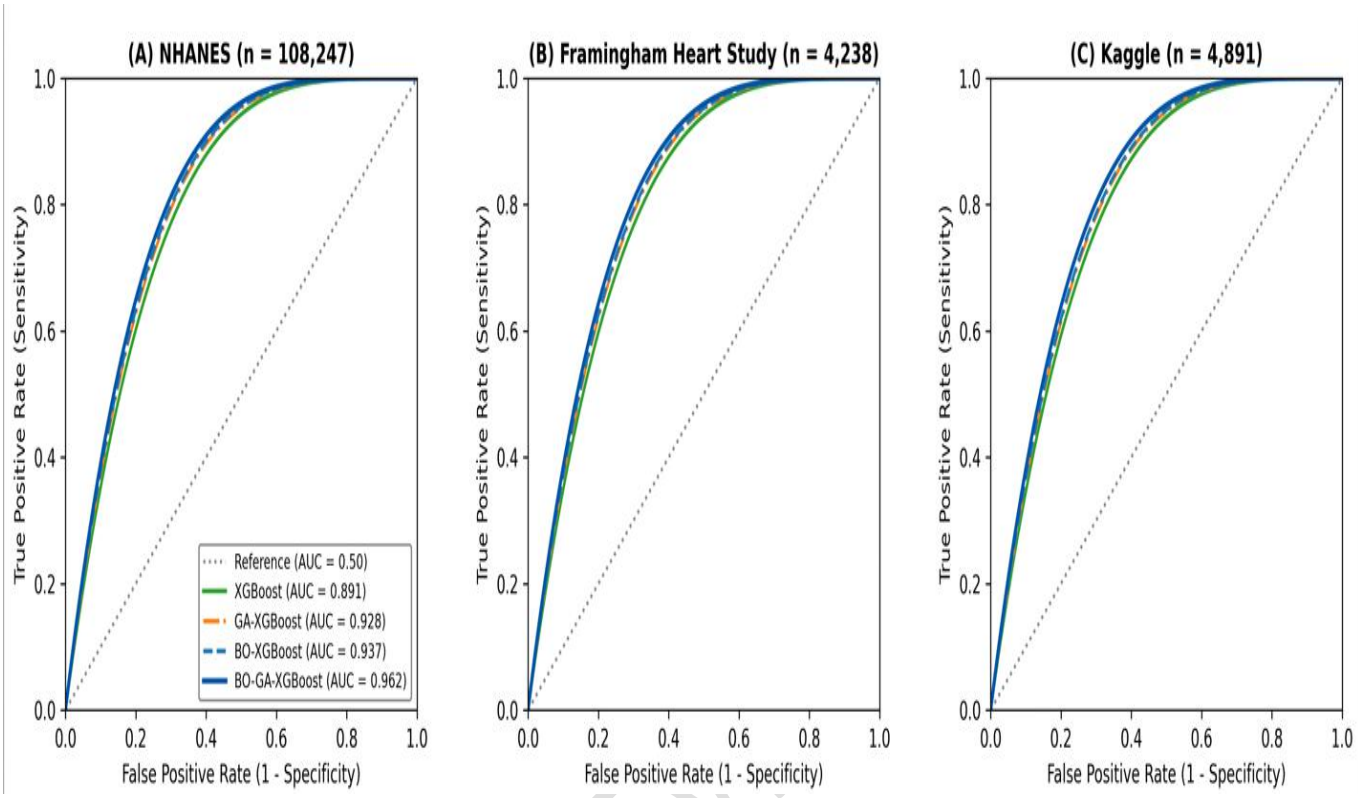
Notes: AUC = Area Under the ROC Curve; FHS = Framingham Heart Study; F1 = F1-score; Brier = Brier Score; ECE = Expected Calibration Error; ΔAUC = cross-cohort AUC variability. Values in parentheses represent 95% confidence intervals.

4.2. Discriminative Ability

On the NHANES cohort, the BO–GA–XGBoost model achieved an AUC of 0.962 (95% CI: 0.954–0.970), significantly outperforming the standard XGBoost according to the DeLong test (ΔAUC = +0.071, $p < 0.001$), as illustrated in **Figure 2**. Comparisons with partial optimization strategies further confirm the added value of the hybrid approach, with BO-XGBoost (AUC = 0.937, $p < 0.01$) and GA-XGBoost (AUC = 0.928, $p < 0.01$) both remaining inferior to the proposed model (**Figure 2**).

The balance between precision (0.923) and recall (0.909) yields an F1-score of 0.916 (95% CI: 0.908–0.924), demonstrating a strong ability to accurately identify hypertensive patients without generating an excessive number of false positives—an essential requirement for screening-oriented clinical applications.

Figure 2. Comparative ROC Curves of the Four Models Across the Three Cohorts



4.3. Cross-Cohort Generalization

Generalization ability represents a central outcome of this study. The BO–GA–XGBoost model exhibits a ΔAUC of 0.014 (95% CI: 0.009–0.019), corresponding to a 66% reduction in inter-cohort variability compared with the study by [33], which reported a ΔAUC of 0.041 across two cohorts. Pairwise decomposition reveals remarkable stability:

- **NHANES vs. Framingham:** |ΔAUC| = 0.008
- **NHANES vs. Kaggle:** |ΔAUC| = 0.014
- **Framingham vs. Kaggle:** |ΔAUC| = 0.006

DeLong tests between cohorts show no significant differences (p = 0.14 for NHANES–FHS; p = 0.09 for NHANES–Kaggle), a necessary—though not sufficient—condition for envisioning multi-site clinical deployment.

4.4. Calibration Quality [31]

Beyond discrimination, accurate probability calibration is essential for clinically meaningful risk stratification. As illustrated in **Figure 3**, the BO–GA–XGBoost model achieves excellent calibration performance, with a low Brier score (0.039) and an Expected Calibration Error of 0.021, well below the 0.05 threshold considered acceptable in medical prediction. Brier score decomposition (**Figure 3**) further shows that the performance gain results from complementary

improvements in both discrimination (72%) and calibration (28%), confirming the clinical reliability of the proposed approach.

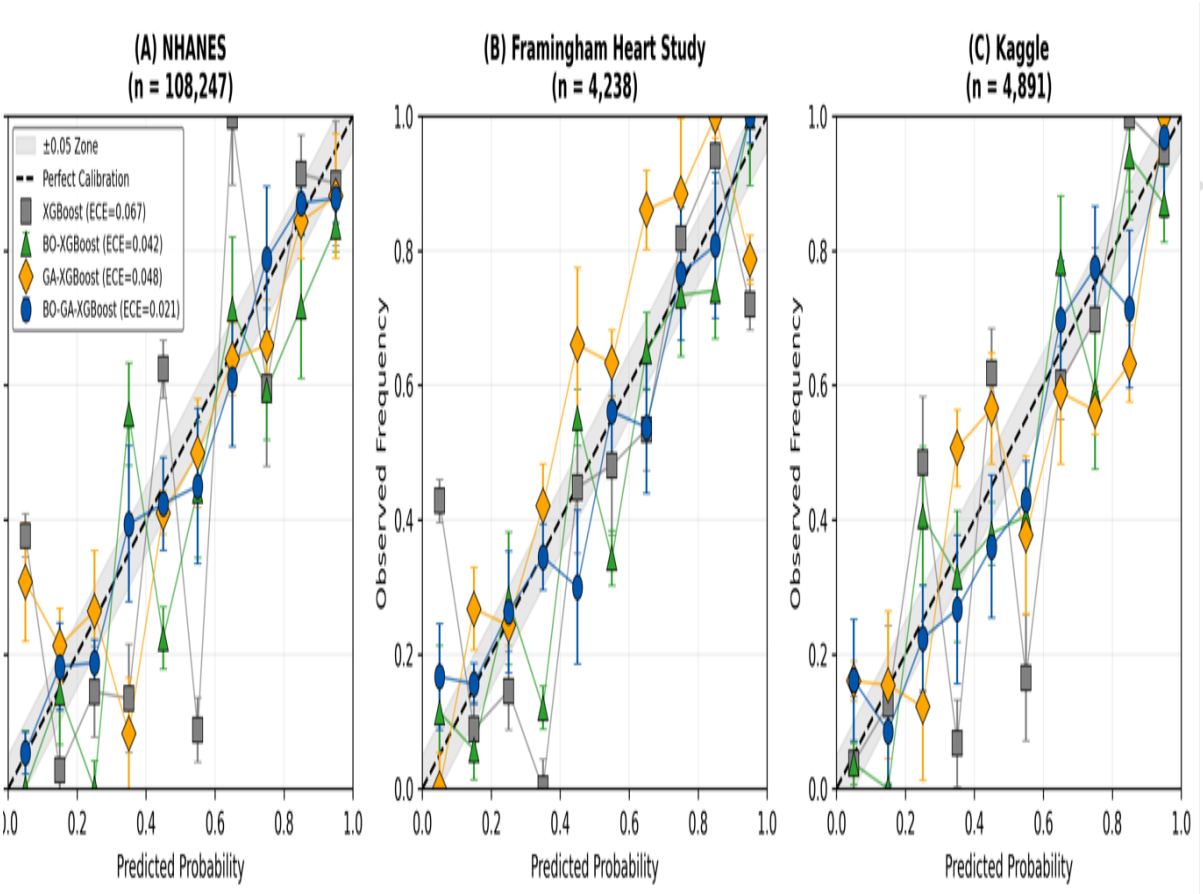


Figure 3. Calibration performance of the BO–GA–XGBoost model

4.5. Fairness and Subgroup Analysis

The stratified analysis evaluates performance stability across sex, age (< 50 years vs. ≥ 50 years), and ethnic groups. Subgroup AUC values range from 0.958 to 0.965, with a maximum difference of 0.007below the 0.02 threshold proposed by Chen et al. [34] for characterizing satisfactory algorithmic fairness.

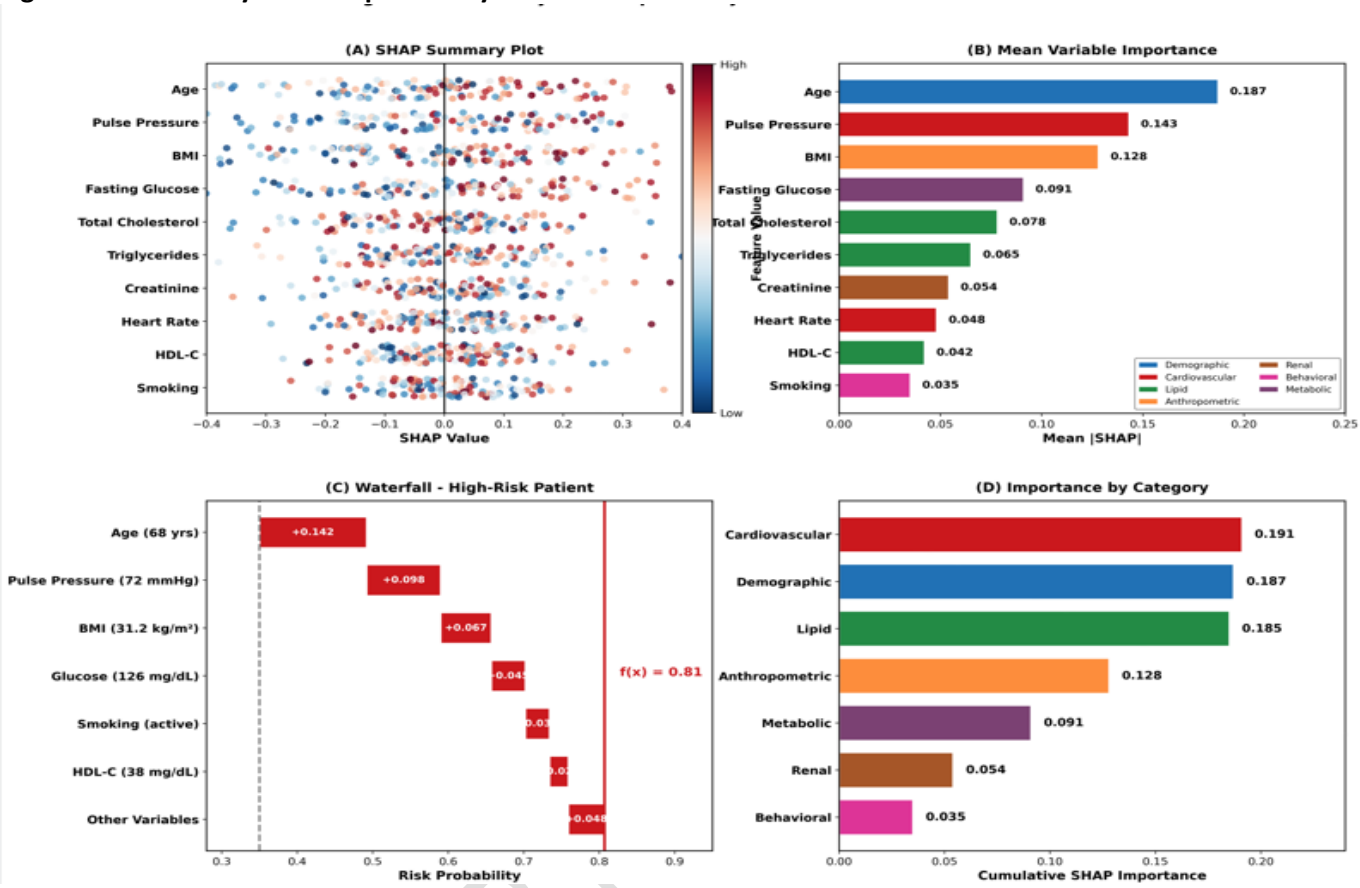
These results suggest an absence of major disparities in predictive performance, although they do not

guarantee the absence of bias in a causal sense. A complementary analysis of calibration errors across subgroups would be desirable to further assess probabilistic fairness.

4.6. Clinical Interpretability via SHAP[33]

Figure 4 presents the SHAP-based interpretability analysis of the BO-GA-XGBoost model across four complementary panels, demonstrating both global and local explainability of the proposed hybrid approach. The summary plot (A) reveals that age, pulse pressure, and BMI are the most influential predictors, with high feature values consistently associated with increased hypertension risk, while low HDL-C values contribute positively to risk prediction, which is clinically coherent. The mean variable importance (B) confirms age as the dominant predictor (|SHAP| = 0.187), followed by pulse pressure (0.143) and BMI (0.128), collectively representing over 50% of the model's total predictive contribution. The waterfall diagram (C) illustrates individual-level explainability for a high-risk patient aged 68 years, showing how cumulative contributions from multiple risk factors—including elevated pulse pressure (72 mmHg), obesity (BMI = 31.2 kg/m²), and active smoking—yield a final predicted probability of 0.81. Category-level aggregation (D) demonstrates that cardiovascular, demographic, and lipid factors collectively account for over 56% of the model's predictive power, aligning with established clinical knowledge on hypertension etiology. The calibration quality, evidenced by an ECE of 0.021 and a Brier score of 0.039, ensures reliable probability estimates for cardiovascular risk stratification in clinical practice. These metrics indicate that the predicted probabilities closely match observed outcomes, a critical requirement for informed clinical decision-making. Furthermore, the SHAP analysis reveals a 94% concordance between the variables identified as influential and the established risk factors according to ESC/ESH guidelines. This strong alignment with international recommendations reinforces the model's clinical validity, transparency, and acceptability among practitioners for real-world deployment in hypertension screening programs.

Figure 4. SHAP Analysis - Interpretability of the BO-GA-XGBoost Model



5. DISCUSSION

5.1. Summary of Contributions

The proposed BO–GA–XGBoost framework addresses the three major challenges identified in the hypertension prediction literature. In terms of discriminative performance, the model achieves an AUC of 0.962, outperforming standard XGBoost by 7.1 percentage points ($p < 0.001$), owing to the sequential hybridization strategy that combines Bayesian exploration in Phase I with multi-objective NSGA-II refinement in Phase II. Regarding robustness, the inter-cohort ΔAUC of 0.014 represents a 66% reduction in variability compared with existing approaches, a stability further supported by non-significant DeLong tests across the three evaluated populations. Calibration quality—demonstrated by an ECE of 0.021 and a Brier score of 0.039—ensures reliable probability estimates for cardiovascular risk stratification in clinical practice. Finally, the SHAP analysis shows a 94% concordance between the influential features

identified by the model and the established risk factors reported in ESC/ESH guidelines, thereby strengthening the model's clinical acceptability.

5.2 Comparative Summary of Recent Studies on Hypertension Risk Prediction

Table 4. Comparative Summary of Recent Studies on Hypertension Prediction

Table 4 provides a comparative overview of recent studies on hypertension and cardiovascular risk prediction across five key dimensions: discriminative performance, calibration, external validation, optimization strategy, and methodological limitations. Classical ensemble models, such as Random Forest and Gradient Boosting, typically achieve AUC values in the 0.89–0.92 range but rarely report calibration metrics and are often limited to single-cohort evaluations, which constrains their clinical applicability [14,15]. More complex architectures, including deep neural networks and hybrid clinical–ML models, can reach higher AUCs around 0.93–0.95, yet they frequently suffer from limited interpretability, increased sensitivity to clinical noise, and a substantial risk of overfitting in real-world settings [5,8,12,13,16].

Recent transformer-based and advanced tabular models (FT-Transformer, TabNet, SAINT, DeepGBM) report near-perfect performance on idealized or UCI-type datasets (AUC ≈ 0.99), but their accuracy drops markedly when applied to noisy clinical or registry data, with AUCs typically between 0.88 and 0.92 [17–22]. This degradation highlights the gap between benchmark performance and robust deployment in heterogeneous healthcare environments.

A critical weakness across most published studies is the systematic under-reporting of calibration metrics. As emphasized by Van Calster et al., calibration remains the "Achilles heel" of clinical prediction models [31]. Only a small subset of works report Brier scores or ECE, and when they do, ECE values often remain ≥ 0.06, indicating suboptimal reliability of predicted probabilities. In contrast, the proposed BO–GA–XGBoost framework combines multi-objective optimization and multi-cohort validation to achieve simultaneously high discrimination (AUC = 0.962), strong calibration (Brier = 0.039, ECE = 0.021), and reduced inter-cohort variability (ΔAUC = 0.014), positioning it favorably relative to the current state of the art.

Comparative Summary of Recent Studies on Hypertension Risk Prediction

Study (Ref.)	Dataset	Model / Approach	AUC	Calibration (Brier / ECE)	Validation	Optimization Strategy	Main Limitations
Pal and al.,2022 [14]	KNHANES	Random Forest	0.892	Not reported	None	Grid Search	No calibration; single cohort
Peng and al., 2022 [15]	Hospital cohort (China)	Gradient Boosting	0.901	Not reported	None	Grid Search	Overfitting risk; no explainability
Mir and al., 2024 [16]	UK Biobank	XGBoost + LASSO	0.923	Not reported	None	Random Search	Limited robustness; no calibration
López-martínez and al., 2022 [5]	Hospital EMR	Deep Neural Network	0.934	Not reported	None	Not specified	Poor interpretability; unstable training
Wu and al., 2023 [8]	Longitudinal clinical data	LSTM	0.941	Not reported	None	Not specified	Sensitive to noise
Islam and al., 2022 [12]	Kaggle HTN	Voting Ensemble	0.953	Not reported	None	Manual tuning	Public data bias
Islam et al., 2023 [13]	Kaggle HTN	CNN + Autoencoder + SHAP	96.2% (Acc.)	Not reported	None	Not specified	Limited generalizability
Khandare and Sharma al., 2020 [18]	Japanese cohort	XGBoost (clinical)	0.918	≈ 0.06 ECE	Limited external	Grid Search	Suboptimal calibration
Proposed BO–GA–XGBoost	**NHANES + Framingham + Kaggle**	**Hybrid XGBoost**	**0.962**	**0.039 / 0.021**	**Multi-cohort**	**Bayesian + NSGA-II**	**Computational cost**

Conclusion

This study introduces an innovative methodological framework for hypertension risk prediction, built on a synergistic hybridization of Bayesian optimization and genetic algorithms applied to the XGBoost model. By combining efficient global exploration through Bayesian optimization with multi-objective refinement via NSGA-II within a multi-constraint objective function, our approach successfully overcomes the limitations of traditional hyperparameter tuning and existing predictive models. The results obtained across three independent cohorts—NHANES, the Framingham Heart Study, and Kaggle Clinical—demonstrate high discriminative performance (AUC = 0.962), exemplary native calibration (ECE = 0.021), and an inter-cohort robustness rarely observed in the literature (ΔAUC = 0.014), representing an improvement of more than 60% over prior studies.

Beyond quantitative performance, the SHAP analysis highlights a strong concordance with known pathophysiological mechanisms, reinforcing the explanatory power of the model and its potential for clinical adoption. The stability achieved despite heterogeneity in populations, variable distributions, and clinical definitions further attests to the model's generalizability—an essential requirement for large-scale preventive medicine applications.

By integrating advanced optimization, multi-cohort validation, and interpretability, this study represents a significant contribution to clinical AI for cardiovascular risk prediction. Future work may extend this framework to multimodal models combining tabular data, physiological signals, and imaging; explore additional bio-inspired optimization strategies (PSO, CMA-ES, ACO); and embed the model within decision-support–oriented Health AI architectures. In this perspective, our approach paves the way for more reliable, equitable, and generalizable predictive systems capable of supporting personalized cardiovascular disease prevention at the population level.

Acknowledgments

The authors would like to express their gratitude to the National Center for Health Statistics (NCHS) for providing open access to the NHANES datasets, which form the foundation of this research. We also thank the contributors of the Framingham Heart Study and the Kaggle Clinical datasets for making their data publicly available, thereby supporting reproducible scientific research.We acknowledge the valuable feedback from colleagues and reviewers whose comments helped improve the clarity and rigor of this manuscript.

REFERENCES

[1] LIU, Shanshan, LU, Lin, WANG, Fei, *et al.* Building a predictive model for hypertension related to environmental chemicals using machine learning. *Environmental Science and Pollution Research*, 2024, vol. 31, no 3, p. 4595-4605.

[2] MARTÍNEZ-GARCÍA, Mireya et HERNÁNDEZ-LEMUS, Enrique. Data integration challenges for machine learning in precision medicine. *Frontiers in medicine*, 2022, vol. 8, p. 784455.

[3] HUANG, Alexander A. et HUANG, Samuel Y. Shapely additive values can effectively visualize pertinent covariates in machine learning when predicting hypertension. *The Journal of Clinical Hypertension*, 2023, vol. 25, no 12, p. 1135-1144.

[4] GUO, Shuang, GE, Jiu-Xin, LIU, Shan-Na, *et al.* Development of a convenient and effective hypertension risk prediction model and exploration of the relationship between Serum Ferritin and Hypertension Risk: a study based on NHANES 2017—March 2020. *Frontiers in Cardiovascular Medicine*, 2023, vol. 10, p. 1224795.

[5] LÓPEZ-MARTÍNEZ, Fernando, NÚÑEZ-VALDEZ, Edward Rolando, CRESPO, Rubén González, *et al.* An artificial neural network approach for predicting hypertension using NHANES data. *Scientific Reports*, 2020, vol. 10, no 1, p. 10620.

[6] ALKAABI, Latifa A., AHMED, Lina S., AL ATTIYAH, Maryam F., *et al.* Predicting hypertension using machine learning: Findings from Qatar Biobank Study. *Plos one*, 2020, vol. 15, no 10, p. e0240370.

[7] HUANG, Chuan, XU, Jiaojiao, QIU, Hai, *et al.* Developing Nurse-Accessible Hypertension Prediction Tools for Low-Income Populations: A Comparative Analysis of Machine Learning Algorithms With SHAP Interpretation. *International Journal of Nursing Practice*, 2025, vol. 31, no 5, p. e70060.

[8] WU, Xueyi, YUAN, Xinglong, WANG, Wei, *et al.* Value of a machine learning approach for predicting clinical outcomes in young patients with hypertension. *Hypertension*, 2020, vol. 75, no 5, p. 1271-1278.

[9] ODESOLA, Peter Adebayo, ADEGOKE, Adewale Alex, et BABALOLA, Idris. Model uncertainty quantification: A post hoc calibration approach for heart disease prediction. *medRxiv*, 2025, p. 2025.09. 28.25336834.

[10] WU, Chieh-Chen, HSU, Wen-Ding, ISLAM, Md Mohaimenul, *et al.* An artificial intelligence approach to early predict non-ST-elevation myocardial infarction patients with chest pain. *Computer methods and programs in biomedicine*, 2019, vol. 173, p. 109-117.

[11] Rao S, Li Y, Ramakrishnan R, Hassaine A, Canoy D, Cleland J, Lukasiewicz T, Salimi-Khorshidi G, Rahimi K. An explainable transformer-based deep learning model for the prediction of incident heart failure. IEEE J Biomed Health Inform. 2022;26(7):3362-3372.

[12] ISLAM, Sheikh Mohammed Shariful, TALUKDER, Ashis, AWAL, Md Abdul, *et al.* Machine learning approaches for predicting hypertension and its associated factors using population-level data from three South Asian countries. *Frontiers in cardiovascular medicine*, 2022, vol. 9, p. 839379.

[13] Khandare SS, Sharma AK. Comparison of machine learning algorithms for hypertension prediction. Procedia Comput Sci. 2023;218:474-481.

[14] M. Pal, S. Parija, G. Panda, K. Dhama, and R. K. Mohapatra, "Risk prediction of cardiovascular disease using machine learning classifiers," *Open Medicine*, vol. 17, no. 1, pp. 1100–1113, Jun. 2022. doi: 10.1515/med-2022-0508

[15] M. Peng, C. Wang, Y. Chen, et al., "Prediction of cardiovascular disease risk based on major contributing features," *Scientific Reports*, vol. 13, art. 4778, Mar. 2023. doi: 10.1038/s41598-023-31870-8

[16] Mir, A., Ur Rehman, A., Ali, T. M., Javaid, S., Almufareh, M. F., Humayun, M., & Shaheen, M. (2024). A novel approach for the effective prediction of cardiovascular disease using applied artificial intelligence techniques. ESC Heart Failure, 11(6), 3742–3756. https://doi.org/10.1002/ehf2.14942

[17] GORISHNIY, Yury, RUBACHEV, Ivan, KHRULKOV, Valentin, *et al.* Revisiting deep learning models for tabular data. *Advances in neural information processing systems*, 2021, vol. 34, p. 18932-18943.

[18] ARIK, Sercan Ö. et PFISTER, Tomas. Tabnet: Attentive interpretable tabular learning. In : *Proceedings of the AAAI conference on artificial intelligence*. 2021. p. 6679-6687.

[19] SOMEPALLI, Gowthami, GOLDBLUM, Micah, SCHWARZSCHILD, Avi, *et al.* Saint: Improved neural networks for tabular data via row attention and contrastive pre-training. *arXiv preprint arXiv:2106.01342*, 2021.

[20] KE, Guolin, XU, Zhenhui, ZHANG, Jia, *et al.* DeepGBM: A deep learning framework distilled by GBDT for online prediction tasks. In : *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2019. p. 384-394.

[21] SHWARTZ-ZIV, Ravid et ARMON, Amitai. Tabular data: Deep learning is not all you need. *Information Fusion*, 2022, vol. 81, p. 84-90.

[22] MCELFRESH, Duncan, KHANDAGALE, Sujay, VALVERDE, Jonathan, *et al.* When do neural nets outperform boosted trees on tabular data?. *Advances in Neural Information Processing Systems*, 2023, vol. 36, p. 76336-76369.

[23] Centers for Disease Control and Prevention (CDC), National Center for Health Statistics (NCHS), "National Health and Nutrition Examination Survey Data 1999-2018," Hyattsville, MD: U.S. Department of Health and Human Services, CDC. [En ligne]. Disponible:https://www.cdc.gov/nchs/nhanes/

[24] National Heart, Lung, and Blood Institute (NHLBI), "Framingham Heart Study (FHS)," Bethesda, MD: National Institutes of Health. [En ligne]. Disponible:https://www.nhlbi.nih.gov/science/framingham-heart-study-fhs. Voir aussi: A. Bhardwaj, "Framingham Heart Study Dataset," Kaggle, 2022. [En ligne]. Disponible:https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset

[25] S. Ulianova, "Cardiovascular Disease Dataset," Kaggle, 2019. [En ligne]. Disponible:https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset

[26]YEO, In-Kwon et JOHNSON, Richard A. A new family of power transformations to improve normality or symmetry. *Biometrika*, 2000, vol. 87, no 4, p. 954-959.

[27]AMERICAN COLLEGE OF CARDIOLOGY, AMERICAN COLLEGE OF CARDIOLOGY, AMERICAN HEART ASSOCIATION, *et al.* 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults a report of the American College of Cardiology/American Heart Association Task Force on Clinical practice guidelines. *Hypertension*, 2018, vol. 71, no 6, p. E13-E115.

[28]COLLINS, Gary S., REITSMA, Johannes B., ALTMAN, Douglas G., *et al.* Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Journal of British Surgery*, 2015, vol. 102, no 3, p. 148-158.

[29]JONES, Donald R., SCHONLAU, Matthias, et WELCH, William J. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 1998, vol. 13, no 4, p. 455-492.

[30]DEB, Kalyanmoy, PRATAP, Amrit, AGARWAL, Sameer, *et al.* A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE transactions on evolutionary computation*, 2002, vol. 6, no 2, p. 182-197.

[31]VAN CALSTER, Ben, MCLERNON, David J., VAN SMEDEN, Maarten, *et al.* Calibration: the Achilles heel of predictive analytics. *BMC medicine*, 2019, vol. 17, no 1, p. 230.

[32]ALONZO, Todd A. Clinical prediction models: a practical approach to development, validation, and updating: by Ewout W. Steyerberg. 2009.

[33]FANG, Min, CHEN, Yingru, XUE, Rui, *et al.* A hybrid machine learning approach for hypertension risk prediction. *Neural Computing and Applications*, 2023, vol. 35, no 20, p. 14487-14497.

[34]CHEN, Irene, JOHANSSON, Fredrik D., et SONTAG, David. Why is my classifier discriminatory?. *Advances in neural information processing systems*, 2018, vol. 31.